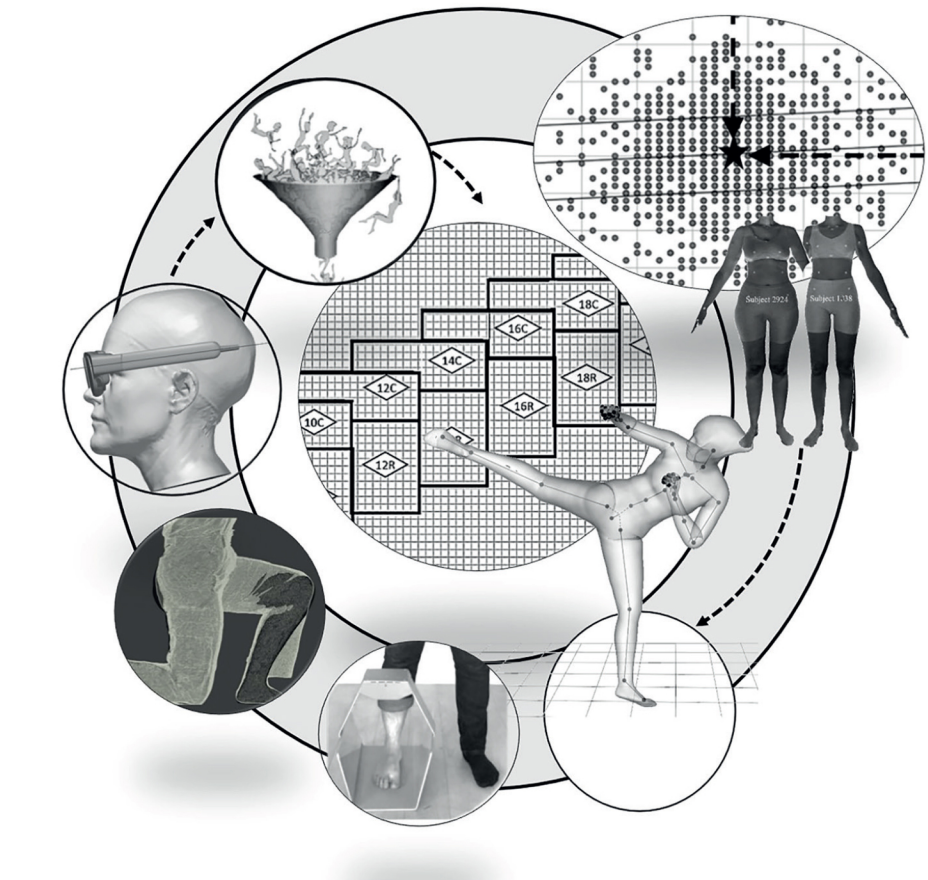
PRODUCT FIT AND SIZING

Sustainable Product Evaluation, Engineering, and Design



Kathleen M. Robinette, Daisy Veitch, Sandra Alemany, and Karen Bredenkamp

CRC Press
Taylor & Francis Group

Product Fit and Sizing

In this book, for the first time, the complexity of assessing fit and using fittings in the product design process is addressed from a scientific and systems engineering perspective. It includes methods to represent the anthropometry of the target market, good practices to develop protocols for more reliable and consistent fit testing, methods for developing and maintaining a fit database, comprehensive statistical analyses needed for fit and sizing analysis, and instructions for selecting and modeling cases for new product development.

Product Fit and Sizing: Sustainable Product Evaluation, Engineering, and Design offers step-by-step instructions for the evaluation, engineering, and design of existing and new products and includes real-world examples of mass-produced apparel, head wearables, and footwear products. It also explains how to develop a sustainable fit standard for fit and sizing continuity for all styles across all seasons and iterations.

This book is intended for industry professionals and undergraduate and graduate education to prepare students for design and engineering jobs. For organizations that purchase uniforms or protective equipment and apparel, it also provides instructions for purchasing professionals to evaluate the suitability of wearable products for their population.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Product Fit and Sizing

Sustainable Product Evaluation, Engineering, and Design

Kathleen M. Robinette, Daisy Veitch,
Sandra Alemany, and Karen Bredenkamp

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Designed cover image: Kathleen M. Robinette, Daisy Veitch, Sandra Alemany and Karen Bredenkamp

First edition published 2025
by CRC Press
2385 NW Executive Center Drive, Suite 320, Boca Raton FL 33431

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

WEAR International (World Engineering Anthropometry Resource) Ltd

ISBN: 978-1-032-49118-9 (hbk)
ISBN: 978-1-032-50243-4 (pbk)
ISBN: 978-1-003-39753-3 (ebk)

DOI: 10.1201/9781003397533

Typeset in Times LT Std
by KnowledgeWorks Global Ltd.

Access the Support Material: www.Routledge.com/9781032491189

Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Preface

Nearly everyone has experienced fit problems with some wearable product and, as a result, few people have confidence in buying something without trying it on. Even if we have the option of trying it on, we may not find any size that fits well. We may have to shop in a department that doesn't have the type of items we need, such as a tall woman forced to shop in menswear or a small person forced to shop in childrenswear department. Being small doesn't mean you want to wear girly pink frilly knickers!

This is also a big problem for retailers who must decide how many of each size to buy and stock for good sales with minimal waste. For online shopping, fit problems result in returns with extra shipping costs. Shipping costs have gotten to be so high that some online apparel companies have been offering to refund a large part of the cost of the product rather than paying for the return shipping! Fit is not only a problem for fashion apparel but even more difficult for protective equipment or special wearables, such as helmets, eyeglasses, continuous positive airway pressure (CPAP) masks, and nitrile gloves, which we may purchase only once or infrequently. This problem exists with larger cost implications for organizations, such as hospitals, fire departments, law enforcement agencies, or the military, having to acquire sizes in large quantities.

Why, after hundreds of years of collecting body measurements, decades of three-dimensional (3D) human scanning and biomechanical modeling, and numerous textbooks on engineering anthropometry, are we still having problems with fit? The reason is a lack of fit data and no guidebook that explains how to get it. There is no database of fit test results and no resource that tells us reliably what product proportioning will fit a given body. The proportioning and sizing for fit success varies depending upon the design, the materials, the target population, the intended function of the product, the style, the other products with which it must interface, and more.

This book describes the process needed to design and assess wearable products effectively and select the best sizes for any population or any individual. We refer to it as the Sustainable Product Evaluation, Engineering, and Design (SPEED) process. The secret to good fit and product functionality is to measure, validate, and document early and throughout the development of the product, who we fit, how well we fit them, and how they relate to our intended wearer population. With this information, it is possible to make informed decisions about the design, the adjustment mechanisms (such as the type and number of pads, elastic straps, lacing, and belts), the number and assortment of sizes, and more to best accommodate your target market with the least amount of sizes and cost.

The authors, with 125 years of combined experience, have put together this textbook to capture and document the best methods and to help others learn from their experience. This is the book we wish we had had, first when we were just starting out in the business, then to use as a quick reference for procedures we only use

occasionally, and finally to use a textbook for our students and for co-workers who were carrying on our work.

We have spent more than three years pulling all the materials together and testing our processes and descriptions to ensure they work well for all types of products and industries. We have even used the draft of the book when explaining processes to our current customers and are sure we will continue to use it after it is published. We hope it will prove to be useful to anyone who wants to create well-designed and well-fitting products.

About the Authors

Kathleen M. Robinette is a research consultant specializing in anthropometry, biostatistics, and fit and sizing for product development and assessment. She has more than 45 years' experience, spearheading the development, management, and transitioning of new technologies in the field of engineering anthropometry and led the field in the development of 3D automated human scanning and modeling for product design and evaluation. She planned, organized, negotiated, and directed the first successful 3D whole body human measurement survey (CAESAR), which produced more than 4,000 whole body models which continue to be used around the world today. She is a fellow of the Air Force Research Laboratory from which she retired after 30 years of service and is an honorary fellow of the Human Factors and Ergonomics Society. She was professor and head of the Department of Design, Housing, and Merchandising at Oklahoma State University 2012–2015 and she established and directed the Human Factors department for Magic Leap Inc. 2015–2017 implementing fit mapping into the product development process. Kathleen has a Ph.D. in Biostatistics and Epidemiology from the University of Cincinnati, an M.S. in Mathematics/Statistics from Wright State University, and a B.A. in Anthropology from Wright State University.

Daisy Veitch is currently the chief technology officer and head of Anthropometry for Anthrotech Inc. and has served as an anthropometry and fit consultant for commercial apparel industry for more than 25 years. She worked with Flinders Medical Center to develop and refine 3D body scanning for medical applications. She is the owner of a U.S. design patent and has registered designs in Australia, Europe, United States, and the European Community. She also directed the Australian National Size and Shape Survey in 2002. She worked in industry for ten years doing technical garment construction and serving as product engineer for an apparel company. As a recognized expert, she was appointed as an International Judge for APDeC 2013 (Asia Pacific Design Challenge; http://apdec.net). Daisy developed her fashion and design skills in Adelaide, Australia, beginning with the Australian Wool Corporation Young Designer Award and Queen Elizabeth II Silver Jubilee Award for Young Australians, which took her to Paris, France where she studied at the La Chambre Syndicale de la Couture Parisienne. She received her Ph.D. from TU Delft (Industrial Design Engineering with a specialization in Medisign). She is a founding member of World Engineering Anthropometry Resources (WEAR) and served as treasurer and secretary-general.

Sandra Alemany is a research scientist at the Instituto de Biomecánica de Valencia (IBV) and founded the Anthropometry Research Group in 2015. She led large-scale anthropometry surveys in Europe using 3D scanning technology and has experience in applying anthropometry to improve wearable fit including footwear, electronic devices, orthotics, insoles, and clothing. Recent developments include a 4D body scanner in movement and the development of two mobile applications to generate 3D body shapes from photographs. She is an expert advisor on European Standardization Committees of Anthropometry and Size System of Clothing and an expert reviewer of R&D projects for the European Commission. She is currently serving as co-chair for the Anthropometry Technical Committee for the International Ergonomics Association. She received her Ph.D. from the University Polytechnic of Valencia in 2023 with research about fit and clothing size prediction from anthropometry.

Karen Bredenkamp currently heads up the Human Factors team at Magic Leap Inc. Karen has more than 20 years' industry experience in anthropometry survey design, data collection, analysis, and implementation in wearable product and workstation design, as well as wearable product fit research as part of the product selection or development processes. Between 2000 and 2016, Karen was employed at Ergonomics Technology, a division of the Armaments Corporation of South Africa (ARMSCOR), where she had a core role in the establishment of the 3D whole-body and foot anthropometry databases for the ethnically diverse and unique South African National Defence Force (SANDF) population. Her activities furthermore involved providing anthropometry and fit inputs and evaluation support for product design as well as purchasing of SANDF clothing, footwear, protective wearable products, workstations, and occupant environments. Karen is currently serving as co-chair for the Anthropometry Technical Committee for the International Ergonomics Association. She has an M.Sc. in Biomedical Engineering from the University of Cape Town and a B.Eng. in Mechanical Engineering from Stellenbosch University, South Africa.

Acknowledgments

The authors would like to thank the many people and groups who supported us. We would like to thank the WEAR Association for bringing us together and suggesting we write a book and for funding the open source digital version. We thank Lorraine Mac Duff for her thorough and thoughtful editing of all our draft manuscripts, Janice Larsen for her insights regarding apparel industry practices, Institute of Biomechanics of Valencia (IBV) and Verna Blewett for help with formatting and reviewing Chapter 5.

Special thanks go to the Footwear Research Team of the IBV. The knowledge synthetized in Chapter 7 is the result of a long research trajectory in anthropometry, biomechanics, and human factors applied to footwear innovation.

We would also like to thank our families who supported us while we worked and met bi-weekly for over three years. We particularly would like to thank Ron Robinette, Henry and Lilian Fellner, Riaan, Emma, and Andru Bredenkamp and Martina and Fernando Villanueva.

Finally, we would like to dedicate this book to our friend and colleague, Prof. Regis Mollard, who supported and encouraged us, brought the four of us together for the first time with his short course in 2007, but who sadly passed away while we were writing this book.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Abbreviations

CAD	Computer Aided Design
CAM	Computed Aided Manufacture
COF	Concept-of-Fit
ISO	International Standards Organization
PCA	Principal Component Analysis
PPE	Personal Protective Equipment
SE	Standard Error
SEM	Standard Error of the Mean
SPEED	Sustainable Product Evaluation, Engineering, and Design
STD	Standard Deviation
TP	Target Population

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1 Introduction

Kathleen M. Robinette, Daisy Veitch, Sandra Alemany, and Karen Bredenkamp

ABSTRACT

This book is about the design and evaluation of anything and everything worn on the body. It is a guide to help designers, manufacturers, retailers, and procurement organizations:

- Solve the persistent fit problems
- Maintain a good fit for all styles of a wearable product and across all seasons or iterations
- Optimize the sizes manufactured or purchased by:
  - Aligning the sizes with the target market
  - Adding missing sizes and
  - Eliminating redundant or unnecessary sizes
- Communicate the range of fit in each size effectively
- Predict correct sizing for individuals based on body measurements
- Ensure products we buy for our store or organization will fit our wearers, and
- Make design, size, and fit decisions with confidence

In this chapter, we introduce the overarching systems engineering approach to wearable design that we refer to as the Sustainable Product Evaluation, Engineering, and Design (SPEED) process. The process is illustrated with a diagram that serves as a roadmap for the book, intended to help the reader understand where they are in the process and where to find the information needed to perform each part effectively.

Nearly everyone has experienced fit problems with a wearable product at one or other time in their life. It is not only a problem for fashion apparel but is even more difficult for protective equipment or special wearables, such as helmets, eyeglasses, Continuous Positive Airway Pressure (CPAP) masks, wearable technology, and nitrile gloves, which we may purchase only once or infrequently and where the fit is critical to its intended use. This problem exists with larger cost implications for organizations, such as hospitals, fire departments, law enforcement agencies, or the military, having to acquire sizes in large quantities. Even if we have the option of trying it on, we may not find any size that fits well. We may have to shop in a department that doesn't have the type of items we need, such as a small person forced to shop in the childrenswear department, or a tall woman forced to shop in menswear. Being small doesn't mean you want to wear girly pink frilly knickers!

DOI: 10.1201/9781003397533-1

This is also a big problem for retailers, who must decide how many of each size to buy and stock. For online shopping, shoppers do not have confidence that the product they buy will fit, and fit problems result in returns with extra shipping costs. Shipping costs have gotten to be so high that some online apparel companies have been offering to refund a large part of the cost of the product rather than paying for the return shipping! These products end up in the garbage dump.

The United States' Environmental Protection Agency (EPA) measures the generation, recycling, composting, combustion with energy recovery, and landfilling of textile materials in municipal solid waste (MSW) and publishes the findings on the official website (EPA, 2020). In 2017, they estimated that clothing and footwear industries in the United States generated 12.8 million tons of MSW, 8.9 million tons of which ended up in landfills. Only 1.7 million tons were recycled. The amount due to wasted sizes is not known, but there is a lot of room for improvement. Kay Liu (2018) estimated the "fashion and textile industry sees a loss of over $500 billion through the take-make-dispose production and consumption model". She stated, "An enormous 80% of the environmental impact is decided at the design stage". While this number is debatable, we agree that there is a huge opportunity for bringing more sustainable products to market via "better" design.

Anyone who is dealing with sizing in the apparel industry today knows that there are still many problems with sizing and fit. For example, in 2019 Emma Spedding (2019) spoke about the industry's sizing problem. She interviewed editors, designers, model bookers, and curve models about fit and sizing and they acknowledged there are still substantial fit and sizing problems, but they are not clear about how to solve them. Some of them felt that they were missing sizes because they were getting so many returns, some thought they might need sizes between the sizes they had because of different proportioning, and some noted that they needed to change but grading is not easy. At least one person felt they were missing out on billions of pounds by not getting it right.

All the commentators agreed there are many problems, but equally, each commentator had differing opinions on how to solve them, and no evidence that their proposed solution would work. This is because they had insufficient evidence about the source of the problem. Poor data and lack of evidence about the source of the problem lead to bad sizing decisions. We don't want you to think that more in-between sizes aren't needed because sometimes they are. However, if they are, we still need to know where and how to create the sizes so they fit people who otherwise would have no fit. We give a method you can systematically apply to find the causes of individual problems, understand them and thus, solve them. We show worked examples of different design and sizing solutions in the subsequent chapters. For example, if the problem is missing sizes, then the correct answer is more sizes, or if the problem is overlapping or duplicate sizes then the correct answer is fewer sizes. Often, the solution is the same number of sizes but with the base size and grading shifted slightly to fit more customers optimally. All problems are not the same and equally, all solutions are not the same.

After decades of research and development experience, we have learned the secret to successful fit and sizing is to measure, validate, and document early and throughout the development of the product, *who* you fit, *how well you fit them*, and *how they relate to your intended wearer population*. With this information, it is possible to make informed decisions about the design, the adjustment mechanisms

(such as the type and number of pads, elastic straps, lacing, and belts), the number and assortment of sizes, and more, to best accommodate your target market with the least amount of sizes and cost.

A word about standards. A universal sizing standard should not be confused with a universal good outcome for consumers. All brands should aim to have a universal good outcome for their shoppers, which means that clothing fits their intended target customer. We call the target customer or user the *Target Population* or TP. A universal sizing standard would mean that all garments labeled, say size 12, would fit the same size people. So, to use Emma's example, if a shopper went to Top Shop and Saint Laurent then they could walk in and both fit a size 12 in each shop. This makes sense if the target customer is the same for both brands. However, groups of target customers differ in products and geography, and therefore a universal sizing standard should be optional. On the other hand, internal fit standards as part of a quality control suite inside each company ensure each new style fits their target customer and are completely essential!

For designers and engineers of completely new wearable innovations, this book helps ensure their first product on the market will effectively fit their TP with less wasted sizes or design components. It enables them to scale the business to other regions or demographics, avoiding failures due to poor adaptation to the new target market. Poor fitting wearable technology not only leads to wasted cost in unnecessary sizes, or lost sales due to poor accommodation of portions or whole demographics groups, but also ultimately, poorly fitted wearable technology could lead to loss of productivity, errors, or safety-critical failures.

For products that are already on the market, this book helps future iterations minimize waste and maximize the percentage of the market accommodated. It explains how to develop a company standard that is maintainable for new styles, new seasons, or new product iterations. We refer to this as the *sustainable fit standard*.

For organizations that purchase large quantities of uniform or standard equipment and apparel, such as hospitals and firefighters, this provides an effective method for evaluating the suitability of wearable products for their population, ensuring better fitting products and minimizing workplace injuries caused by poor design or sizing.

SUSTAINABLE PRODUCT EVALUATION, ENGINEERING, AND DESIGN PROCESS

The Sustainable Product Evaluation, Engineering, and Design (SPEED) process is illustrated in Figure 1.1. There are three main sections: (1) inputs, (2) the design loop, and (3) the sizing loop. The inputs section is shown at the upper left of the diagram, and the design and sizing loops are shown in the center, with the design loop being the outer loop and the sizing loop the inner loop. We look at this process when we are starting with a general idea about what product we want to design on the outer loop, then through iterative testing, narrowing down our options and refining the design until we arrive at the final product in all its sizes and configurations.

In Figure 1.1, we present the entire process and the basic order in which the steps occur as if we are starting from the beginning of a product design. If we are working with a new product or product type, we might need to start this process at the beginning, establishing requirements, the concept of how it should fit, the TP, etc. However,

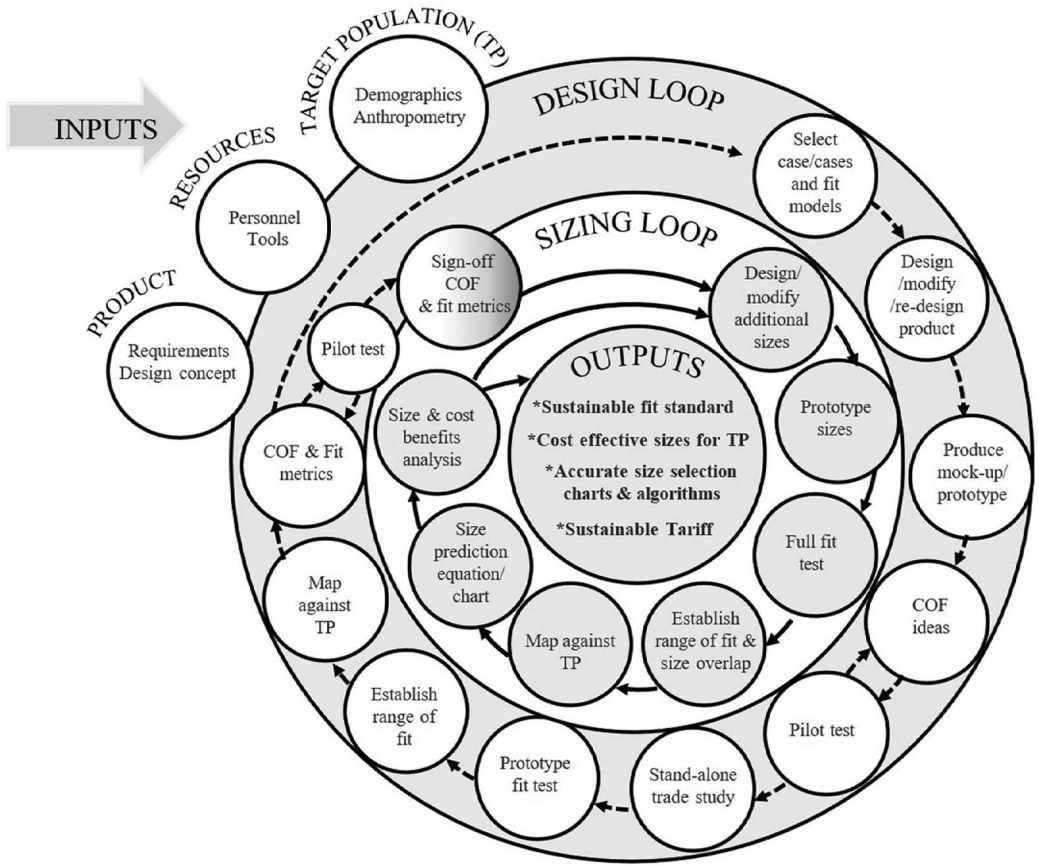


FIGURE 1.1 Sustainable Product Evaluation, Engineering, and Design (SPEED) process.

it is common to use the process at different points, depending on the maturity of the product and for companies with existing products, or for organizations who are simply purchasing products for their workforce it may only be necessary to use parts of the process. For example, if we are working with an established company that already has existing products and a fit model representing the base size some of the first steps have already been done. If they have never done a prototype fit test, we might start the process by evaluating their fit concepts [we call this the Concept-of-Fit (COF)], using a pilot test and then, when the procedures are finalized, begin a fit audit.

Pilot tests are pre-experiment guidance tests that evaluate our test procedures before we begin testing. We include pilot testing at two points in our process; once early in the design loop and once between the design loop and the sizing loop. The former ensures we are on the right track as we begin, and the latter is to ensure something hasn't changed as the product evolved.

The COF is a description of how the product should fit and how to evaluate the fit. The initial COF is usually a part of the design concept and is often not documented but is instead just an idea in the mind of the designer. The finalization of the COF requires input from the manufacturer and any stakeholders to ensure that some important aspect of fit is not overlooked or misrepresented and to ensure everyone agrees. This is best done by having all stakeholders (such as designers, buyers, quality assurance people, patternmakers, human factors people, user-interface people, and senior management) observe fit assessment against the COF on test subjects and have them sign off on it when possible. This ensures there is no misunderstanding to be discovered at the end when it is too late. This fit assessment is usually one part of the pilot test.

We have a more detailed discussion of the COF in Chapter 2, where we discuss what makes a good COF and what to avoid. We also have COF examples in Chapters 5 through 7.

A *fit audit* is an assessment of an existing product to determine if the sizing is suitable as is or needs to change to accommodate the TP. This will tell us if we can drop some sizes, add sizes, or adapt our size range to accommodate more people in the same number of sizes.

We usually like to begin each fit audit with a pilot test to ensure everyone is clear and on the same page about what constitutes a good fit, and to verify that our measurements and questionnaires are capturing what we need. While the COF may seem good on paper, when we see it applied to people with varying body sizes, we often find we need to make changes or refinements. The audit might then go through a design loop with prototype fit testing or skip the design loop and go to the sizing loop. This depends on our confidence in the design and the reason we are doing the audit. For example, if the organization evaluating the fit is not the manufacturer and will not have the option of changing the design there is no need to go through a design loop. The sizing loop will identify design flaws as well as indicate which sizes to obtain or purchase or even indicate if a product is worth purchasing at all.

For some mature products, the manufacturer might have a fit standard. A *fit standard* is part of a suite of quality control and quality assurance standards. Its function is to ensure consistency in fit and sizing between styles within a brand or company. It includes such things as body measurements, blocks/slopers, manikins, pattern measurements, human models, grading, and CAD models. Sustainable fit standards

usually build on existing fit standards. A sustainable fit standard has been verified with and flows from sizing loop fit studies. This enables the company to create excellent tariffs and sizing prediction tools to match each product with each market for consistently good sizing and inventory optimization outcomes.

When there is an existing fit standard, the fit audit might go through the sizing loop to either verify the fit standard is sustainable or to arrive at a new more sustainable fit standard. If there is a need for a new standard, the next step for improving the sizing is to select a case that meets the new standard and begin a design loop. In Chapter 5, we have an example of an apparel item, a woman's pant, that started the process with COF ideas, went through a few pilot tests and a prototype fit test to improve the product, conducted a full fit test to optimize the sizing, and finally arrived at a sustainable fit standard for follow-on styles and seasons.

For new designs or items, some parts of the design loop are typically repeated many times as the design matures. For example, trade studies are done for every design change that might impact fit, such as new components or materials. Sometimes these are done as part of a prototype fit test and sometimes they are done as stand-alone trade studies. When a trade study test indicates the need for a design change, the process might skip back to the "modify product" step after the test. If we know there is a need for a design change, there is no need to establish the range of fit or map against the TP until the change is made and tested. We move to establish the range of fit once we are confident in our design. If components or materials we had planned to use are determined to be too expensive or no longer available, we may have the need for a trade study after a product has already gone through sizing assessment. In this scenario we might skip right to a quick stand-alone trade study of the candidate materials to select the new one. If the new item does not change the fit, there may be no need to do additional prototype or sizing loop fit tests.

Inputs

By inputs, we refer to all the things that might be needed before starting the design or evaluation. This is the subject of Chapter 2. Inputs are divided into three groups:

- Product
- Resources
- TP

The product group inputs are knowledge about the product requirements, constraints, design concepts, and the initial COF. Things such as per item cost limitations, manufacturability, user requirements, product performance requirements, or key performance indicators (KPIs) are included in this group.

The resources group inputs include data collection tools, personnel and their training, data collection logistics, facilities needed, and database management and maintenance plans. Database management and maintenance plans are often overlooked when an organization initiates testing. It can become a nightmare to try to organize all the many datasets collected over time or to track the decisions made as a result of testing. It is best to plan for collating and grouping data from the start. This not only enables easier searching for data and test results but also helps track design changes and the reasons behind them.

The TP group of inputs includes the demographic and anthropometric description of the TP as well as data resources to be used. Data resources can include both existing samples and plans for collecting new samples from the TP.

Design Loop

The design loop has two parts: (1) case selection with prototyping and (2) testing. It begins with case selection. The Human Factors and Ergonomics Society's publication "Guidelines for Using Anthropometric Data in Product Design" (Dainoff et al., 2004) introduces the concept of cases. A case is a single individual to be represented in a product design or evaluation. This representation can take three forms: (1) a list of measurements of an individual, (2) a 3-D (or 4-D) model of an individual, or (3) the actual individual. The actual individual is called a fit model or a live model. Often all three representations are used.

Chapter 3 covers the first three steps in the design loop and presents comprehensive details for selecting, representing, and using cases. This includes:

- Identifying base size cases with raw data or aggregate data
- How to select key variables
- How to use bivariate analysis and plots
- How to do and understand a PCA analysis
- How to compute and use z-scores
- Selecting additional cases and grading
- How to do and use multivariate regression
- Representing cases with physical or digital models
- Using cases

The rest of the design loop consists of testing to inform the design. We use a loop rather than a straight line because we continue to iterate the design until we are satisfied with its fit and performance. As we learn about the product, and what works and what doesn't for our TP, we modify and improve our design.

There are three types of design loop tests: (1) pilot tests, (2) trade studies, and (3) prototype fit tests. We present them in this order because it is the order of simplest and easiest to most complex and challenging. This is not necessarily the order in which they need to be done. Tests are done when a decision is needed, and it is common for unscheduled decisions to be needed. In other words, the complete design loop is not necessarily fully completed before a design modification is made and a new design loop iteration started. If the first pilot test reveals the need to make immediate changes or adjustments, the remaining steps in the design loop would be skipped and the design loop would restart beginning with modification or redesign of the product or procedure. This is true for subsequent steps as well. If changes are determined necessary at the conclusion of a step in the design loop some of the follow-on steps might be skipped until the change is made and the loop starts again with the modified mock-up or prototype. The design loop should also start again if there are changes for reasons other than fit and performance issues, such as manufacturability or material cost if these things are believed to affect fit.

Trade studies are simple comparisons of a small number of treatment options, such as paddings with different firmness, or two different types of adjustment mechanisms,

for the purpose of narrowing down design options. A trade study can be done as a stand-alone test or can be done as part of a larger prototype fit test. Doing a stand-alone trade study can enable us to drop some of the options before doing a larger more complex test.

Prototype fit tests incorporate some of the same tests as trade studies and more. While simple trade studies of some components may not depend on fit, prototype systems do. That means body size and shape are necessary factors, the analysis is more complex, and usually, there are many different analyses needed corresponding to the different aspects of fit according to the COF. Prototype fit tests determine how well the product is performing for all types of people in our TP. This will usually require more subjects than we need in the trade studies testing.

If there are fit issues, we analyze them and determine how to resolve them, either with design changes or sizing changes. If the changes needed will substantially modify the system, it may be necessary to redesign or modify before examining the range of fit in the first size or sizes. If the first size or sizes seem to fit some people well, then we examine who is fit and who is not, to understand the range of fit in the first size or sizes. If the first size or sizes need to be adjusted, then these changes will be made, and the new size or sizes evaluated again. If, on the other hand, there are no major issues that need addressing except for finalizing sizes, then the wearable is ready to move to the sizing loop.

When the testing reveals that the design concept is feasible, it is time to analyze the range of fit within a size and use that range to map the size against the TP sample. This is called *fit mapping*. Fit mapping reveals if the sizes are well-placed and if there is a need for additional sizes or adjustments. If the first size falls at a spot that is off from the main part of the population it is sometimes necessary to modify the product in a major way and re-start the design loop. If it appears to be in a reasonable location, then the process moves to the finalization of the COF and the creation of any additional sizes or adjustments.

Chapter 4 has step-by-step procedures, experimental design, and statistical analysis methods for all testing. This includes the testing in the design loop and testing in the sizing loop.

Sizing Loop

The sizing loop uses fit testing of the final product in at least one size along with sampling from the TP to determine the best size assortment and adjustability features. The final full fit test is done in the sizing loop with fit mapping against the TP. The purpose of the design loop is to optimize the design. The purpose of the sizing loop is to verify that the sizes and adjustments are effective and to optimize the sizes and adjustment mechanisms. By the time we arrive at the sizing loop, we should have a mature design that we are confident is good. For the sizing loop, we should have a complete and functional product with all components and functional features.

The sizing loop is where fit audits and most size validations are done. The outcomes of the sizing loop are:

- Validated product fit in all sizes
- Size prediction algorithms, charts, and procedures
- Tariffs (how many of each size to produce, purchase, or stock)
- Sustainable fit standards for future styles, or product iterations

A *sustainable fit standard* is a fit calibration framework that is used as part of the quality control process to ensure the fit is maintained for all products of a type. It is essential to ensure a consistently good fit for the TP while minimizing unnecessary sizes and size duplication. It is determined and defined after a fit audit and may be redefined if the TP changes, the company changes its brand identity, or the product is sufficiently different in some way that may require a new fit audit.

If an effective sustainable fit standard has been defined (based on the results of prior testing), the fit to the TP can be maintained with mini-fit tests of the prototypes for new product styles, or versions.

Whenever a new product is developed the developer makes some assumptions about how many and which sizes will be needed. We have found that this assumption is almost always flawed, even for products that have existed for a long time. For example, Robinette and Veitch (2016) illustrate how the apparel industry's most used grade could be adjusted very slightly and would accommodate 15% more of the TP, which represents more than a 40% improvement in sales opportunities. The sizing loop not only tests the assumptions but also permits the developer to make informed decisions about whom to fit, whom to risk not fitting, and how to adjust the product to improve it. This might be an improvement for just one product, or it could be documented in a new sustainable fit standard for multiple products.

After the design loops are complete and the COF is finalized, the sizing loop begins with the creation of any additional sizes deemed necessary from the design loop testing. Next, a full fit test is done. This test will typically have the greatest number of subjects and sufficient representation from each demographic category in the TP.

The full fit test results, particularly the range of body measurements that fit for each size, are mapped against a body measurement sample from the TP to create the fit map. The fit map is used to determine the percentages accommodated in each size, the degree of overlap in the sizes, and the degree to which the TP is covered with the entire size assortment. It should be noted that the fit test results can be mapped against many different TPs provided the demographic categories for each were tested. For example, if the original TP was estimated to be 80% under age 30 and 20% over age 30 but it was later learned the TP would be 50% under age 30 and 50% over age 30 the results can be re-weighted to represent the new percentages. If the original population was from the United States, but the product will now be sold in Australia the fit results can be mapped against an Australian sample.

Fit mapping enables risk assessments for dropping sizes, spreading sizes further apart, and adding sizes. In other words, it allows informed assessment of which users will not be able to obtain a good fit if a size is omitted and thus whose business may be lost. This can be compared to the cost of producing the size.

Fit mapping test results and TP maps are also used to create size selection charts and prediction equations that will assist the user in obtaining the correct size in an easy-to-understand manner. This makes it easier for marketing to communicate sizing more accurately to both new and existing customers and improve customer confidence for internet sales.

Experimental designs, sampling and test procedures for the sizing loop testing are provided in Chapter 4. This includes how to map against the TP, how to create size selection charts, and how to create size prediction equations.

An example of the use of the sizing loop and the amount of improvement that might be expected was illustrated in the Navy Women's Uniform study discussed

above (Mellian et al., 1991; Robinette et al., 1991). Dr. Robinette and Ms. Mellian were only asked to validate the "improvement" in fit with the addition of sizes. However, they did more. They did a full fit audit of the sizing. This assessed who was accommodated, who was not, and how that mapped against the target wearer population, the Navy female recruits.

Robinette and Mellian measured each of the women (anthropometry) and recorded their demographics in addition to assessing fit. From this information, they determined who was accommodated in the sizes, who was not accommodated, which sizes were unnecessary, and what sizes were needed that were not available. Based on the fit audit they were able to understand how to resolve the fit issues and accommodate 99% of the recruits, a huge improvement from the baseline of fitting 25% of the recruits, without increasing the number of sizes. The product of the fit audit was a new sustainable fit standard, or sizing standard, for pants and skirts.

This analysis revealed two things: (1) there was no sustained fit standard and (2) there were two body shapes that the existing Navy sizes did not accommodate in any of the pant-and-skirt styles. We determined there was no sustainable fit standard since the women fit in as many as four different sizes in the four pant-and-skirt styles.

Pattern analysis revealed that the sizes 10 and 11 had identical patterns. This was also true for sizes 12 and 13 and similarly the size 14. When questioned, the manufacturer said that the client (the Navy) had asked for more "in-between sizes" to help the problem of poor fit. However, no spec, guidance, or instructions on how to create the extra sizes was provided. The manufacturer compared the size 10 and 12 patterns. The pattern alteration to create an in-between size 11 would have been so small that it was less than the sewing tolerance for each size, so they decided to exactly replicate the size 10 pattern and call it size 11. The customer (the Navy) asked no questions and thus, the manufacturer didn't say how they added the in-between sizes. As a result, the first time this duplication came to light was during the fit audit. In summary, there were exact duplicate sizes in the size range with different size labels to fulfill their client's instructions of creating in-between sizes. Clearly, this attempted solution didn't improve the number of people who achieved a fit and instead led to both confusion and the additional cost for the client of doubling the inventory held.

In addition, the size 10/11 in the blue skirt was equivalent to the size 12/13 in the white skirt. For example, subject number 395 achieved her best fit in the size 11 for the white skirt, but size 13 for the blue skirt. Her best fit size was size 12 for the white pants, but size 10 for the blue pants. If we consider the duplicate sizes this becomes size 10/11 for the white skirt, size 12/13 for the blue skirt, size 12/13 for the white pants, and size 10/11 for the blue pants. The color of the item is based on the uniform of the day. They would wear a white or blue uniform but would never wear part white and part blue. If we look at the sizes by uniform, we see the result in Table 1.1. Even considering the

TABLE 1.1
Subject 395's Sizes by Uniform and Item Type

Subject 395	Pants	Skirts
White Uniform	12/13	10/11
Blue Uniform	10/11	12/13

TABLE 1.2
Subjects with Different Body Shapes

Variable	Subject 2924	Subject 1838	Difference
Stature (cm)	174	174	0
Waist Circumference (cm)	75	74	1
Hip Circumference (cm)	117	102	15
Bra Size	36B	36B	0
Age (years)	35	32	3

duplicate sizes she wore two different size lower body garments in each uniform. This means the base size starting point for the grade was different depending on the style. In other words, they did not have a standard, or they did not maintain one if they had one.

The fit audit revealed that the two body shapes not accommodated were: women with the same waist but larger hips, and women with the same waist but smaller hips. Subjects 2924 and 1838 shown in Table 1.2 and Figure 1.2 exemplify these shapes. Subject 2924 represents the larger hip subjects who were not accommodated in the

FIGURE 1.2 Different shapes need different size ranges.

TABLE 1.3
Navy Women's Uniform Sizes Before and After Fit Audit

Pants and Skirts Sizing	Before Fit Audit	After Fit Audit				Legend
	Size	Size				
			4 M		J	Junior (narrow hip vs waist)
	6		6 M	6W	M	Misses (middle hip vs waist)
	7				W	Women (large hip vs waist)
	8		8M	8W	XS	Extra Short
	9				S	Short
	10	10J	10M	10W	R	Regular Length
	11				L	Long
	12	12J	12M	12W	P	Petite
	13				T	Tall
	14	14J	14M	14W		
	15					
	16	16J	16M	16W		
	17					
	18		18M	18W		
	19					
	20		20M			
	Lengths	Lengths				
	XS					
	S		P			
	R		R			
	L		T			
Total Number	15 * 4 = 60	20 * 3 = 60				
Percentage Fit	25%	99%				

original size range and subject 1838 represents the original body proportioning. These two women need the same waist size in a pant but different sizes for the hips.

The larger hip/waist ratio body shape is sometimes referred to as a "pear" shape or "curvy". The pear-shaped woman has substantial problems finding close-fitting pants, skirts, and dresses that fit in the commercial market to this day.

This fit issue was resolved by adding two new size ranges representing the two shapes that were not accommodated and dropping the duplicate sizes the Navy had added. The Navy felt the women would not like sizes called plus hip or minus hip, so it was decided to call the large hip sizes "women's" sizes (W) and the smaller hip sizes "junior's" (J). The sizes before and after the study are shown in Table 1.3.

The new sizing, which had the same number of sizes that the Navy had proposed, was tested with additional subjects (a process called fit validation), and verified to have achieved a fit for 99% of new female recruits without the need for major alterations. This is nearly a fourfold improvement over the original fit and illustrates how the SPEED process can reduce waste (drop unnecessary sizes)

while at the same time increasing the accommodation and satisfaction of the TP. This represented a cost saving for the Navy of several million dollars as well as streamlining the issuing of new uniforms to recruits, optimizing inventory and many other benefits.

You might think that this type of situation hardly ever happens, but it is surprisingly common considering the client's (in this case the Navy's) core business is not uniforms and often clients rely on the manufacturers' expertise. The manufacturers know they are not solving the problem but simultaneously do not have the right information to effectively offer alternative solutions. Specifically, they are missing information and expertise in fitting their clients' TP and have no sustainable fit standard to guide them. This means the manufacturer will not point out the issues and instead slavishly follow instructions to keep their clients happy and their contracts in place. These situations are completely avoided with a sustainable fit standard based on a fit audit.

HOW TO USE THIS BOOK

This book is intended to be a guidebook for all product fit and sizing aspects of the wearable product design as well as a reference book for refreshing our minds about the best practices for different procedures. It is structured to help a newcomer to the fieldwork, through the whole process while also enabling experienced readers to skip around to whatever part is of interest. It is divided into two sections. The first section, Chapters 2 to 4, contains instructions for carrying out the best practices. The second section, Chapters 5 to 7, is a section with industry-specific explanations and examples.

Chapter 2 is all about all the things we need to know or think about before we start the SPEED process. This includes tools, materials, personnel, training, resources, and inputs like requirements and constraints. It includes things to consider before starting that we want to have or keep track of later, such as how the data will be saved and stored. We will also talk about TP sampling and the COF definition which can occur throughout the process, not just before we start. Since we might have a good sample from the TP before we start, and we should have some COF then as well, we include these discussions in Chapter 2. However, it is often the case that a good sample from the TP might be gathered as we do fit testing or when we do a sizing loop fit audit, and the COF is tested and refined at least twice in the process and not just before we start.

Chapter 3 is about selecting and using cases. Cases are numeric, digital, or physical representatives and are essential components to both design and sustainable fit standards. Numeric cases are things such as a list of body measurements in a specification. Digital cases can be digital models or copies of people, body segments, or manikins. Physical representatives can be live models, or physical manikins or body forms. For some products, all three types are used, and it is very important that they be well selected, evaluated, and represented.

Chapter 4 contains most of the testing and analysis procedures for both the design and sizing loops. It has a general overview but then breaks down the experimental design and analysis by type of test, and these are organized into design loop and sizing

loop sections. This chapter explains how to do fit mapping, select the sizes, predict the size of best fit, and create a tariff to estimate how many of each size to produce or stock.

The remaining three chapters put the process into the context of three different industries, mass-produced apparel, head and face wearables, and footwear. Each contains a section that discusses industry-specific issues and terminology followed by a section of examples we refer to as case studies that illustrate how the process is applied in that industry.

Chapter 5 discusses mass-produced apparel because mass production has many opportunities to improve fit and reduce waste. It begins by describing industry practices and how they can be improved with the SPEED process. This is followed by case studies illustrating some of the ways the SPEED process is implemented. Examples include a fit audit of an existing pant, the use of 3D scanning overlays to assess fit, evaluation of body armor before purchasing, and the use of a prototype test to fix a grading issue.

Chapter 6 is focused on wearables for the head and face, such as helmets, XR headsets, and head worn PPE. Small differences in people can have a large impact on fit and performance in this area, and products often need to follow the contours of the head and face closely, therefore accommodation of head and/or face size and shape variance could be extra challenging. These products often have complex, expensive and/or time-consuming production requirements, resulting in more time spent on the initial design from digital cases, and less fit iterations. Initial fit iterations are also more often done using non-functional or representative (not final) product materials. This chapter describes the extra challenges for this region of the body with several examples illustrating how some companies used the SPEED process to overcome them.

Chapter 7 is all about footwear. The foot is perhaps the most complex body segment to accommodate in a product. Each foot has 26 bones, 33 joints, 107 ligaments, and 19 muscles. This does not even include the bones in the ankle and leg that connect to the foot. In addition, the feet support the body weight and transfer it to the ground and are important for controlling body movement with efficiency and stability. The footwear industry also has some standard tools and practices that have been used effectively in the past so they are important to maintain. This chapter reviews the current practices and explains how the SPEED process can be used to improve upon them. Then it provides two case studies illustrating how to implement the improvements for casual and fashion footwear, and for innovative footwear products.

Whether your interest in the subject is driven by an economic lens, optimizing the safety, performance, or aesthetic of the wearable product, this book aims to provide you with the current best practices through the life cycle process. You should come away with “how to” knowledge of the SPEED process, which includes concepts such as defining the product inputs, selecting a case/fit model(s), drafting the COF for the product, conducting fit studies of various types, mapping the range of sizes against the TP and providing a size range among other outcomes. The examples discussed in this chapter make for a compelling motivation to apply this process effectively. We have drawn upon decades of experience, research, and applied knowledge to share with you. Regardless of your role in parts, or all of the SPEED process, we hope it will be a fascinating and satisfying experience that will help you to learn, apply, and contribute to the improved engineering, design, and evaluation of wearable products.

REFERENCES

Dainoff, M., Gordon, C., Robinette, K. M., & Strauss, M. (2004). *Guidelines for Using Anthropometric Data in Product Design*. Human Factors and Ergonomics Society.

EPA. (2020, August). *Facts and Figures About Materials Waste and Recycling/Textiles: Material-Specific Data* [Official Website of the United States Government]. https://www.epa.gov/facts-and-figures-about-materials-waste-and-recycling/textiles-material-specific-data

Liu, K. (2018, September). *Undressing Waste's Issues*. https://www.redress.com.hk/events/2018/9/27/talk-undressing-wastes-issues

Mellian, S. A., Ervin, C., & Robinette, K. M. (1991). *Sizing Evaluation of Navy Women's Uniforms* (Technical Report AL-TR-1991-0116). Air Force Systems Command. https://apps.dtic.mil/sti/citations/ADA249782

Robinette, K. M., Mellian, S. A., & Ervin, C. A. (1991). *Development of Sizing Systems for Navy Women's Uniforms* (Technical Report AL-TR-1991-0117; Issue AL-TR-1991-0117). Armstrong Laboratory, Air Force Systems Command. https://apps.dtic.mil/docs/citations/ADA250071

Robinette, K. M., & Veitch, D. (2016). Sustainable Sizing. *Human Factors*, *58*(5), 657–664. https://doi.org/10.1177/0018720816649091

Spedding, E. (2019, May). *Editors, Designers and Model Agents on the Fashion Industry's Sizing Problem*. https://www.whowhatwear.com/fashion-industry-sizing-problem

2 Inputs and Getting Started

Kathleen M. Robinette

ABSTRACT

This chapter discusses the things to know, prepare, obtain, document, and assess before starting product design or evaluation. It serves as a checklist or a reminder with handy tips for the more experienced person and as a detailed guide for the novice. It is divided into three sections that match the inputs in the Sustainable Product Evaluation, Engineering, and Design (SPEED) process: (1) product, (2) resources, and (3) target population (TP). The product section discusses what we need to know about the product such as the product concept, how it is supposed to function, how it is expected to fit, and any constraints. The resources section discusses the tools needed, the personnel needed and their training, the facilities and test site considerations, as well as planning for data collection, data management, and maintenance. The TP section discusses how to represent and characterize the TP before, during, and after data collection.

Testing and gathering data on human subjects is a large part of the SPEED process enabling informed decisions throughout the process. Testing tells us what works and what does not, what is good and what is better, and what the users like and what they do not like. If we know these things early, it is less expensive to make changes than if we wait until the design is complete. If we wait until we are evaluating sizing for the finished product, it could be too late. Careful preparation allows us to avoid wasting time and money while doing this. Or as Benjamin Franklin once said, "A stitch in time saves nine".

A large part of the preparation involves minimizing, controlling, and accounting for variability we do not want (error) such that with the help of probability theory and statistics, we can make good decisions. Whenever measurements are performed, no matter how carefully or scientifically it is performed or the quality of the measuring instrument, measurements are always susceptible to error and uncertainty. This is true even for measuring things that don't live, breathe, and change all the time. In the Metrology and Engineering fields, this is called Measurement uncertainty. *Measurement uncertainty* is an estimate of the level of accuracy and precision with which a measurement can be taken with a given tool or process.

Measurement uncertainty is one type of error. There are many sources of error including the tools used and their precision, the measurers and their training, environmental conditions, changes in the person being measured, measurement location, type of measurement, measurement method, clarity of measurement description, the sample size relative to the population size, and more. Unfortunately, these sources do not cancel each other out, but the error expands with each additional source. Therefore, it is important to do things to minimize or at least manage

DOI: 10.1201/9781003397533-2

errors from each source. This begins with good preparation, which is a large part of this chapter.

PRODUCTS

Before we start the design and/or testing, there are two categories of things we need regarding the product: (1) requirements and constraints for the product and (2) the design concept including the Concept-of-Fit (COF). This includes describing the product, what it must do, who will use or wear it [target population (TP)], and any limitations in time, money, or other factors. As the product is tested, new requirements will reveal themselves and some of the pre-set requirements will change. However, through the course of the design process, having clear and documented requirements, in the beginning, helps track changes and understand the reasons for the changes. Poorly defined or unclear product goals can also result in wasted time and money. It is much more difficult to hit a fuzzy, moving target than a clear, stationary one. As Charles F. Kettering once said, "A problem well stated is a problem half-solved".

Requirements and Constraints

The best requirements have input from all the different types of stakeholders. A stakeholder, in this instance, is someone who has an interest or concern in the product, its fit, and its function. This can include people from senior management, product development, sales, marketing, engineering, design, legal department, and potential customers. Getting their input up front can avoid problems later and save time getting the product to market.

Established companies have probably been through the process of documenting and reviewing requirements for previous products. They may know their customers well and have a good idea about the requirements for the new product. Therefore, this step may be easy for them.

However, new start-up companies seem to have difficulty with this part. They frequently create a prototype before understanding and defining the requirements and the TP. It is reported that more than 90% of new start-ups fail and the number one reason new start-ups fail is because they built something no one wanted (Garplid, 2013). In other words, they failed to validate the product requirements with potential customers and stakeholders. This emphasizes the importance of not only clear but also validated requirements. Before defining requirements, it is important to know what your customers want or need badly enough that they will spend money on it, and it is also important to know what is achievable for a cost that will enable the business to make enough money to survive.

For example, mass-production apparel manufacturers want a product that sells well, usually by making the wearer look good while allowing movement and keeping costs to a minimum. This is not a product that perfectly matches the size and shape of the body. They also have some strong opinions about what looks good that may or may not conform to what their customers think looks good. Whether we agree or disagree with these opinions, it is good to know something about them before we define our requirements.

To avoid making something that no one wants and being one of the 90% of start-ups that fail, a good place to start is with some of the questions from the Heilmeier catechism (Heilmeier, 2021) and answer the following:

- What are you trying to do?
- How is it done today, and what are the limits of current practice?
- What is new in your approach and why do you think it will be successful?
- Who cares? If you are successful, what difference will it make?

To answer these questions, we must start with a literature review or a review of what has been done before. Then to answer the remaining questions we need to interview people who have done something similar and people who will care. The latter are people from groups or organizations you believe will want the product. Kim Goodwin's book on design, even though focused on software, has extensive detail regarding communicating with potential customers and users and establishing requirements that apply to any kind of product for human use (Goodwin, 2009).

The purpose of interviewing is to understand needs and wants, that is, requirements. It is not about learning if they want your concept of a proposed product. In fact, it is best if only a proposed solution is mentioned, not the detail of the proposed product because this will bias their responses. The questions should be about learning what is being done now that the product will improve upon, what issues they are having, and what kinds of things they believe are needed. For example, imagine you are planning to build a wearable, vital signs monitoring device. You may have in mind a vest, an arm band, or a head band, but it is best if the person being interviewed does not know your plan. Instead, tell them you are doing research on vital signs monitoring and want to ask them about their experiences with it. Some of the questions you might ask include following:

- What types of devices do they use?
- How happy are they with them?
- What issues do they have with them?
- If they could improve them, what would they do?
- How close the devices need to be to a particular point on the body to give accurate readings?
- If they could have any product they don't currently have, what would it be?

It is also important to interview other stakeholders such as the people who might fund the development, company managers, people from marketing departments, people from purchasing department, and engineers who understand what is involved in producing such a product. If, for example, the product you plan to build is deemed by the engineers to only be possible by using very expensive components that make it unaffordable, it is best to find that out in the beginning, so you can try to find a less expensive but doable option. Interviewing possible stakeholders has an added benefit. Not only will you learn a lot about what the most important characteristics of the wearable need to be but also you will find funding sources and potential collaborators in the process.

One of the best programs for learning how to do this is the NSF Innovation Corps (I-Corps™). It was developed to help new start-ups be successful. It is an intense several-week course established to "reduce the time and risk associated with translating promising ideas and technologies from the laboratory to the marketplace" (National Science Foundation, 2020). It uses customer and industry discovery through 100 or more interviews before deciding on the product and establishing the requirements. We know from personal experience that it is common for new start-ups to completely change their product by the end of the course and then they end up with a viable company for the long term. One of the key resources for this program is the book by Stephen Blank about having a successful start-up (Blank, 2020). Other countries, like Australia, have similar start-up help for self-employment provided by their government (https://whatsnext.dewr.gov.au/try-something-new/starting-small-business).

Key performance indicators (KPIs) are quantifiable measures of product performance and these serve as the list of requirements in the requirements document. They can be things the product is being designed to do, such as monitoring heart rate or protecting from aerosols. Or they can be things the product must be able to do while at the same time doing what it is designed to do, such as being comfortable and stable or enabling the user to reach or see something. A KPI requires some sort of metric that indicates success or failure. For wearable products, some examples include the following:

- Ninety percent of the target users should find it acceptably comfortable after 3 hours of wear
- Will not shift more than 2 cm during exercise
- Product temperature will not exceed 40 degrees
- Lenses will not contact the eyelashes
- Product will not impede full range of arm movement
- No more than 10% blockage of peripheral vision
- Must completely cover all skin without gaps when sitting or standing, or
- Ninety percent of the target users like the way it looks

This last one will make some engineers scowl, but our experience has shown that customers will pay lots of money for something that they think makes them look good even if it is extremely uncomfortable. For example, women wear shoes with extremely high heels that not only hurt their feet but can even permanently damage them. Women (and some men) wore corsets that made it difficult to breathe for centuries. We had a recent example where women bought two pairs of the same pant, one to wear at work where they stood up all day and one in a larger size to wear to drive home. They liked the pant that looked good on them while standing but it was too uncomfortable to sit down in it. This customer called the smaller size their "standing-only pant".

In addition, people will not buy a product, no matter how comfortable or effective, if they do not like the way it looks. We did a study of helmets and had two identical helmets except for the outer color. One was white and one gray. Despite being identical, the subjects insisted the gray one was more comfortable. Sometimes the way the product looks can make or break it.

It is also important to list or know about KPIs that may or may not be fit-related. These types of KPIs can be trade-offs with fit and have an impact on cost versus benefit decisions. Some examples include following:

- Fabric shall not cost more than X amount per meter
- Must be able to replace batteries without tools
- Total cost to build shall not exceed X amount
- Design concept will be a sleeveless vest or helmet-style or not look like a helmet

This is a design process so the design itself changes. The changes will be guided by the requirements. The requirements may also change as we learn what works and what does not. By having a requirements document at the start, we can track changes and the decision process.

The requirements should also describe who the intended or expected wearers will be. We refer to this as the TP. The TP definition will be used to guide the selection of subjects for design and testing, as well as size selection and cost/benefit analysis. A good definition of the TP is important to ensure that the product will accommodate the most users in the fewest number of sizes. When we get to the cost/benefit analysis, we will need a good sample drawn from the TP to make informed choices regarding which and how many of each size to produce (Robinette & Veitch, 2018).

The TP definition should describe the demographics of the expected users. The purpose of demographics is to ensure that the sizes and shapes of the relevant groups of people are adequately represented and to ensure that time and money are not wasted on groups who will not be users. Demographics generally include things such as gender, age group(s), and geographic user region. It might also include things specific to the product such as people with health conditions, people who participate in a particular sport, or people from an occupational group such as firefighters or software developers. People in different occupations can have very different body proportions.

Let us consider some examples of demographics that impact size and proportions. In Table 2.1, we see the mean stature values from three countries; Italy, the United States, and The Netherlands; calculated from the raw data from the CAESAR™

TABLE 2.1
Stature Statistics from Three Countries

				Differences (mm)		
Stature (mm)	**Italy**	**United States**	**Netherlands (NL)**	**IT-USA**	**USA-NL**	**IT-NL**
		Males				
Mean	1736	1777	1815	–41	–38	–79
Std. Dev.	67	80	87	–13	–8	–20
N	413	980	565			
		Females				
Mean	1611	1639	1680	–28	–41	–69
Std. Dev.	62	74	76	–12	–2	–14
N	388	1178	700			

study (Robinette et al., 2002). These were data collected by the same measuring teams using the same tools and methods as part of one data collection series in the time frame from 1998 to 2000, so any differences are not due to measurers, tools, or time frames. They reflect differences in the populations of the three countries.

The means from the different countries are all significantly different at $\alpha = .01$. As can be seen, even though Italy and The Netherlands are both European countries, the Dutch have much larger statures than the Italians, 79 mm and 69 mm difference for males and females, respectively. Therefore, making a pant for the Italian population, it is likely to be too short for the Dutch population. The means for the United States fall between the Dutch and Italian means, however, they are still significantly different. This illustrates just how important the definition of the geographic region for the user population can be.

In their article on firefighter anthropometry Hsiao et al. (2014) provide us with a good example of occupational differences. They estimated that "On average, male firefighters were 9.8 kg heavier and female firefighters were 29 mm taller than their counterparts in the general U.S. population". They further noted that "They also have larger upper body builds than those of the general U.S. population". The fact that male firefighters are not substantially taller than the general U.S. population, but they are substantially larger in the upper body is an indication that there are proportional differences. Therefore, to get the proportioning right for firefighting apparel, we should use firefighters for our design cases and fit models.

Hsiao et al. (2021) demonstrated there are large differences between law enforcement officers (LEOs) and the general U.S. population. In this article, they illustrate the difference using a three-dimensional (3D) scan of a case that is near the mean for male LEOs compared to a 3D scan of a case that is near the mean for the general male population of the United States. This allows us to see not only the size but also the shape differences.

Once we have a definition of our TP in our requirements document, we must gather samples from it or have a plan for how and when we will get one. In the design and sizing loops, we will represent the TP using cases and samples taken from it or a similar population.

Design Concept

A design concept must clarify where on the body the product is intended to be worn, how it is expected to function, how it is expected to look, any other items with which it must interface, and what are the expectations regarding what will constitute a fit. If this is a new kind of product, it is reasonable to expect that some of these concepts will change during development, but if we document our starting concept, then we can track the changes along with the reasons for them. When we have a record, we can keep on a steady course even with changes to the design and engineering team, and it permits reconsideration of design trade-offs as the product evolves.

The concept of what constitutes a good fit versus a poor fit and methods for assessing or measuring it is referred to as the COF. It is important because it defines the priorities while integrating and balancing the look, function, form, and fit. Sometimes features like the look and comfort become trade-offs, so the COF helps prioritize one over the other. Thousands of decisions go into product development and the COF

helps the product development team understand from the outset where the areas of focus are and where compromises are needed. There are three parts to a COF:

1. Fit requirements
2. Tools or metrics to assess the requirements
3. Level of acceptability or unacceptability (pass or fail)

Fit criteria will differ depending on the item being assessed and who is assessing. So, the COF for a fashion item will look very different from an item of personal protective equipment (PPE). Judgment of fit depends on the assessor's point of view and priorities. The wearer judges fit on themselves, the company judges fit through the lens of the brand identity, and PPE is judged largely on its performance. The COF should consider everyone's opinion that matters.

Table 2.2 illustrates one example of a COF for a head wearable. In this example there are three fit categories: (1) location, (2) comfort, and (3) stability. There is a list of tools or metrics to be used to assess the requirements and there are pass or fail criteria for each. This example comes from early in the design process when the mock-up was not functional. It was a simple 3D printed shape with the size, shape, weight, and fitting properties of the design concept. Once a functional prototype was available, the COF changed to replace location measurements with functional assessment of the person's ability to see and use the wearable as needed. The location pass/fail criteria were approximations for the mock-up, but they were sufficient to enable the refinement of the form factor, weight distribution, size, shape, and fitting mechanisms without having to produce expensive fully functioning prototypes. This saved money and time.

A good COF must state how the quality of the fit will be measured. This might be done using a questionnaire and/or objective fit measurements such as measuring parts per million (ppm) leakage in a mask. Regardless of the method used to measure and score, the pass/fail value should not be arbitrary. In other words, do not set a pass/fail value in advance if we do not know the value in advance. Instead, have the investigator conducting the test make a pass/fail judgment.

This may seem like common sense but setting arbitrary pass/fail criteria is a surprisingly common problem for fit testing. Engineers are particularly uncomfortable with opinions. They want a number. However, if we do not know the number, then an arbitrary number can be horribly wrong. For example, in the fit test of a flight helmet, the manufacturer had specified that the pupil of the subject had to be exactly 2 inches below the edge roll (the part of the helmet above the eyes in the front). When the fit test began, everyone failed. No one could get the helmet placed in that location. Furthermore, the pilots did not want the helmet to be placed in that location because it limited their ability to look up, a maneuver that could save their life. It was determined that the helmet fit well as it was and needed no changes. It was the pass/fail fit criteria that were wrong.

Another example was demonstrated by a law enforcement organization that was testing a ballistic vest to determine which vest sizes to buy. They specified that the bottom of the vest should be between 0 and 1 inch above the navel to fit. When the vest was pilot-tested, few LEOs wore it that way. They preferred it to rest on their utility belt where the weight was supported and where they would have greater

TABLE 2.2
Example Concept-of-Fit for Non-Functioning Mock-Up

Concept-of-Fit for Non-Functioning Mock-Up of the Head Wearable

There are three fit categories: location, comfort, and stability. ***All three must simultaneously pass fit to be an overall pass.***

Location Location-fit will be measured using 3D scanning and measuring the eye location and head horizontal orientation with respect to the display. The process is as follows:

1. The subject will be scanned without the wearable with the subject looking at his or herself in the mirror to establish the natural gaze plane (scan 1). Pupil landmarks will be identified in the scan.
2. Subject will be scanned with the head wearable in place (scan 2) and landmarks will be identified indicating display location and orientation.
3. The head wearable will be 3D scanned (scan 3). Landmarks will be identified in the scan indicating display location and orientation.
4. Scan 2 will be registered to scan 1 in the software tool by matching the visible areas of the subject's head and face.
5. Scan 3 will be registered to scan 2 in the software tool by matching the visible areas of the wearable.
6. Using the software tool, measure the distances between the scan 1 pupils and the scan 3 display.
   a. Shortest straight-line distance
   b. Distance horizontal to the display orientation
7. Calculate the angle between the straight-line distance and the horizontal distance.

Pass criteria:

Horizontal and straight-line distances are less than 25 mm.

Angle between horizontal and shortest straight line is greater than 10 degrees.

Comfort Comfort will be assessed using an ordinal scale questionnaire instrument at two points: (1) immediately after the wearable is first fitted (5 minutes) to determine if adjustments are needed and (2) after 30 minutes of wear for the final fit score. The fit fails if any one of the following conditions are true:

Subject would not be willing to wear it for 4 hours or more.

Subject scores the overall fit as poor or fail.

Subject scores it a 1 for pressure in any area (not satisfied at all – pressure is very distracting).

Stability Stability will be assessed using an ordinal scale questionnaire instrument at two points: (1) immediately after the wearable is first fitted (5 minutes) to determine if adjustments are needed and (2) after 30 minutes of wear for the final fit score. The fit fails if the wearable nearly fell or slipped out of position and did not return during normal wear or after roll, pitch, and yaw movements.

coverage and protection. When we asked the organization why they had the 1-inch limitation, we were told that most vital organs were above the navel and they did not want the vest to ride up into the neck when the person sat down. Riding up was not an issue and their theory about fit was flawed. The organization's COF outcomes resulted in less coverage and poorer fit. A better COF was to allow it to rest on the utility belt and ensure it did not interfere with the neck when seated. These kinds of issues are quickly resolved with the COF mini-tests or a pilot test.

In these two examples, the pass/fail criteria were objective measures that did not require the subject's opinion, but they were not good measures of fit. They illustrate that objective measurements are not necessarily better than subjective opinions. Objective measurements can sometimes give a false sense of confidence if the interpretation of the number is based on arbitrary or theoretical pass/fail criteria. If we make the judgment based upon how well the subject said he or she could perform their job or tasks, then we can test our theory about fit, adjust the location, tightness or looseness, and so on angle to match where it does fit best and can use the fit results to improve the design.

Things to consider when drafting a COF include (1) product performance requirements, (2) use environment (occupation, indoor vs outdoor, etc.), and (3) integration and compatibility with other things to be worn or used with it. A procedure for the construction of fit criteria is described in detail in a fit mapping manual by Choi et al. (2009). It involves breaking down the criteria into a list of all the requirements and measurements that should be assessed. Then systematically translating that list into a consistent and measurable form by which fit can be evaluated and quantified. This can be helpful for complex items like protective gear to ensure nothing important is overlooked.

Regardless of how careful we are when we create the COF, it is important to test it in a pilot test. The pilot test helps us find flaws in our scoring, reveals differences of opinions about fit, and helps us standardize the opinions we get from the questionnaire thereby minimizing the error. It also helps to educate new personnel who join the team after the COF has evolved. New people may want to revert to a COF that has been rejected or refined. When the COF and changes have been documented and tested, we have the evidence to show the new people, so we can avoid vicious circles and repetition of mistakes.

RESOURCES

After we have a design concept and a list of requirements and constraints, we can begin to identify and obtain the necessary data collection, analysis, and management tools, as well as the personnel and the facilities. The process of identifying and obtaining resources is to: (1) identify the data we need to gather, (2) evaluate the tools needed to gather and analyze that data, (3) down-select and obtain the appropriate tools, (4) identify the skills needed to gather and analyze the data, (5) recruit or train personnel to have those skills, and (6) secure the facilities needed for conducting the testing.

Tools

Tools are needed for data analysis and management as well as for data collection. We estimate and control for some types of error using probability, statistics, and experimental design. Some of the most common statistics used to estimate error include the standard error (SE), the standard deviation (STD), and the standard error of the mean (SEM). The smaller the error the more likely we will be to be able to find true differences between people, components, treatments, performance, or products. Therefore, it is important to have statistical analysis software available that will permit us to factor out the error and understand our results.

Data analysis tools must enable us to do statistical analysis, extract information from and/or edit 3D or four-dimensional (4D) data files and visualize product-to-subject

relationships. Data management tools must enable us to organize and manage data files from all testing, so they can be effectively searched, compared, and re-analyzed throughout the design process and for new products or iterations. Data collection tools must be affordable and of sufficient quality to provide relevant, accurate, and reliable data.

Analysis and Data Management Tools

Some statistical analysis and management tools are essential for any data collection and there are many software packages available. We use two in this book, Excel™ and SPSS®, but there are other good options as well, including SAS®, STATA®, and Minitab® to name a few.

Excel™ is a spreadsheet software package that is part of the Microsoft 365® suite. We use it mostly to save and manage our data. One advantage is the data from Excel™ can be exported and imported to and from most other software packages, including online questionnaire survey tools such as Qualtrics XM® and Google Forms®. Excel™ has a statistical analysis add-on that is capable of some level of statistical analysis and graphing, as well, although for statistical analysis, it is not as intuitive to use as SPSS®. An example of an Excel™ data spreadsheet is shown in Table 2.3. We also made a raw data sample of adult women from the U.S. population available on Routledge.com in the Support Materials tab for the book. It contains

TABLE 2.3
Example of a Data Set in a Spreadsheet

Subject Number	Gender	Age	Acromial Height Sitting (mm)	Ankle Circ. (mm)	Armscye Circ. (mm)	Bizygomatic Breadth (mm)	Chest Circ. (mm)
26	Male	28	575	245	452	150	1042
57	Male	38	615	262	422	152	950
69	Female	29	590	238	402	137	872
70	Female	36	594	238	390	143	1000
73	Male	32	665	257	488	156	1085
90	Female	40	535	238	343	139	834
98	Male	24	583	239	397	145	862
107	Male	32	613	256	436	156	970
112	Male	33	561	252	476	157	1017
160	Female	26	532	208	350	130	789
169	Male	40	613	259	471	155	1074
171	Male	28	727	294	483	151	1007
185	Female	32	607	227	345	140	812
195	Female	58	526	212	346	132	795
209	Male	41	594	261	452	163	985
214	Female	23	564	222	358	132	885
231	Female	27	581	230	382	138	909
240	Male	38	614	255	410	148	892
244	Female	22	595	229	355	135	820
249	Female	23	537	220	398	140	897
253	Male	36	586	259	439	156	974
270	Female	25	560	236	401	150	1009

anthropometry, demographics, and some fit test variables including size of best fit and pass versus fail fit scores for two types of pants. This data is not a random sample and not intended to be used for design purposes.

In this book, we use SPSS® for most of the statistical analyses, plots, and graphs. This includes frequency plots and graphs, chi-square analysis, correlation analysis, bivariate charts, linear regression, stepwise discriminate analysis, and more. We also use SPSS to tailor and weight our datasets and we explain how in the TP sampling section. SPSS can import and export our Excel™ file datasets, and it has a text file which is a recording of what was done, so we can track and document the analysis. If we plan carefully, we can use SPSS to re-group and re-analyze data or to combine data from multiple tests.

Often when we start fit testing, we are not thinking about whether we will use the data for future products, but these data sets are valuable resources that can be used again and again. Not only will we want to compare the results from this product to the next one but we will be continually adding to an anthropometry resource that can be used to represent our TP in the future. Therefore, it is important to think about how to save the data for future use and combine it with future data collections.

Subject numbers are assigned to give each individual anonymity and protect their privacy. They are also used to track and connect different types of data collected in a test or study. When assigning subject numbers, we need to consider that the same person might be used in testing multiple times or for multiple tests. Therefore, each person should be assigned a number or identifier that is used for all testing. This will avoid counting some people more than once in an analysis inadvertently. Giving each person one and only one unique number requires tracking numbers by some identifier in a separate file, but it allows us to easily follow a person's progress through multiple tests.

The data collected for some tests might also include things that do not fit into a spreadsheet such as photos, videos, or 3D scans. Each subject may have this data in separate files, so it is a good idea to ensure the file names indicate both the subject number and the associated test and test condition. For example, the polygonal mesh 3D scan file for subject number 0007 in the study called "csr" and in the standing pose a was saved as "csr0007a.ply". This file name contains the study name (csr), the subject ID number (0007), the pose (a), and the file format (ply). When there are hundreds or even thousands of photos or scans, the ability to search the filename is essential. It can also be helpful if there is a column or columns in the data spreadsheet that list the file names for any photos or scans taken.

The same subject should retain the same unique identifier if they are measured multiple times. For example, if the subject is a fit model in the fashion industry, then they might be measured once every six months to track any change in their body size as well as trying on multiple garments. It is important to keep track of which repeat is which so they can be evaluated later. This is done by creating file naming conventions. For example, the naming convention might include the following:

- The study or test identifier, e.g., MZ
- The subject's unique identifier or subject number, e.g., 0105
- The date in reverse order, e.g., YYYYMMDDTTTT
- The garment code or name, e.g., P77998
- The size, e.g., size 8

TABLE 2.4
Example Product Naming System for Management of Bra Comparison Data

Study ID	Subject	Date	Garment Code	Size	Scan File Name
MZ	0155	20230317	GenGar Bb000	NA	MZ015520230317Bb000NA
MZ	0155	20230317	GenGar S4564	Small	MZ015520230317S4564SS
MZ	0155	20230317	GenGar S4564	Medium	MZ015520230317S4564SM
MZ	0155	20230317	GenGar N8845	Small	MZ015520230317N8845SS
MZ	0155	20230317	GenGar U001	Medium	MZ0155202303172U001SM
MZ	0155	20230317	GenGar LB44563	S8	MZ015520230317LB44563S8
MZ	1026	20230318	GenGar Bb000	NA	MZ102620230318Bb000NA
MZ	1026	20230318	GenGar S4564	Small	MZ102620230318S4564SS
MZ	1026	20230318	GenGar N8845	Small	MZ102620230318N8845SS
MZ	1026	20230318	GenGar U001	Medium	MZ1026202303182U001SM
MZ	1026	20230318	GenGar LB44563	S8	MZ102620230318LB44563S8

With this naming convention, the name of the scan file becomes a combination of the names, such as MZ015520230317Bb000NA. A separate file is needed to keep track of the names and identifiers. Table 2.4 is an example of an Excel™ sheet used to track the files.

The tests were described in detail in another text document. This included explanations of the coded names as well as the coded responses.

If we use the same variables or measurements in different studies, it is important to use the exact same name and exact same responses. This includes upper case, lower case, hyphens, and abbreviations. If, for example, gender is recorded as M or F or NA in one study but male, female, and NA in another we cannot merge them. M and male would be analyzed as different responses. If we call a variable Hip Circ. in one study, but Hip Circumference in another these would not be automatically merged.

If the variables are different, then we need to give them different names. For example, if we record Comfort on a scale of 1 to 10 in one study and 1 to 5 in another, these could be merged but the analysis would be incorrect. In this situation, we should give the variable a name that corresponds to the difference, such as Comfort10 versus Comfort5.

Usually, companies plan to have future products and/or new iterations of the same product and they will want to compare them. This is particularly true if we want to use a sustainable sizing standard. Therefore, it is important to include a variable in each data set that indicates the version of the product tested. This might require more than one variable if there are multiple configurations and sizes for a single test iteration.

There should also be a document that is associated with each configuration that describes it. This record should include both a written description of the product with its components and physical measurements. It can also include patterns, CAD files, and 3D scans. If there will be upgrades or new versions of the product in the future this information will prove invaluable. Anything that may affect fit should be described. This can include materials and their properties, adjustment mechanisms, and components.

Data Collection Tools

There are two measuring tool categories essential for any fit test: (1) questionnaires and (2) physical instruments. Questionnaires are used for self-reported measurements, demographics, subjective fit questions, and product version and/or size being tested. Physical instruments are used for anthropometry, product dimensions, and objective fit measurements such as dexterity performance testing of gloves, leakage testing of protective masks, or pupil visibility testing for eye tracking in virtual reality goggles.

Three types of data are *essential* for all fit tests: (1) one-dimensional (1D) anthropometric data, (2) demographics, and (3) subject fit questions. Any measurement that results in one number, such as waist circumference, stature, and weight is an 1D measurement. We need 1D data for statistical analysis of the relationships between body size, demographics, and fit. Two-dimensional (2D) and 3D data are nice for individual subject fit visualization, but statistical analysis tools understanding population fit ranges within a size using 2D or 3D data are not available.

Demographic data should include age, gender, and ethnicity at a minimum. These are used to ensure the appropriate representation of all groups in the sample and the TP. Other data may also be essential for a product, depending on the product.

Subjective questions about fit are essential. There are no objective measurements or human models that can tell us if someone is uncomfortable or able to perform tasks while wearing a product. Computer models feel no pain and can do the impossible. Human subjects, such as fit models in the apparel industry, who are familiar with the company's products also contribute other valuable expert commentary. We can and often do take objective measurements, but they do not replace subjective assessment.

3D scans or other visualization tools are not essential for testing, but they are extremely valuable and should be considered. They help us understand body shape differences as well as the interface between the person and the product when we superimpose scans with and without the product. This is particularly helpful when there are fit issues that don't seem to make sense. Visualizing body shape and location within the product is not possible with 1D measurements. To simply visualize the relationship, the scanner does not need to be precise or accurate so some of the less expensive tools can suffice. However, if we want to measure the location of the body within the product, then we need calibrated, precise, accurate, and reliable scanners as well as special software tools for superimposing the scans.

Product measurements are also important for at least four reasons: (1) to ensure we are testing the size it is supposed to be, which will identify production errors, such as sewing and cutting errors, (2) to document the dimensions for when we make changes during the design process, (3) to identify errors in the pattern, such as base size and grading errors, and (4) to have a record of what we did for future reference. The product measurements can be saved in the questionnaire, or if there is a CAD file or pattern and the measurements are correct, the CAD file or pattern serves as the record.

We start identifying the data we need with the list of fit questions we need to answer according to the COF and proposed instruments we need to answer them. These, in turn, suggest to us what anthropometric and demographic data are relevant. We typically start with a long list of measurements and tools and must pare it down. More is not always better. It is more difficult to find volunteers if the time they spend with us is longer than about an hour. This is especially true if we want them to come back for further testing or if we want to test people in multiple sizes. The list is pared down by

evaluating the quality of the data produced by the different instruments and selecting the combinations of measurements that provide the best answers.

Before finalizing the questionnaires and physical measurements, it is important to conduct a practice run-through, which we call a pilot test. Pilot testing helps us refine the test procedures and practice to make the test run smoothly. Once the actual testing starts, no changes must be made.

The process of selecting and assessing questionnaire instruments is different from selecting and assessing physical measuring instruments, so we discuss them in different sections. Since we start our list of measurements with fit questions from the COF, we begin with a discussion of how to create good questions and evaluate questionnaires for accuracy and reliability. After the questionnaire instrument section, we proceed to the anthropometric measurements and discuss how to select measurements and evaluate measuring tools. This is followed by a discussion of potential objective physical fit measurement tools and how to evaluate them.

Questionnaires

While we would love to have objective ways to measure all aspects of fit or tools that do not require an opinion, some very important things must be measured with a good set of subjective questions. For example, comfort is a subjective opinion that cannot be measured objectively and requires a questionnaire. We can measure related things objectively, such as the amount of pressure something exerts on the body, but we cannot know how much is enough or too much unless we ask the wearer. Each wearer is different in many ways that we do not have the ability to measure objectively. It is not just that a person's tolerance or preference is different. There are also unmeasurable physical differences. For example, nerve ending distribution (the ability to detect pressure) is different depending on the location on the body and it varies from person to person. Fingertips are more sensitive than the center of the palm and some people's fingertips have more nerve endings than others. Also, human shapes differ so the location of the wearable with respect to the body will also differ. Finally, even if we could measure the number of nerve endings on a living person, we still would not know how well the signals are being transmitted to the brain. This can also differ depending on the area of the body and from person to person. *For fit testing, subjective measures are often the most important ones.*

When the subjects represent potential customers, their opinions can be one of the most valuable pieces of information. The fit assessment can help us with marketing as well as fit and the most important fit question might be, "Based on the fit, would you buy this product?". So, the questionnaire used for fit assessment is very important.

A good fit assessment questionnaire has the following properties:

1. It contains questions for both the subject and for the investigator.
2. It is quick and easy to score.
3. The questions with a scale are answered with a standard numbering system.
4. The numbering system maintains the same response order.
5. It contains fit area questions for every aspect of fit in the COF.
6. It has at least one overall fit score question.
7. There is a place to record optional comments.
8. The questions are asked in a way that does not bias the answer.

TABLE 2.5
Sample Fit Test Data Collection Form

Demographics

Subject Number: _____
Place of Birth (menus of choices)
Gender (select one): M F No response
Date: _____
d/m/y
Age at last birthday _____
Birth Date: _____
d/m/y
Ethnicity (select one): W B A O
Occupation (menu of choices)
Hispanic (select one): Y N
Handedness (select one): Right Left Ambi

Anthropometry

Height (cm)	_____
Weight (kg)	_____
Neck Circumference Base (cm)	_____
Shoulder Breadth (cm)	_____
Chest Circumference (cm)	_____
Waist Circumference (cm)	_____
Waist Back Length (cm)	_____
Waist Front Length (cm)	_____

Fit Scores

Size	First Menu of choices	Smaller Menu of choices	Larger Menu of choices
Subject's Rating[a] (circle one)	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
Fitter's Rating[a] (circle one)	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5

Fitter Assessed Best Fitting Size (circle one) 6 8 10 12 14 16 18

Fitter's Best Fit Size Score (circle one) Pass Fail

Subject Comments:

Fitter Comments

[a] Rating 1 = Extremely Bad to 5 = Extremely Good.

An example of a simple fit test data collection form is shown in Table 2.5. In this example, we allow for fit testing of up to three sizes on each subject. The first size selected might be selected randomly or it might be what the investigator deems to be the predicted best-fitting size. Each size worn is rated on a Likert scale of 1 to 5 by the subject and by the fitter. Subjective data can be expressed as a number by using

Likert scales and this makes it easy to record and analyze. We have found that a scale of 1 to 5 is sufficient, but sometimes our customers have insisted on other ranges such as 1 to 10. This is acceptable too, but it does not really improve the result and we usually must group the numbers into two or three categories anyway.

Note there is no place for the subject's name, identification number, or contact information. This is required to protect their privacy. The subject number is something we assign and if we want to use the same subject for multiple tests, we must keep a separate secured file with the name and contact information associated with the subject number.

While the subject's opinion about things like comfort, appearance, personal preference, and ease of use is essential, it is also important to have the fit test investigator record his or her opinion about the fit. The subject rating is important for understanding how the subject feels, which is something the fitter cannot see. The fitter, on the other hand, is someone who knows the agreed upon COF and will give a consistent rating from one subject to the next. The fitter's rating is important for standardizing the scoring and is the most important one for analysis.

To quickly obtain consistent answers it is important to have a set of standardized responses, such as a pull-down menu, a check box, or something to circle rather than relying on the investigator to type or write down words or comments. If the subjects or investigators must write out information or type textual answers, the information often will not get recorded, will have typos, or will be difficult to analyze. Checkable responses are standardized and automatically coded.

Standardizing the response numbering and order for questions that have a scale or order (called ordinal variables) in the response will reduce the chance of making a recording error and make it easier for the data analyst to understand the answers. There are two types of ordered fit questions: one that goes in only one direction from bad to good and one that goes in two directions, bad on both ends and good in the middle. In the first type, always word the questions so the largest number is the best number. For example, "waistline at the back has no bunching or horizontal wrinkling" can be scored 1–5 where 1 is strongly disagree and 5 is strongly agree.

The second type of question could also use the 1 to 5 numbering, with 1 and 2 being poor in one way and 4 and 5 being poor in another, with 3 being just right. However, it is better to use negative numbers for the too small or too short responses, positive numbers for too long or too big, and 0 being just right. For example, "sleeve length is correct when standing normally" can be scored with –2 being extremely too short, –1 is too short, 0 is perfect, 1 is too long, and 2 is extremely too long. Having the numbering system also makes it easier for the people analyzing the data. They will know from the numbers without having to read the question.

In the simple form in Table 2.5, we only have an overall fit rating, but it is often helpful to capture more information about areas of fit issues. When we want to understand the sources of the problem and the part of the COF that was the source of the fit failure, we need to have questions that provide the details. To do this, questions with standardized, checkable answers are used for each issue in each area. These kinds of additional details help us understand the reasoning behind the fit scores.

An example of this type of fit scoring is shown in Table 2.6. This was taken from a fit study of body armor for women (Zehner et al., 1987). Here we see four

TABLE 2.6
Example of Area Fit Scores Question

Rate the Armor Fit in These Areas	Subject's Rating					Investigator Rating				
Circs.	–2 = Very small to +2 = Very large					–2 = Very small to +2 = Very large				
Chest	–2	–1	0	+1	+2	–2	–1	0	+1	+2
Waist	–2	–1	0	+1	+2	–2	–1	0	+1	+2
Neck	–2	–1	0	+1	+2	–2	–1	0	+1	+2
Armholes	–2	–1	0	+1	+2	–2	–1	0	+1	+2
Length	–2 = Very short to +2 = Very long					–2 = Very short to +2 = Very long				
Torso	–2	–1	0	+1	+2	–2	–1	0	+1	+2

circumference areas that were included in the concept of fit and one length area. There is a place for both the subject and the investigator to score the fit in these areas.

After each size is rated in the different areas, it is a good idea to have the fitter decide which was the best-fit size and to give it a pass or fail score, as we did in the previous example (Table 2.5). This is important. It may seem redundant, but it makes the analysis clearer, faster, and easier. Sometimes people will get a good fit in more than one size or they will not get a good fit in any size. This can be difficult to decide after the subject is gone. With this method, the decision is made while the subject is present by a trained fitter who knows what is most important. Then within a few minutes, we can have some preliminary results to show management or whomever it is who wanted the results yesterday.

It is important to have a space for comments for two reasons: (1) issues arise that were not anticipated and (2) subjects sometimes want to make comments and will give better responses if we have a space devoted to listening to them. Sometimes they can have issues that are unrelated to the study and if they do not feel we are listening it can affect their responses.

Table 2.5 includes two comments sections, one for the subject's comments and one for the investigator or fitter. However, while important, comments should be optional because a comment section cannot be relied upon to identify consistent fit issues.

It is important to design the fit questionnaire, so we do not lead the subject to an answer or introduce bias in some way. Choi and Pak (2005) have a nice summary of potential issues in questionnaires. Some common ones for fit measurement are seen in Table 2.7.

Whether we decide (or have an opinion) about what constitutes a good performance or test score before or after the testing, it is best to record the actual score and not just the decision about the quality of the fit performance. It is sometimes tempting to only record if it passes or fails the performance test, but that limits our ability to re-evaluate later. Perhaps in the future, we might have a different purpose for the product for which the score requirement is higher or lower. If we have the score, we can adjust for that. In other words, we might measure the offset or product positioning but not decide what positioning or size is best until analyzing all the data, particularly the subjective questionnaire scores.

TABLE 2.7
Common Issues and Solutions in Drafting Fit Questionnaires

Issue	Alternative
Use of vague words like regularly or occasionally	More specific wording such as twice a week, etc.
Leading questions such as "Are you uncomfortable?"	Rate your comfort on a scale from very uncomfortable to very comfortable
Changing scale, such as 1 to 10 has 10 as best for one question but worst for another	Use the same scale (e.g., 1 to 5) and always have the order the same
Response fatigue from questionnaire that is too long	Minimize the questions as much as possible

In the past, questionnaires were first recorded on paper and then entered into electronic spreadsheets later. Now there are numerous tools available to collect data and automatically save it as a spreadsheet for analysis. Some common ones are Microsoft Forms®, Google Forms®, and Qualtrics XM®. The questionnaire can be accessed anywhere and data entry can occur on a tablet or laptop and the data automatically collated into a spreadsheet file. It is also possible to create a questionnaire and send it out for response via email. This can be useful for follow-up questionnaires once the product is on the market or for a pre-test survey to find potential subjects.

Regardless of whether the forms are paper, electronic, or on-line, it is important to record the exact wording of the questions and the definitions of the responses before starting data collection. This record should be saved with the questionnaire responses for reference. Otherwise, we can end up with questions for which a number is the response, but we won't know what the number means. Is the number 1 a good score or a bad score? Also, the questions in the saved spreadsheet typically have abbreviated names and it can be difficult to remember what the original question was, especially several weeks after data collection is complete. This record should include a copy of the blank questionnaire or at least the order of the questions in the questionnaire to help in the analysis.

It is a good idea if the questionnaire and the definition of the responses are recorded as part of the COF, along with any photos or videos that help explain the agreed upon responses. This helps reduce debate after the testing and is useful for future reference for new items or testing.

Demographic questions are used in the analysis in three ways: (1) to identify fit issues specific to one group, (2) to weight the sample to make it more representative of the TP, and (3) to verify that all groups within the TP will be accommodated. Age, gender, and ethnicity are the most common demographic questions as we know there are body size and shape differences associated with these different groups.

Some demographics help identify fit issues but are not typically used to characterize a TP. These include experience with similar products, handedness, garment sizes worn, hair length and coarseness (for head wearables), beard growth (for oxygen masks), fitness level, and health status. It is important to have enough detail about each subject to be able to understand fit issues or preferences during fit testing and ensure all groups are accommodated.

Experience can be very important for the quality of the subjective feedback. For example, if the wearable is to be worn like a helmet, someone who had experience wearing helmets is likely to be able to provide a more reasonable comfort and useability assessment. They understand that some discomfort is to be expected and some tightness can reduce annoying slippage.

In experimental design and analysis, categories of groups that we wish to verify are adequately accommodated are referred to as blocks. For example, if we are designing a wearable to be worn by both men and women, the gender variable would be treated as a block variable. In the analysis, we would determine if fit outcome differed depending upon gender. If so, we would look at the outcome for each gender separately to see which outcome is better and to understand why.

The race or ethnicity question is not clear-cut and can be sensitive. We ask these questions so we can be sure that we have adequate representation from the categories to ensure all groups will be adequately accommodated. Different racial or ethnic groups can have very different shapes and proportions so accommodating one may not accommodate the others.

Which category to place someone is sometimes unclear. Many people have mixed ancestry, for example, and might fall into multiple groups. In those instances, it is useful to select whatever group seems to be predominant in terms of shape or body proportions. The purpose of the grouping is to ensure that all shapes and proportions are equally accommodated; hence, this is used to select the group. If this is difficult to decide, it is important to allow them to answer "other", "not sure", or "don't wish to answer". However, it is best to have few subjects in these undetermined categories. If we cannot verify that, we have sufficient representation in each group, then minority groups can end up being under-represented which can result in poorer fit for minorities.

It is also *not* helpful to have many groups with all kinds of combinations of ethnicity, such as Asian-Black-Hispanic or Indian-Black-Pacific Islander. Having dozens of categories is essentially putting off the decision about groupings until after the subject is gone when it is harder, if not impossible, to decide. This can be equivalent to having no groupings at all. If we cannot decide which group predominates for an individual, we place them in the "other" category.

Hispanic is a question best asked separately from ethnicity in the United States. Many Hispanics who were born in Caribbean countries have African ancestry. Hispanics from Mexico and Central America often have Native American ancestry. Hispanics from Spain usually have white European ancestry. These groups have very different sizes and shapes. If we ask if they are Hispanic separately from ethnicity, we can better capture the differences and ensure we get good representation for everyone.

If our TP has people from a wide geographic region, it is important to sample from throughout the region and to record their birthplace. This can be ascertained with one question or multiple questions such as country of birth and state or province born.

Multinational studies may need to account for differences between countries. Not only do we need to consider language differences but also sometimes the responses need to be different as well. For example, in the CAESAR study (Blackwell et al., 2002), the question about jacket size had different sizes in the United States than in Italy, because the size numbering systems were different as shown in Table 2.8. This is the type of issue that is quickly identified and resolved with a pilot test in each country.

TABLE 2.8
Same Question for Men in Two Different Countries

Question in the United States: What is your most common jacket size?									
30 or smaller	32	34	36	38	40	42	44	46	48 or larger
Do not know	No response								
Question in Italy: What is your most common jacket size?									
46 or smaller	48	50	52	54	56	58 or larger			
Do not know	No response								

Anthropometry

We use human body measurement (anthropometry) to select people to represent our starting size (our cases), to create models of people for use in our design process, and to assess who achieves a fit and who does not. This enables us to determine where and how much to modify our product to fit the most people in the fewest number of sizes. Once we have our list of fit questions, we can start to list relevant anthropometric measurements we want to consider collecting. We typically start with a long list of measurements and tools and must pare it down.

When choosing anthropometric measurements and measurement tools, the three most important criteria are (1) relevance to the product fit, (2) measurement quality, and (3) comparability to measurements from prior databases of TP. Just like any good science project, we begin this by reviewing what others have done. The prior studies give us ideas about what is possible with the different tools.

People all over the world have been collecting and storing anthropometric data for decades and there are quite a few documents describing how they took their measurements and which tools they used. Some of these documents, such as the Anthropometry of Air Force Women (Clauser et al., 1972), the 1977 survey of U.S. Army Women (Laubach et al., 1977), and the CAESAR, Final Report, Volume II: Descriptions (Blackwell et al., 2002) are freely available for download from the Defense Technical Information Center (DTIC) website (https://discover.dtic.mil).

LaBat and Ryan (2019) have a nice list of 1D anthropometric measurements for different body areas and the WEAR Association has a collection of more than 65 datasets available for free to members (www.bodysizeshape.com). These are helpful places to start.

The International Standards Organization (ISO) has documents available for purchase such as the ISO standard 7250 contains measurement descriptions for standard measurements for international surveys to enable comparisons between datasets (ISO 7250-1:2017, 2017), and the standard for 3D scanning methodologies (ISO 20685-1, 2018). The International Society for the Advancement of Kinanthropometry (ISAK) has a measurement manual that may be purchased and offers training in taking measurements pertaining to sports science.

It is helpful to take some measurements the way others have in previous studies when we want to compare our data to them or if we want to use one of the previous

data sets as a starting TP. However, we should not limit ourselves to those measurements and tools. Those studies most likely had goals that were not related to our product, so they cannot be expected to always take the best measurements related to our product. For example, many studies that are interested in health, body composition, or kinematics take the waist measurements in a horizontal plane located at Omphalion (the belly button), which is perfectly reasonable for those applications. However, the waist of our clothes does not sit at Omphalion and is not a horizontal measurement. If the waist where the product will be worn is the most relevant, we should take the more relevant measurement.

Things to look for in the measurement descriptions include (1) the tool used, (2) the posture or pose of the subject, (3) the positioning of the tool, and (4) the direction of the measurement. For example. in the 1968 survey of United States Air Force (USAF) women (Clauser et al., 1972), the posture for the Sitting Height measurement was, "Subject sits erect, head in the Frankfurt plan, upper arms hanging relaxed, forearm and hands extended horizontally". The tool used was an anthropometer. The positioning of the tool was "firmly touching the scalp" and "from the sitting surface to the top of the head". The direction was vertical. Photos are also helpful.

If we take our measurement with a different pose, tool, positioning, or direction it can result in different measurement values, making them seem larger or smaller when they are not. One common example is when we use 3D scanners to take a 1D measurement versus using a manual tool such as a tape measure or a caliper. These different tools can result in very different values simply because of the tool used. Manual tools touch the body introducing some compression that influences the measurement values, whereas scanners do not. Tape measures span some of the peaks and valleys along the way but scanners may not or may not in exactly the same way. Therefore, even if both types of measurements are good quality measurements, it does not mean they are the same.

An example of the effect of posture or positioning was described in the 1968 USAF study mentioned earlier. In addition to the Sitting Height with the subject sitting erectly, they also took a Sitting Height measurement they called Sitting Height, Relaxed. There are slight differences in the postures for these two measurements, because the subject sat relaxed rather than erect, and the result was a 13 mm difference in the average values for the two.

Once we have an idea of what measurements we want to capture; we must consider the tools to use to capture them. Most anthropometric tools can be classified as one of two types: (1) manual and (2) imaging. Manual tools are tools used to measure the subject directly. Manual tools examples include the following:

- Anthropometers
- Stadiometers
- Spreading calipers
- Sliding calipers
- Tape measures
- Pupillometers
- Headboards
- Footboxes
- Grip measurements
- Ring-size finger measurement tools

Imaging tools are tools that capture a copy of some parts of the subject, such as points from some surfaces and/or named landmark on the surface, and the measurements are taken from the copy. Imaging tool examples include the following:

- 2D imaging
  - Overlaid photographs
  - Photographs with grids
- 3D imaging
  - Stereophotogrammetry
  - Whole body scanners
  - Head scanners
  - Face scanners
  - Foot scanners
  - Hand-held scanner
- 4D imaging (3D + time)
  - Motion tracking of landmarks
  - Whole body scanners with motion frames
  - Face scanners with motion frames

2D imaging, or use of 2D photographs, was used before we had 3D imaging. It is not very accurate for measuring but can be a quick and inexpensive way to visualize the product on the subject. Two photos are taken and overlaid in a software tool such as PowerPoint™ with one image made mostly transparent. However, these overlays are very crude and it can be difficult to see or measure the relationship. 3D imaging is better for visualizing the interface. Visualizing and measuring the interface is addressed in the physical fit measurement section.

Stereophotogrammetry tools capture pairs of images from cameras that are in known, calibrated locations with respect to each other and stereo viewers to identify points on the surface of the body. Prior to the 1990s and the advent of 3D scanning technology, this was used to get body segment volumes and 3D landmarks for estimating human biomechanics before the advent of 3D body (Herron et al., 1976; McConville et al., 1980). While 3D scanners offer a faster and more accurate technology for capturing the body surface, these early studies defined landmarks and body segmentation methods that are essential for measurement extraction and human modeling today.

When we refer to 4D scanning, we mean 3D scanning with time added or full body scans in motion. This is different from motion tracking, which tracks a limited set of landmarks over time rather than the body surface. Motion tracking might be used as a physical fit assessment method or a performance assessment method, but it is not used for body measurement itself. Therefore, it will be discussed in the physical fit measurement section.

We select tools for the data collection by evaluating and comparing them for:

- Tool quality
- Data quality
- Personnel and software requirements
- Time to results
- Cost

Tool and data quality must be assessed for each measurement to be taken. Some tools are best for one type of measurement but not for others.

Good quality tools are those with a high enough resolution for the measurement and the ability to stay calibrated. Resolution is the closeness of the settings, gradations, or points. When gradations or points are close together it is called high resolution and when they are far apart, they are called low resolution. If, for example, the closest tick marks on one instrument are 1 mm apart, it is said to have a 1 mm resolution. An instrument that has tick marks every 0.1 mm would be considered higher resolution. Some measurements require higher resolution than others.

Calibration is the process of determining, checking, or rectifying the settings or gradations on a measuring instrument or other piece of precision equipment. The tools we select must be able to be calibrated, maintain calibration for the entire test or study, and be calibrated for all sizes and shapes of subjects. Some tools deteriorate or get damaged, so we need to be able to ensure they are working properly throughout data collection. For example, a scale for measuring weight (body mass) should be checked to ensure it is outputting values that are accurate by placing objects of known weight on the scale. If it is out of calibration, it means it is outputting the wrong number. A good instrument should be able to be adjusted to ensure it provides the correct values for the full range of subjects or objects to be measured. Some instruments may be precise and correctly calibrated within a range of sizes but not be outside that range. For example, 3D scanners can have a scanning area that is precise and calibrated but outside that area, they can be imprecise or difficult to keep calibrated. This happens when a scanner is designed for a standing posture with the arms at the sides but intended to be used for a seated posture or with arms akimbo. Also, sometimes the scanners are calibrated with a small- or medium-sized manikin and the point cloud is dense enough, but for a large person the points can be spread too thin. Other tools have issues as well. A scale for measuring weight may only be calibrated to 400 lb or 180 kg. So, it won't be sufficient for measuring morbidly obese people. A tape measure may not be long enough for measuring things such as Vertical Trunk Circumference. These are things that must be checked before an instrument is selected for use in a study. A good instrument should also stay calibrated for at least a full day.

Good quality data is precise, accurate, and reliable. Precision is the consistency of measurement. Accuracy is the ability of the tool to capture the true measurement we are trying to capture. Reliability is the ability to get the same measurement on live subjects under consistent conditions. Reliability is affected by precision and accuracy of the tools but is also affected by the fact that we are measuring people who are constantly changing.

The difference between precision and accuracy is illustrated in Figure 2.1. In Part a of this figure, we see three measurement values that are the same, but they are not the true value. These are precise but not accurate. This can happen if a tool is not assembled correctly or if the starting point or end point for the tool is different than the starting point or end point of the measurement. For example, an anthropometer's end point for a measurement is the bottom of the tool's arm, but when we measure "Elbow Height, Sitting" we want the measurement to the top of the tool's arm. The arm is 10 mm thick so the measurement reading on the anthropometer will be consistently off by 10 mm. In Part b, we see three measurements that have an average

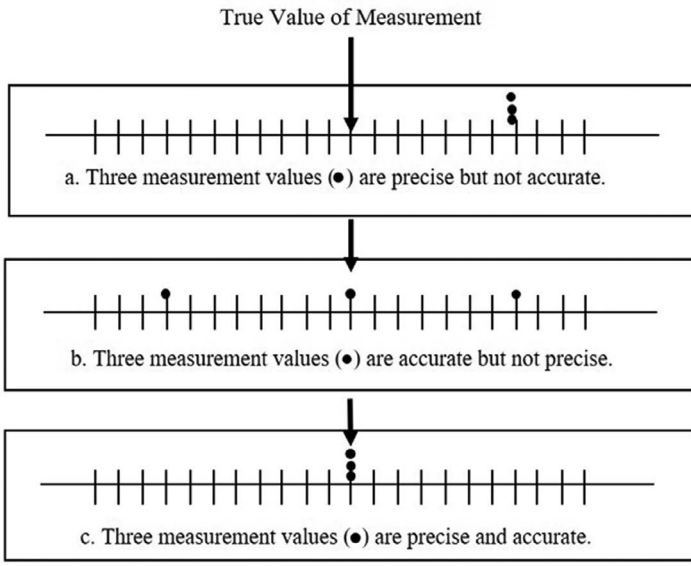


FIGURE 2.1 Precision and accuracy illustration.

that matches the true value, but they are three very different values. These are accurate on average but are not precise. In Part c, we see three values that are both precise and accurate. This is our goal.

Tool and data quality are assessed and compared by first using fixed static objects of known size before we then assess them on live subjects. People change constantly so the assessment with static objects gives us an idea about how good a measurement we can get if we don't have something that is breathing and moving. The objects used to assess the precision of an instrument must also reflect the size complexity of the body or body areas to be measured with it because the precision can be different for different sizes and shapes.

Precision can be limited by the resolution of the tool. Some 3D scanners, such as the scanner used in the CAESAR project (Robinette et al., 2002) have a resolution of 3–5 mm. If we need a precision of 1 mm or less, for facial measurements, for example, a 3–5 mm resolution would not provide measurements precise enough for our purposes. Also, each time a scan is done, the points that are collected can fall in different places on the body, so measurement precision for specific landmarks can be further degraded.

Some aspects of accuracy can be assessed using static objects but many of the things that affect accuracy cannot be effectively separated from other things that affect reliability of the measurements. Therefore, accuracy and reliability are usually evaluated at the same time on live subjects. For example, if we do not have any way to know what the "true" anthropometric measurement is on a live subject, we cannot evaluate accuracy separately from reliability.

Reliability is affected by all sources of measurement error including resolution, calibration, precision, and accuracy. It is affected by posture or pose with some

postures being difficult to maintain, for example. They can be affected by instrument or measurement positioning, the visibility of landmarks, and the quality of algorithms for landmark detection and measurement extraction. It can also be affected by post processing. This is particularly an issue for 3D scan tools which may require a transformation of the data into a format that changes the body surface such that measurements may no longer be accurate.

Reliability is also affected by the measurer or observer whether that measurer is human or an artificial intelligence (AI) software tool. Human measurer's skill level, training, alertness, and attention to detail can affect measurement reliability. This can vary depending upon the training of the measurers, so reliability assessment is done when the human measurers are trained. Training is discussed in further detail in the personnel and facilities section.

For imaging tools with AI software, reliability includes the skill level of the software tools and the effect of subject positioning in the image or scan. Many 1D measurements and 3D landmarks are difficult for an AI tool to properly or consistently locate on live people which results in the inability to get the same measurement value each time a person is measured. For example, it can be particularly problematic to locate landmarks for larger people whose body surface is far away from underlying bony landmarks or who may have more complex surface shape due to fat folds. It can also be difficult to find landmarks reliably using AI on people who have lots of dark hair that does not scan well or who have extreme asymmetry. The mis-identification of landmarks can result in a biased sample as well as an unreliable one. This assessment should be done before deciding what tools to use and it is included in the imaging tool assessment section.

A tool can be very precise, accurate, and reliable for some measurements but not reliable for others. We have had 3D scanning tools for measuring the head and face since the mid-1980s (Blackwell & Robinette, 1993; Robinette, 1986) and we have had whole body scanners since the 1990s (Robinette et al., 1999). Scanners are valuable because they provide information that we cannot get from manual measurements, particularly fit visualization. However, they are inadequate and inaccurate when used to extract or derive many 1D measurements. Each measurement should be tested to see if they are the same when using different tools, that is, a scanner versus a tape measure (ISO 20685). An ISO standard for 3D anthropometry indicates that *not* all measurements in the ISO basic body measurement standard 7250 (ISO 7250-1:2017, 2017) are "…well-suited to extraction from 3-D scanned images".

In Table 2.9, we give some examples of measurements for which scanners are poor tools and some for which they are the best tools. Measurements that have areas hidden from view, such as Chest Girth at Scye and Crotch Length, are best taken with manual tools. Measurements obscured by hair such as Head Breadth are also best taken with manual tools. It can also be easier to measure many different poses rapidly with manual tools. On the other hand, 3D scans capture the shape as well as the size, so they are better for shape comparison, for comparing subjects with and without the product in place, and can be better than 1D for 3D modeling. If there is a need for a physical or digital copy of an individual, processed 3D scans can be entered into a computer software tool for designing or 3D printing.

It is essential that some of the measurements we capture are 1D for a range of fit analysis and statistical comparison of prototypes and components. This includes all

TABLE 2.9
Best and Worst Tools for Select Measurements

Measurement	Best	Alternative	Not Recommended	Comments
Chest Girth at Scye	Steel tape	Cloth tape	3D scanner	Underarm obscured from view in scan
Head Breadth	Spreading caliper	Sliding caliper	3D scanner	Hair obscures measurement and location must be detected by palpation
Crotch Length	Steel tape	Cloth tape	3D scanner	Crotch area obscured from scan
Arm Length Shoulder to Wrist	Steel tape measure	Cloth tape	3D scanner	Not a scanning pose and requires spanning some surface contours
Right Acromion 3D Location	3D scanner with manual pre-scan marker	3D Faro Arm	3D scanner without pre-scan marking	This is located through palpation and cannot be found reliably from surface only
Contour Comparison/ Overlay	3D scanner	2D photos	1D tools	It is impossible to align locations for comparison with 1D tools

the dimensions we think might be the best predictor of the best-fitting size. For these measurements, it is usually best to use manual tools such as tape measures and calipers. They do not require special software or programming knowledge to use, understand, calibrate, and test for reliability. These tools are inexpensive and our customers or users are likely to have these tools as well, making them suitable for size selection.

It is also helpful if at least some of the 1D measurements use the same tools and methods as previous studies, particularly those we want to use for representing our TP, so we can compare our subjects to theirs. 1D measurements extracted from 3D scans can be acceptable (provided they meet the measurement quality criteria), but they are not the same as measurements taken with calipers and tape measures so making comparisons to previous studies is compromised.

Some of the pros and cons of manual versus 3D imaging tools are shown in Table 2.10. Because each type provides things that the other does not, we often find it best to have both scanners and manual tools.

When measuring with manual tools, the body positioning can be adapted and changed for each measurement during the process of taking a sequence of measurements. This is not as easy to do for 3D body scanning. When 3D body scanning technology is used, the body scan is only performed in a limited number of postures

TABLE 2.10
Comparison of Manual Tools Versus 3D Scanners That Are Calibrated and Precise to at Least +/– 2 mm

	Manual Tools	3D Imaging (Scanners)
Cost	$25 to $5000	$5000 to $300000
Tool Availability	Widespread	Limited
Databases available	More than 50	Less than 5
Calibration/precision assessment	Quick and easy	Time consuming, complex, and sometimes not available
Postures and poses	Many	Limited
Circumferences	Reliable	Not reliable
Heights	Reliable	Only reliable to visible and repeatable landmarks
Point-to-point distances	Reliable	Reliable
Pre-measurement landmarking	Required	Required
Post-measuring processing	None	Required
Time to statistical analysis	Minutes	Hours/days
Visualization of actual body shape	Not available	Free software tools available
Visualization of actual product on actual body	Not available	Free software tools available
3D landmark locations	Not available	Available with software tools
Watertight model	Not available	Available with software tools
Revisit the subject	Not available	Available with software tools

from which the information is subsequently extracted. For instance, the standard ISO 20685 proposes four scanning postures to calculate the body measurements included in ISO 7250-1:2017. New 4D scanning systems are being developed that may improve this, but they are not yet widely available.

Tool assessments are more complex for imaging tools than for manual tools. While manual tools typically have one simple data output, each imaging tool can have many kinds of outputs and each of the outputs we intend to use must be assessed. Imaging tool outputs are not standardized, so assessment methods can be different for each tool. In addition, when the data is modeled or translated to a new format, the data can be changed. If we plan to use these formats, we need to assess them as well. Therefore, we explain how to assess manual tools separately from imaging tools.

Manual Tool Assessment

The resolution of manual tools is easy to determine. It is simply the spread of the tick marks on the tool or the spread of the 2D grid lines. It is usually best to use metric tools with resolutions in millimeters, centimeters, or even tenths of millimeters rather than tools with English units such as inches or fractions of inches. Metric tools are easier to record correctly since the data must be in decimals to analyze.

Manual tools are calibrated and assessed for precision using gauges and weights. The weights are used for calibrating scales. Durable, precisely machined gauges such as the one shown in Figure 2.2 are used for calipers, anthropometers, and tape

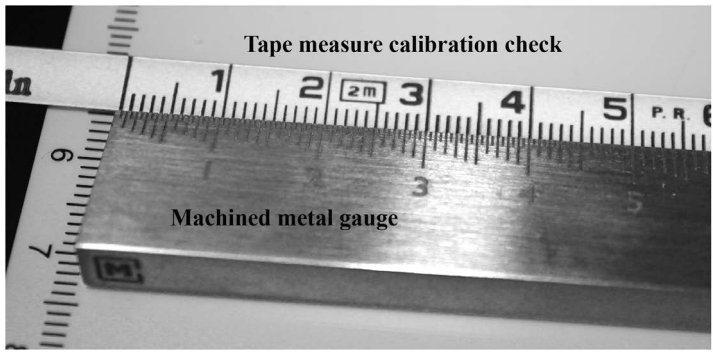


FIGURE 2.2 Calibration gauge.

measures. This example is precisely machined brass and is used to check rulers and tape measures, such as the steel tape shown in the photo.

Several things can affect calibration over time. For example, if the instrument is made of cloth, like some tape measures, it could stretch over time taking it out of calibration. Narrow metal ones, such as the one shown in Figure 2.3, get the most reliable measurements, since they do not stretch much over time like a cloth tape and being thin (this one is 6 mm in width), they fit into smaller spaces than wider tapes. The tape shown is made by Lufkin with one side in metric units with mm marks which is the side we use. The other side looks like inches but is not. It is intended for measuring diameters of pipes and tree limbs or trunks and is cm to mm. This is an example of a precise instrument that is not accurate because it is not measuring

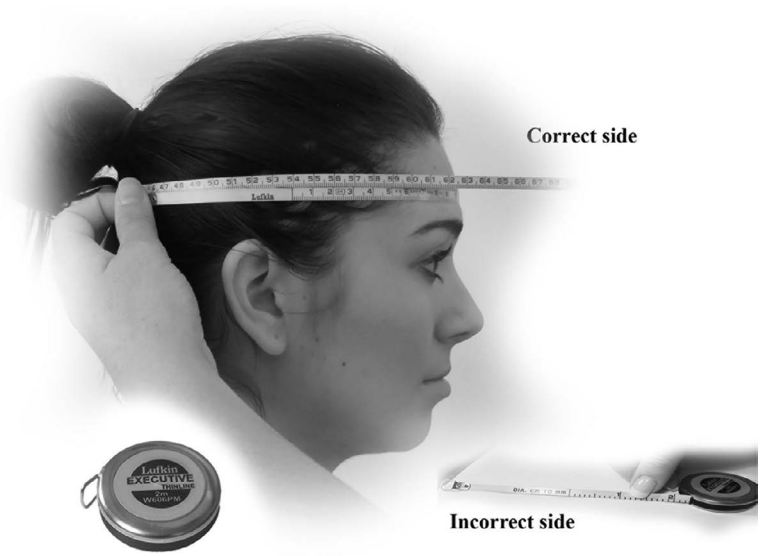


FIGURE 2.3 Lufkin 6 mm wide metal tape.

FIGURE 2.4 Anthropometer and sliding caliper.

circumference. It may be accurate for diameters but is not accurate for circumferences. Unless you want a diameter from a circumference, this side should not be used. It illustrates another reason for checking the calibration. Sometimes, the instrument is not measuring what we think it is.

It is important not to drop an instrument which can bend or knock it out of calibration. As explained in the Army measurer's handbook (Hotzman et al., 2011), if an instrument is dropped, a gauge should be used to check its accuracy before using it again. They also caution us not to lean an anthropometer up against a wall or table because it can easily fall and be damaged.

Most anthropometers are multi-tools that can be configured into spreading calipers. One example is shown in Figure 2.4. The anthropometer comes in pieces that can be assembled to measure heights from a seated or standing surface or configured to measure breadths and depths. The portion of the tool must be assembled correctly and read on the correct side to give the correct reading for each type of measurement. The heights configuration, such as the one used for the measurement shown, indicates the measurement to the bottom of the bar. The other side indicates the measurement to the top (or the inside) of the bar for the caliper breadth reading. It is important to check the calibration each time instruments are assembled to avoid getting this mixed up.

Accuracy and reliability using manual tools are done by repeatedly measuring live subjects. This should be done by a skilled and trained measurer. An unskilled measurer will add variability to the measurements that are due to the lack of skill rather than the tool. Luckily, repeated measuring for many of the common manual tools has been done by others. We can review what they have done when evaluating what tools to use (Gordon & Bradtmiller, 1992; Gordon et al., 1989; Hotzman et al., 2011). These studies quantify measurer or observer error, and the result is an assessment of the level of accuracy and reliability that can be expected when skilled

measurers use the tools for the measurements they included. If we plan to use a different tool or take a different measurement, we should follow the same procedure to assess them as these authors used.

Imaging Tools Assessment

Assessing imaging tools for quality is more complex. There are two categories of assessment for imaging tools: (1) assessment for visualization and (2) assessment for measurement extraction. Sometimes, we only want an imaging tool for visualizing shapes and relationships between products and the body. This can be extremely valuable for understanding fit, but it is essentially a subjective fit assessment tool and any results recorded might be part of the subjective fit questionnaire. If we are only using the scans to visualize the fit, we might not need dimensional accuracy, so we can make do with an uncalibrated inexpensive hand-held scanner and with minimal file format translation, we can visualize using free or inexpensive tools, such as Blender™ or even superimposed 2D photographs. Quality assessment needs only be a check to verify that we can see what we need in the images. This might require an evaluation of the resolution in the images but not an evaluation of calibration, precision, accuracy, or reliability.

If, on the other hand, we plan to use numerical data produced from imaging, we need to first understand how the tool works, what data is being produced, how it is produced, and how it is processed. Because imaging tools also require specialized software, we must also assess the quality of the software and the ability to use the data in the software tools we are using for designing and producing our products. In other words, does the data import work in our CAD modeling or 3D printing and manufacturing tools? The best data in the world is worthless if we cannot use it.

These procedures can be explained by explaining the assessment of 3D body scanning tools. 3D body scanning technologies obtain a set of 3D points on the surface of the body that is referred to as the *point cloud*. This point cloud is not a surface and the points in it are not connected to each other. The points are spread across the surface and the distance of the spread is called the resolution. Unlike manual tools that have a single resolution that does not vary, the resolution for 3D scanners varies from scanner to scanner, from person to person, from scan to scan, and can be different in different areas of the body. Therefore, it is much more complex to evaluate.

Larger people have point clouds with more points in them than smaller people and the location of the points in the point cloud is different every time a person is scanned. In other words, the first point in one scan is not in the same place as the first point in another scan. To relate one scan to another, we need some identifiable points that are the same in every scan. We call these landmarks. We use named landmarks that can be seen in a scan to help us orient the points in the scan and relate one scan to the next. If a landmark is not clear without marking the body, we add markers on the body to make it visible. This requires that the markers are visible in the software tool we plan to use to view the scans, and we need software to determine the 3D coordinates of the landmarks. The resolution of the scanned points affects the precision and accuracy of the landmark location because there may not be a scanned point at the precise location of the landmark. Also, the scanned points change every time someone is scanned, so the point that is closest to the landmark can change from scan to scan.

Robinette and Daanen evaluated the precision and reliability of extracting landmark locations and measurements from the landmarks for two different scanners when the landmarks are marked before scanning and visible in the scans for identification (Robinette & Daanen, 2006). They scanned each subject three times in each scanner and evaluated the differences between 1D measurements calculated as distances between landmarks. The measurement repeatability (reliability) was within the range of repeatability with manual tools. The reliability results for the two scanners were similar but one was slightly better than the other.

Robinette and Daanen evaluated only measurements that were the shortest distances between two points because these types of measurements are invariant to the point of view and the axis system used. These are calculated using all three coordinates (x, y, and z) for both points. This is different from calculating a difference between two points in one direction such as Glabella to Pupil in the fore-aft direction only. It is also different than treating each individual coordinate for a landmark as a separate variable, which is how they appear in a data spreadsheet. In general, *single coordinates for landmarks should* ***not*** *be used as independent variables.* They are subject to point of view or orientation error. There is no standardizable axis system that can make the individual coordinates or directions such as up, down, back, or front comparable from one person to the next.

This issue is illustrated in Figure 2.5. This shows three head orientations and the effect it has on the distance in the Z direction (fore-aft) measurement Glabella-z to Rhinion-z. This is the same person in all three places. The head is simply rotated. The landmarks have not moved with respect to each other and the distance between the landmarks as calculated with all three coordinates has not changed. However, the values for the z distances are quite different, all due to orientation uncertainty or point-of-view error.

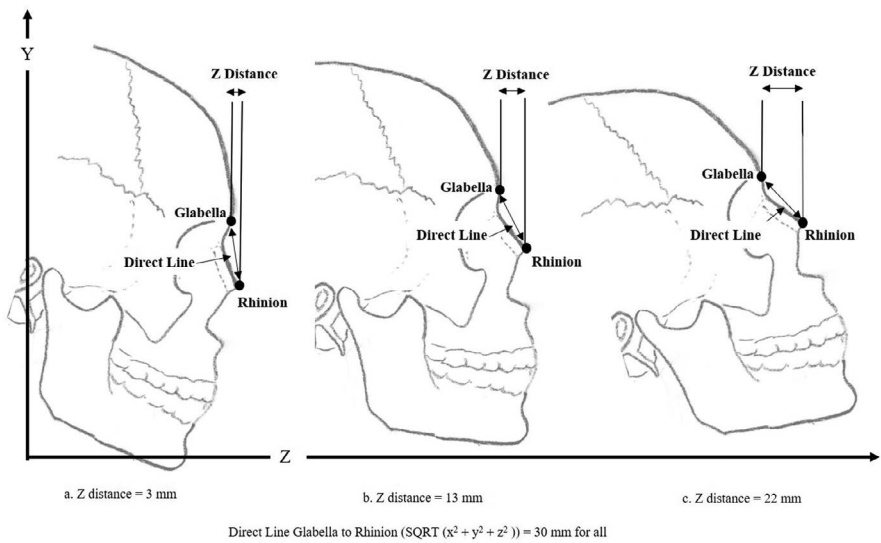


FIGURE 2.5 Point-of-view error illustration.

These types of measurements are particularly a problem for the head and face for three reasons: (1) head and face measurements are small, so small accuracy errors have a large impact, (2) there are few reliable landmarks on the top and back of the head for orienting the point of view or axis system, and (3) the head is roughly a sphere that can be rotated multiple ways with landmarks that vary independently of each other. As a result, the inaccuracies introduced by orientation are larger than the differences between people, which means the measurements are unreliable, or worse, misleading if they are used. If measurement errors are large compared with the true differences, then the differences observed between two subjects could be due purely to error rather than to a genuine difference.

Whitestone and Robinette (1997) review the many ways people have tried orienting the head in a standard way so that the 1D measurements could be used. None of the orientations provide accurate or reliable 1D measurements. There is one exception to this for product design purposes. That is if the axis system and orientation are defined by the product when the actual subject is scanned in the product and both the subject landmarks and the product landmarks are visible in the same scan. This is not the same as placing a CAD model of the subject in a CAD model of the product, because the CAD model placement is hypothetical, not real, and can be subject to point-of-view or placement error.

Another important issue with using single coordinates from 3D landmarks has to do with correlation. Since all landmarks move together, all the z distances change together which means they are correlated. The fact that they are correlated can give the impression that the distances are important. However, it merely indicates that there is a substantial point-of-view error. This topic is discussed in further detail in Chapter 6.

This is particularly a problem for principal component analysis (PCA) of head measurements because these correlated errors can overwhelm any other true correlations between measurements and make the results meaningless. It may look like a component is a face depth or the eye depth component but really it is just indicating that the depth errors all vary together. When the head is rotated up, they all get smaller and when the head is rotated down, they all get larger. More information about PCA analysis and this issue is provided in Chapter 3.

When evaluating 3D scanners for resolution, calibration, and precision, it is necessary to evaluate both the hardware and the software components. The scanners Robinette and Daanen used had been assessed for resolution, calibration, and precision using the same calibration and precision gauges and both had been calibrated prior to scanning. These assessments revealed they had similar point resolutions, but they had differences in precision and in what parts of the body surface were missing from the scan. To get a full 360-degree scan of a person, multiple scan heads that can view all around the body simultaneously are required. These scanners used similar scan heads and technology, but they used different software tools to process and combine images from the different scan heads.

Calibration gauges for 3D scanning are more complex than for manual instruments. They require calibrating in multiple directions, for different shapes, for point precision, and point resolution throughout the complex surface. One of the calibration gauges used for calibrating and checking the calibration of the scanner during the CAESAR™ project is shown in Figure 2.6. This gauge has both curved surfaces

FIGURE 2.6 3D scanner calibration gauge.

and flat surfaces and it has markers and indents to ensure the pre-marked landmarks would be visible in the scan. The width of the largest part is approximately the width of the largest subject expected. Small and large portions were used to ensure sufficient accuracy, precision, and resolution in all body regions.

This gauge was specific to this scanner and project. It may be necessary to construct a different calibration object and process depending on the product, the measurements of interest, the type of scanner, and the target market. The CAESAR™ project was the first large scale 3D scanning survey and the calibration process was developed along with the data collection protocol.

In the ISO standard on 3D scanning methodologies (ISO 20685-1, 2018), the calibration gauge is called the "test object". There is no guidance about the test object other than it should have known dimensions and be similar to dimensions found in humans, much like the gauge used for the CAESAR™ project.

The best scanners have a calibration system and process provided by the manufacturer. Whether one is provided or not, it is good to have a calibration process that can be checked periodically. Just like with manual tools, things can happen that can knock it out of calibration. If that happens, all measurements taken from it can be affected.

Because the scanners employ light that bounces off the surface and triangulation to find the surface, the point clouds have missing information where the surfaces are hidden (e.g., under the armpits and at the crotch), and large breaks in the data where surfaces are horizontal to the cameras used to capture the light. These missing areas are called occluded areas and some typical ones are shown in Figure 2.7.

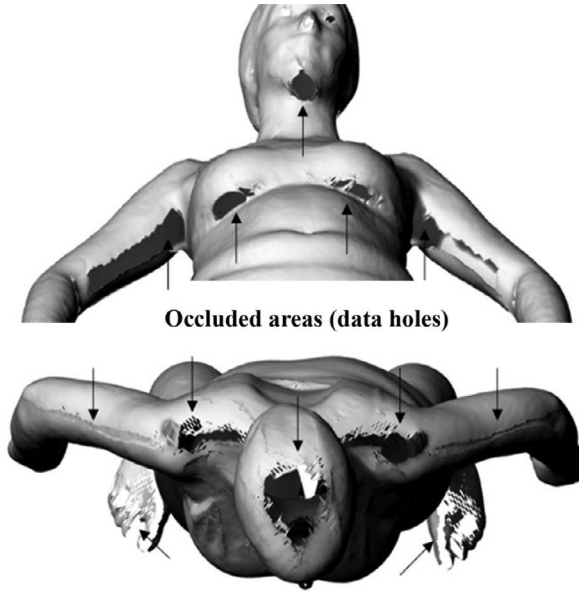


FIGURE 2.7 Typical occluded areas in a body scan.

The estimation of the surface to fill the holes introduces fake information that should be identified and tracked during the processing. In the case of big holes, such as the holes in occluded areas shown in Figure 2.7, the lack of anatomical shape of these patches can be evident. These occluded areas are affected by a person's pose as well as their body size and shape. Therefore, when assessing accuracy and reliability for measurements of interest, we must consider a variety of sizes and shapes.

One of the biggest differences between manual tools and scanners is the data processing required and its effect on the measurement values. The measurements we get from manual tools typically do not need additional processing or calculation, but scanners require some post-scanning processing. If we are using accurate scans to identify landmark locations, we will need to translate the scan into a format that is readable by a software tool for identifying landmark locations. In addition, some software tools, particularly CAD tools, require watertight models which necessitate more processing than a simple format translation. When we measure with a scanner, we are measuring a digital copy of the person. The quality of the copy impacts the quality of the measurement. To get a good quality copy, we need good quality processing. This does not necessarily come with the scanner when it is purchased.

We must connect the points in the point cloud in some way to get a surface. If we create a surface, we are adding data points that were not captured, so they are not the true surface. This makes them less accurate than the actual points in the point cloud. We must take this into account when we evaluate accuracy.

A variety of algorithms (public or proprietary) are used to generate a 3D body mesh that connects the points to create a surface. One is called a polygonal mesh, which connects points to create a surface with flat triangles. Another is called a

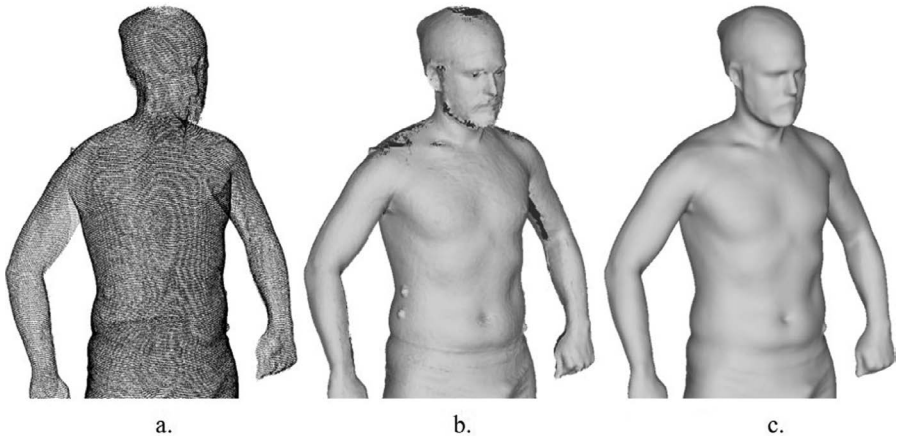


FIGURE 2.8 Three versions of a scan: (a) point cloud, (b) polygonal mesh, and (c) NURBS watertight model.

Non-Uniform Rational B-spline Basis Function Surface (NURBS), which connects the points using curved surfaces or some other surface estimation process. The polygonal mesh uses the original points and adds flat surfaces to fill the holes. The NURBS method moves some of the original points to better fit a curved surface and adds surfaces to fill holes. Both add points that are not in the scan and NURBS also drops some of the original points to create a smoother surface. This makes a more natural looking surface, but it smooths away feature details in the process and is less accurate.

The original point cloud is compared to a polygonal mesh and a NURBS watertight model in Figure 2.8. The polygonal mesh in this figure is not watertight. In other words, the larger holes between points remain to be filled. Sometimes investigators choose to use curved surfaces in these areas. The watertight NURBS model has all the holes filled but it has smoothed some of the surfaces such as the ears, eyebrows, and the bumper landmark markers on the side.

Sometimes the surface estimation connects the wrong points. An example is shown in Figure 2.9. The fake surface added at the crotch area using existing algorithms, such as Poisson reconstruction, may affect a search and location of a landmark to measure, for instance, the crotch height of the inseam leg length. Hole filling also adds fake surfaces at several parts of the body scan. The top of the head is typically an area of missing information. The reconstructed surface can introduce error in the head surface affecting a relevant measurement such as the stature.

Watertight objects are objects with completely closed surfaces (no holes in the surface) and are needed for CAD compatibility. The point cloud data from a 3D scanner might be saved in one of several standard file formats. The most popular are *stl, *ply, and *obj. Software that deals with these formats is typically conceived for animation of characters and they add meshes and surfaces using creative tools. However, the CAD software is used for industrial engineering works with solid entities and parametric surfaces. The standard files to support these are usually *iges or

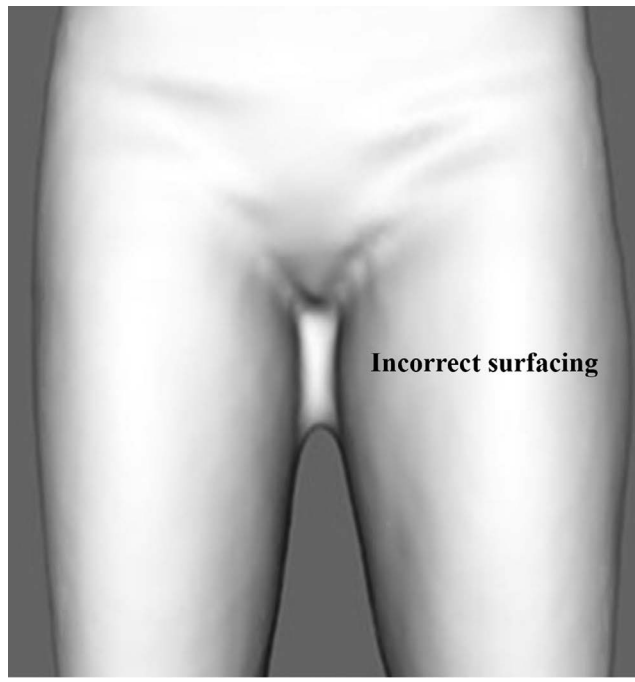


FIGURE 2.9 Example of incorrect connecting of occluded area.

*step. A raw 3D mesh converted directly from triangles to surfaces is not usable for a parametric CAD environment. Each triangle of the 3D body mesh is converted into a single surface, obtaining a model made of thousands of surfaces difficult to operate if the software can import it at all without crashing. The conversion of a 3D mesh into a single or a few numbers of parametric surfaces is not trivial and can introduce inaccuracies in the data. In addition, even if it can be converted and imported into CAD, it can still be very slow or can crash the system. So, if this is something desired be sure to test to ensure it works before you buy into the system.

Even if a watertight version of a scan is produced, it is important to keep the original scan. The two versions can be compared, so we know which points are actual scan points and which are modified or created by the model. This is one way to assess of precision and accuracy. This can be particularly important when using a CAD model to create our product prototype and is one reason why the CAD models inaccurately portray fit such that live subjects are needed for testing.

The definition and control of the posture and body positioning during the measuring session are perhaps of the greatest sources of measuring error and difference. Postures need to be defined well enough so they can be duplicated, and they must be able to be maintained for the duration of the measurement. The postures possible and maintainable differ for manual versus scanning tools.

This is an area where 4D body scanning with 3D surface details (not just motion capture of landmarks) excels. Not only can we obtain many postures in one scan but also we can evaluate actual movement and changes. 4D scanning also permits

improved estimation of occluded areas because something might be occluded at one point but be visible at another.

The basic postures often used in traditional anthropometry (e.g., standing with feet together and arms extended and close to the body) are often not suitable for scanning due to the number and location of occluded areas. This impacts the comparability of 1D measurements extracted from scans to 1D measurements taken with other tools.

For example, for full body scanning in the standard postures, the armpits and crotch are typically occluded (Kouchi, 2014; Bragança et al., 2018). To get data in these areas, the arms and legs must be separated and the separation sometimes must be quite large for some of the larger subjects. When this is done, the measurements related to the shoulders and hips change (Gill & Parker, 2017; Kouchi & Mochimaru, 2005; Mckinnon & Istook, 2002). Kouchi and Mochimaru measured two subjects using a motion capture system as they move their arms from 0-degree arm abduction angle (vertical position) to an abduction angle of 90 degrees (horizontal position) (Kouchi & Mochimaru, 2005). They found that Acromial Height remains constant up to an approximate angle of 20 degrees, after that, it increases substantially. For both subjects, it increased approximately 20 mm from the 20-degree angle to the 50-degree angle. Biacromial Breadth got smaller as the abduction angle was increased and was noticeable at just a 5 degrees of abduction of the arms. For subject 1, the Biacromial Breadth decreased from approximately 440 mm to 390 mm between the abduction angles of 20 and 50 degrees, a difference of 50 mm. For subject 2, the Biacromial Breadth decreases from approximately 410 mm to 370 mm between the abduction angles of 20 and 50 degrees, a difference of 40 mm.

Regarding the position of the feet, the optimal separation to ensure the integrity of the data, avoiding hidden areas as much as possible, is 10.16 cm according to the study carried out by Mckinnon and Istook (2002). However, this position of the feet apart increases the measurement of the hip contour with respect to the natural posture, by almost 1 cm, depending on the height at which the hip measurement is taken (Gill & Parker, 2017).

Another aspect of posture is the change of the body as the subject breathes. For some measurements, particularly for measurements or scans of the torso, it is important to monitor and/or control breathing during measurement. Manual measurements, such as Chest Circumference, are done by recording the measurement at a specific point in the breathing process. For example, in the 1968 USAF survey, the Bust Circumference was recorded at the "point of maximum quiet inspiration" (Clauser et al., 1972). This would be the largest value of the tape when the subject is quietly breathing.

In the ISO standard for basic human body measurements (ISO 7250-1:2017, 2017), the breathing conditions are not specified with each individual measurement description. Instead, there is a statement that "Chest and other measurements affected by breathing should be taken during gentle breathing".

With 3D scanning, the scan is often not instantaneous, so watching the breath and taking the measure at one specific point is not possible. Breathing must be controlled in some other way (Daanen et al., 1997; Kouchi, 2014; Lu & Wang, 2010; Mckinnon & Istook, 2002; Perkins et al., 2000; Pheasant, 2014). Breathing can introduce

variations in the armpit, bust, and under bust circumferences greater between 1.5 and 3 cm in the inhalation or exhalation phases. Maintaining normal breathing is recommended to reduce variation in these measurements due to respiration. For scanners that take several seconds to complete a scan, we will not know the point at which the breathing process was scanned.

Nowadays, the progress of computer vision technology enables the use of real-time detection of joints in video images that can be used to help the posture adoption during the scanning session. An alternative used in some studies is a special physical support for the hands. This system needs to be adjusted to the arm length of each subject and later on should be deleted from the body scan to avoid interferences during the measuring extraction.

The position and direction of the measurement are defined using reference planes (defining horizontal, vertical, etc.) and landmarks. Landmarks are reference points that indicate a location on the body. We use them for measuring, modeling, and for tracking motion or animating the body.

The landscape of current standards related to anthropometry and 3D body scanning is very diverse since they are developed by several international organizations and targeted at different applications (McDonald et al., 2018). The main focus of the standards is the definition of landmarks and body measurements: ISO 8559-1 (2017) and ASTM D5219-15 (2015) provide body measurement designations and definitions for garment construction, ISO 7250-1 (2017) for ergonomic design and ISAK for shape tracking in health, sports, and fitness. All these standards always refer to anatomical landmarks of the body to define the dimensions. The recent standards ISO 18825-1 (2016) and ISO 18825-2 (2016) provide measurement and landmark definitions for virtual models used in digital fashion. They are the first standards addressing the definition of body measurements, landmarks, body parts, and joint definitions from a natively digital perspective.

Whether we are measuring with manual tools or scanners, it is important to locate and pre-mark landmarks before scanning. It is important that pre-marked landmarks are visible in the scans for identification after scanning. ISO 20685 states that "Landmarks should be marked on the skin, and then identified with dots or other techniques that can be seen on the displayed image, and distinguished using the available software" (ISO 20685-1, 2018).

Some scanners that include images of the body surface can detect flat stickers of a contrasting color, such as those shown in Figure 2.10 placed before scanning. However, some scanners do not have the ability to see flat stickers or markers and others have areas of the surface that are at such an angle to the scanner that a flat mark on the surface cannot be seen. In these instances, the only way to "see" a landmark is to place a 3D volumetric marker such as a sphere or cube of known and calibrated size. Some examples of 3D markers as visible in a scan are also shown in Figure 2.10.

Volumetric markers make artificial bumps on the surface, so they need to be removed during the processing stage to avoid undesirable artifacts on the measuring extraction. It is recommended to use the minimal size for these markers to be more precise in the landmark location. 3D body scanners that include image capture of the surface are better for the measurement of physical markers. It is possible to use flat landmark stickers that do not perturb the body surface.

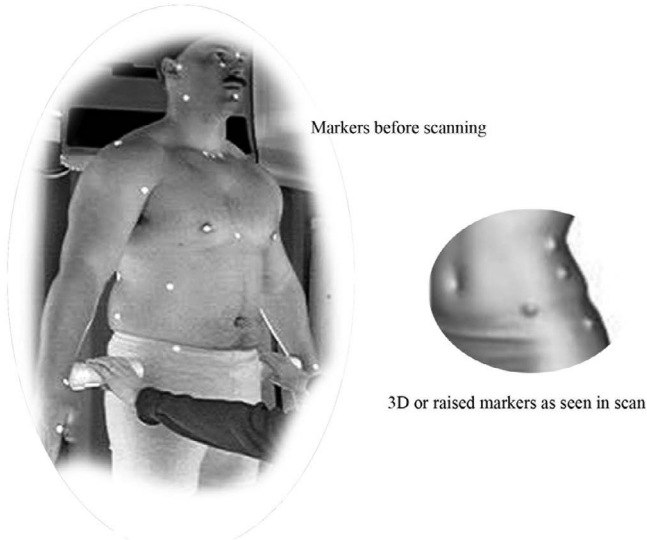


FIGURE 2.10 Markers for landmarks.

Identifying landmarks can be done manually, automatically, or a combination of the two. All methods require software tools for identifying landmarks and these tools do not necessarily come with the scanner. It is important to verify with the scanner manufacturer that landmark recognition tools are available and to test them to ensure they work effectively.

For example, for the CAESAR™ project, the 3D landmarks were identified manually using custom-developed landmarking software (Blackwell et al., 2002). The landmarks had been pre-identified with stickers before scanning and the landmarks were identifiable from surface images without having to create a watertight mesh. This process is illustrated in Figure 2.11.

The scanner, the Cyberware WBS™, had color cameras in addition to the data scanner. The color cameras took photos of the surface. The photos were aligned to the data file in the software tool, Integrate (Burnsides et al., 1996). This made the landmark stickers visible. Figure 2.11 shows the photos overlaid on the data file and the software tool process for identifying the 3D coordinates of a landmark. A person points, with a cursor, to a sticker that was placed on the subject before scanning and a window pops up with a closer view of the point. The viewer then selects the center of the point to identify the landmark location and moves to the next point. This manual "point-picking" process was used for the first surveys and the process was later automated.

Manual landmarking of the body prior to scanning, as was done for the CAESAR™ project (Robinette et al., 2002), can be time-consuming and is not feasible for a person who is self-scanning. Therefore, several scanner manufacturers provide software tools to automatically obtain a set of measurements from a scan. In this case, the landmarks are automatically calculated using geometric features of the body shape (e.g., prominent points, curvatures, and minimum/maximum circumferences).

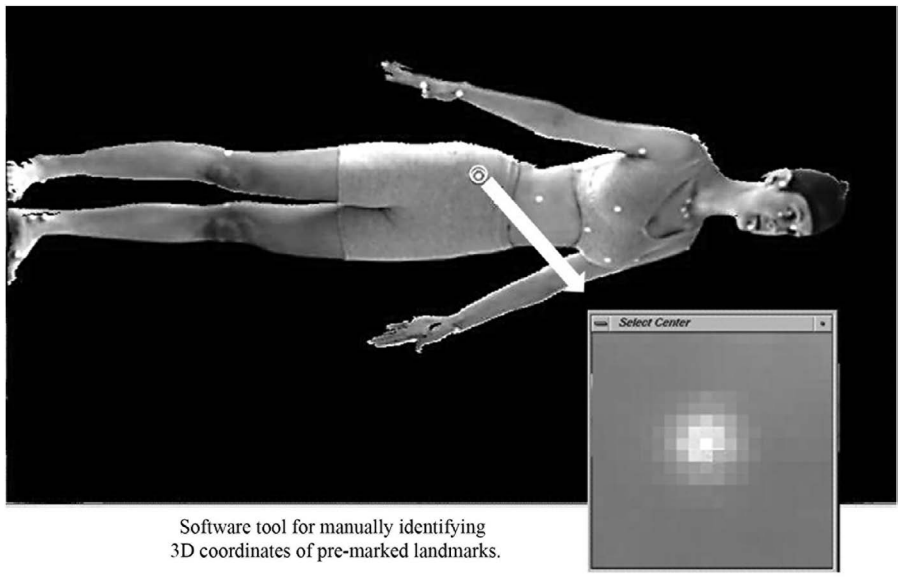


FIGURE 2.11 Manual landmarking directly from original scan.

Caution must be used with these tools because they produce less robust and accurate results. Some geometric features used to calculate the landmarks automatically are not consistent among different body types.

We illustrate this in Figure 2.12 where we show the automatically extracted minimum horizontal waist girth locations for two subjects as well as their preferred waist, which is where they prefer to wear their clothes. Subject 1 is a thin subject and the minimum horizontal girth of the torso is much higher than where she wears the waist of her garments. Subject 2 has a different torso shape such that his minimum torso girth is at his hip level. With automated measurement extraction, this could be the location of the waist girth that is returned automatically because without landmarks, the only direction the computer knows is horizontal.

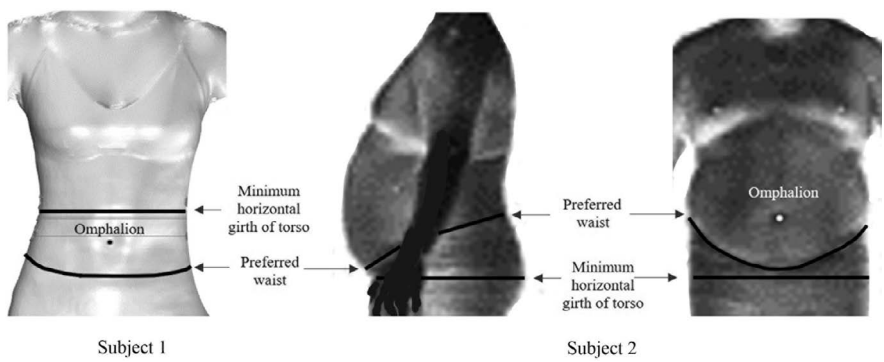


FIGURE 2.12 Inconsistency of automated waist measurement locations for two subjects.

Most scanner validations are only done on thin subjects, like subject 1, or on manikins that are proportioned like a thin or fit subject. As a result, the extracted measurement location and reliability assessment may not reflect the true location and reliability on all body types. This can lead to poor or missing data on some segments of the population and bias in the results.

The Omphalion landmark is visible in the scans, so it might be used as a landmark. It might provide a more consistent measurement, but it is not a reliable indicator of where the waist of the wearable will be worn. Many functional and relevant measurements for wearables are complex and horizontal and vertical directions are not sufficient. Yet, without multiple landmarks to guide the computer measurement, the only options are to specify horizontal and vertical as the measurement direction. In this example, we had a horizontal measurement, but the waist where we wear our clothes is rarely if ever horizontal. We typically prefer the waist of our clothes to be lower in the front than in the back. Even for subject 1, the measurement dips in front. For subject 2, the difference between the front and the back is even more extreme. It is important for wearable development to have identified and visible landmarks sufficient to determine the location of the wearable.

The errors due to the automatic processing of scans depend on several criteria including (1) the availability of landmarks that are visible without pre-marking, (2) location and direction definitions that hold up for all body shapes and sizes, and (3) the robustness and consistency of the algorithms. The definition of measurements of traditional anthropometry based on anatomy does not work for most automatic digital measurements.

Even if we have reliable pre-marked digital landmarks for the 3D body scan, it is still necessary to use mathematical functions to calculate measurements and, except for the shortest straight distance between two points, the translation of the anthropometric definitions to mathematical algorithms is not standardized. As a result, the interpretation to implement the definition of body measurements in algorithms varies among commercial systems and research studies.

All factors described earlier influence traditional and digital measurements in different ways producing a lack of compatibility between both methods (Han et al., 2010; Kouchi, 2014). This is confirmed by several published studies that compare traditional and digital measurements of a sample group of subjects being particularly relevant for the body contours (Kouchi, 2014; Kouchi & Mochimaru, 2005; Lu & Wang, 2010; Perkins et al., 2000).

A reliability study of anthropometric methods done by the IEEE 3D body processing group compared traditional manual methods and new 3D body scanning technologies. The study was done in two phases measuring in total 132 subjects by five different experts on manual anthropometry, four different body scanners, and four mobile applications (Ballester et al., 2020). The results indicated significant biases of up to several centimeters between pairs of 3D body scanners. This could be related to the different algorithms used by each 3D body scanner to calculate the body measurements.

Only direct straight-line distances between anatomical points marked by palpation for both traditional and digital measurements are comparable. Robinette and Daanen (2006) demonstrated that if subjects are pre-marked before scanning and the

measurements taken from the scan are point-to-point distances between these marks, then the 1D measurements can be accurate and reliable. The straight-line distance between two points is not affected by direction, surface contours, or hidden areas; therefore, they have the least amount of measurement error of all the measurement types, provided the landmarks are accurately placed and precisely identified in the scans. Any measurements that follow along the surface such as a Waist Circumference or a length that simulates one taken with a tape measure will not be as reliable.

Physical Fit Measurement Tools

Objective fit tools measure the location, stability, and performance effects of the human interface with the wearable in a way that does not require an opinion. All wearables are meant to be worn in a particular location on the body and for some wearables, this location is more specific and critical than others. For example, some have components that must be visible to the eyes, and some have components that must be in the right place with respect to the ear canal for the person to hear, and so on. For wearables with stringent location requirements, the wearable must be able to stay in that place and continue to function properly while the person is wearing it for the expected activities. We refer to this as stability. Performance means the wearer can do their tasks well and safely, without interference due to poor fit. For example, if a coverall is too loose or too tight, it might hinder reach and movement. Or, a protective suit may gap and expose the skin during movement.

Both stability and performance measurement require testing with the subjects moving or dynamic testing. There are a variety of tools and methods for this type of testing and we review a few here that have been used to provide some ideas for interface measurement. This is not an exhaustive list. It is merely intended to offer some suggestions. New wearable products can be unique, so specialized test methods may need to be developed.

Location measurement can be as simple as measuring with a ruler the amount of fabric offset from the body, as was shown in the fit-mapping manual by Choi et al. (2009). They showed how they pinched the sides of a coverall at the hip level and measured the width of the fabric in the pinch on each side. They called this amount the ease and suggested fit scores based on the range of ease, with a good fit having 2.5 to 4 inches of ease. The decision about how much ease is acceptable can be decided before testing or after. Regardless, it is best to record the actual measurement just in case we might change our minds about what is acceptable later. Other fit criteria may be more important, such as the ability to move or reach, that might change our minds about what is acceptable. For example, the acceptable fit might not look as pretty, but it might function better.

If we want to visualize and measure the location of the product when worn to ensure proper placement for performance, such as an optical display, then 3D scanning with and without the product in place may be essential. This is something that is usually not possible to measure reliably with calipers and tape measures. Examples of software tools for superimposing the scans or images include Blender™, PowerPoint™, or Photoshop™. For rigid items like helmets, accurate 3D scanning can enable us to measure, with precision, the location of key landmarks within the wearable, by overlaying the scans.

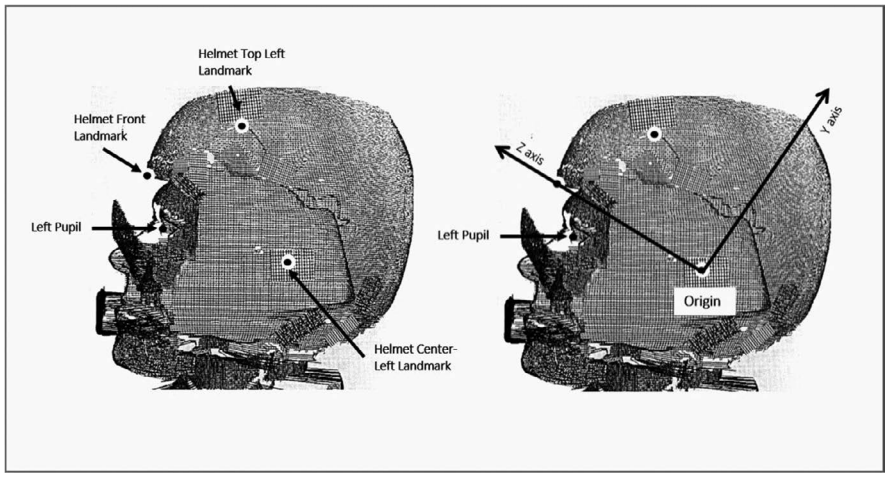


FIGURE 2.13 3D scan of subject in a helmet with helmet-based axis system.

Scanners or imaging systems that do not have precision, such as inexpensive handheld scanners or 2D digital photographs, do not have the precise measuring capability but can enable us to visualize the location of the product on each person and this has proven to be extremely useful for understanding complex fit issues.

Figure 2.13 illustrates a 3D pupil location measurement with respect to a helmet with a visual display. At the left is an image of the 3D scan of a subject with the helmet in place and his pupil landmark has been identified. The helmet also has five pre-marked landmarks, one in front and two on each side. These are used to put the data into a helmet-based axis system for measuring the locations of important landmarks. This axis system is shown at the right in the figure. The Helmet Center-Left and Helmet Center-Right landmarks define one helmet orientation axis (x). This axis is not visible because it is perpendicular to this 2D view, but it is located at the origin. A line from the Helmet Front landmark to the x-axis defines a second axis (z). The first three points together also define a plane, the xz plane. A third axis is defined as perpendicular to this plane and starting at the center of the left and right Helmet Center landmarks. The center point of the three axes is called the origin, and it is the point that has the value of zero for all directions, denoted (0,0,0).

The helmet-based axis system allows us to measure where each subject is located in the helmet. This can also help to track the independent movement of the wearable and the body or slippage if the scan is taken before and after physical activities. This works well for wearables that have some rigid components that are visible in a scan but not as well for soft wearables.

When combined with fit scores that identify areas of concern, these visualizations help us understand both underlying causes and solutions. For example, in Figure 2.14, we show the estimated distribution of right pupil locations in a helmet that was calculated by overlaying scans and measuring the locations for more than 100 subjects. The distribution is represented with two ellipses, one enclosing 95% of the subjects,

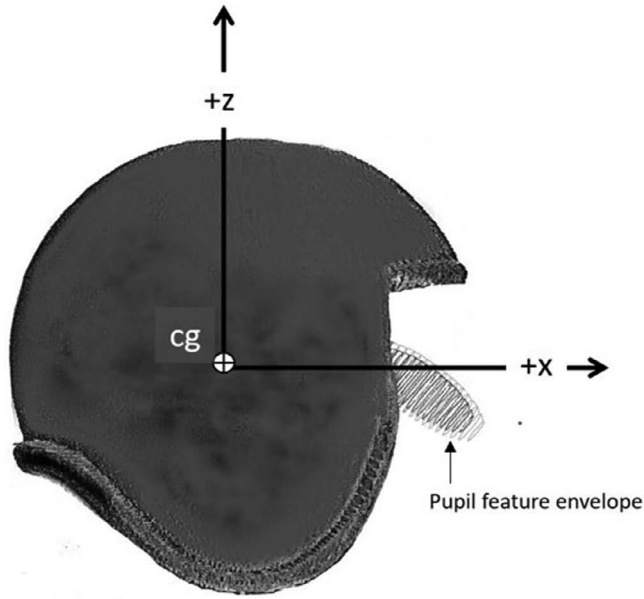


FIGURE 2.14 Range of pupil locations in a helmet.

and one enclosing 99%. These ellipses are referred to as feature envelopes, the feature being the right pupil in this example.

The overlay and recording of location in the helmet enabled us to see both the range and the location of the pupil variability which allowed us to design the adjustments or sizes more accurately. In this example, the pupil location was intended to be 2 inches below the edge roll but the true location ranged from 2 to 4 inches. We also see that there is an angle to the location which means helmet rotation is a factor. This was unexpected when we modeled this in CAD.

In Figure 2.15, we show a simple tool to measure the location of the bottom of the sleeve from the wrist landmark. It is a quick and standardized, repeatable way to identify and record if the sleeve is long, short, or just right.

Stability is more difficult to measure than location. The most common and often the most effective method is a subjective assessment and a questionnaire. This can be assessed by asking subjects questions about how much it moved and recording responses on a scale or by having the investigator score the movement or slippage. For this method, we need a prescribed set of movements or activities for each subject to perform. The activities should be related to whatever activities they would be expected to have while wearing the wearable.

An alternative is to devise some method to measure how much the product moves during activities. In their report on the fit assessment of three helmet systems (Blackwell & Robinette, 1993), the authors used two stability assessment methods, one subjective with a questionnaire and one objective with a measurement of the movement during activity. For the subjective assessment, they had the subject do a set of movements and the investigator scored the movement on a scale of 1 to 5,

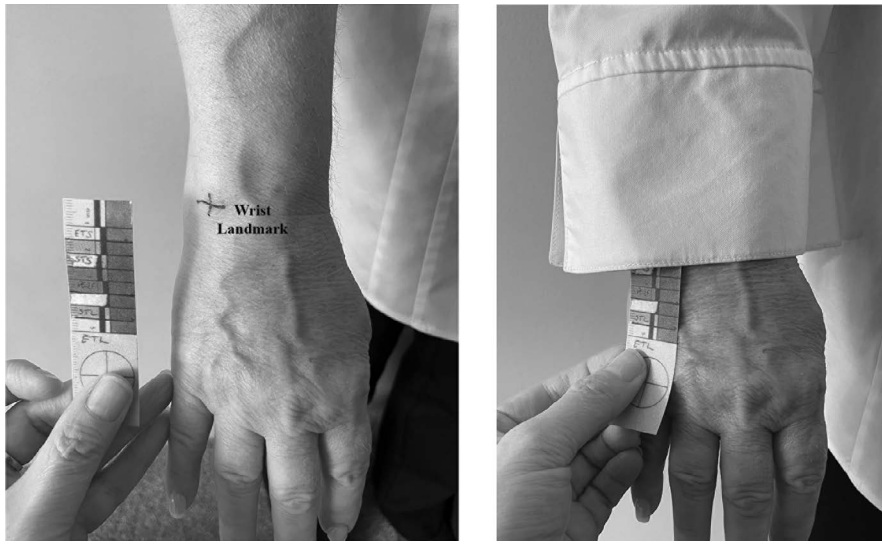


FIGURE 2.15 Garment length location tool for sleeve length.

with 1 representing the lack of movement and 5 representing excessive movement. For the objective assessment, they attached a ballpoint pen to the helmet in such a way that it would draw on the subject's forehead as the helmet moved around during the same movements. They scored the movement in 5 mm increments, from good (1–5 mm) to poor (20 mm or more). The helmet would re-seat itself to its original position when the subject stopped moving such that having the ink path showing how much it travelled during movement was more effective than simply measuring before and after movement.

Motion capture systems and 4D measurement systems have become more precise and capable than they were in the 1990s and can provide a good way to measure stability for some products. It will depend on the product, the way it is worn, and the visibility of both body surface and wearable landmarks. If the appropriate body and wearable landmarks are visible during movement, these tools might be an alternative to consider, particularly if they are also suitable for performance measurement of the product so that added expense can be justified.

Performance measurement requires us to measure whether the subject can do the activities they need to do well and safely with the wearable performing as intended. This is very dependent on the wearable. For example, we might measure if the mask seals out toxins by measuring the ppm of particles that get inside a chemical protective mask while the subject does physical exercises. This would require a special measuring chamber as was used for the testing of the MCU-2P mask (Case et al., 1989). We might measure if the glove allows us to do our job using a series of dexterity tests related to the job as was done for comparing chemical defense gloves (Robinette et al., 1986). We might measure if the full display is visible by having a subject read parts of the display. We might measure if a sensor is placed properly, if we can detect the oxygenation level.

Performance-based measurement includes the measurement of ability with the wearable, such as if a subject can perform the necessary tasks like reaching, bending to pick something up, or driving a car. Choi et al. have some good examples for assessing mobility or the ability to move and perform tasks (Choi et al., 2009). They describe a motion or a task and list the pass/fail criteria as a three-level assessment: pass, pass with difficulty, or fail.

Reach and grip strength have been included in anthropometric measurement surveys such as the survey of Army Women in the 1970s (Laubach et al., 1977). These measurements can be taken with and without the wearable to determine if the wearable interferes with reach or strength abilities.

Personnel and Facilities

Before we begin the design process and the iterative testing required, we need to acquire, train, and assess the personnel we need and acquire facilities that are safe, secure, equipped, maintainable, and accessible. There are several aspects of accessibility. First, both our physical spaces and our software tools must be accessible to people with disabilities. Second, we must have the proper clearances and approvals to use the facilities and conduct testing with human subjects. Finally, we must have a shared data space that is accessible to all members of the team and properly maintained for long-term use.

Personnel

Personnel, who are skilled and trained for data collection, analysis, and management are needed for good outcomes of these processes. It is usually best if we have more than one person doing data collection and testing, and it is a good idea to have one person who functions as the team lead who understands the whole process. When we have a team of people, some of them can specialize in certain areas, such as data collection or data processing, and not have to be trained and skilled at everything. The team members' roles include the following:

- Team lead/project manager: This person coordinates the project, manages the team, and solves problems as they arise. This can include presenting proposals and results to company management, conducting quality assurance audits, answering detailed questions, making decisions when there is debate, and collecting body scans, photographs, or other information to illustrate and document fit issues.
- Trainer/senior anthropometrist: If the individual team members have not had anthropometric training, this person would be the individual who provides the training. This is someone whose measurement experience qualifies them to be the standard against whom other measurers on the team are compared during the team training.
- Measurer/recorder: This is a person who measures and/or records body measurements and demographics.
- Fit assessor: This person conducts the fit assessment on the subject. This may involve body scanning with and without garments.

- Greeter/logistics: This is someone who greets the subject, explains the project, and ensures the consent forms are completed. In addition, they will manage the flow of subjects to the measuring teams.
- Analyst: This is someone who has experience in statistics and analysis of data.

This list is not all inclusive and sometimes, one person might perform more than one role. For anthropometric measurement, data collection is best to plan to have at least two investigators working together because the measuring is faster and more reliable. For example, for manual measuring, the placement of the tools with respect to the subject can be viewed from two directions at once making it more precise. Also, one person can be measuring, while the other records the values and gets tools ready for the next measurement. This can cut data collection time in half.

If the data collection will be ongoing for several weeks, it is valuable to have trained backup data collectors. It is difficult enough to get subjects to show up without having to cancel appointments because a team member is ill or on leave. If all team members are trained for all data collection types and roles, it enables people to trade-off jobs to stay fresh or take breaks.

Data analysis can be done by someone who is not part of data collection, but they will need to understand how the data are collected. It is also helpful if the data collection team members understand a bit about how the data will be analyzed. That will help them make better decisions when issues arise.

Ideally, the designer or engineer for the product should be the fit assessor or at a minimum be involved in the fit assessment. If not possible, they should at least be present for the pilot testing as this is where the main learning about the product occurs.

All data collection team members should be trained and practiced for three reasons: (1) to learn how to interact with subjects, (2) to collect high good quality data, and (3) to save time and money on testing. Trained teams can collect data quickly, accurately, and with a smooth transition to analysis meaning reliable results will be ready rapidly.

There are two types of training: individual investigator training and team training. Individual training brings each team member up to the level of training they need for their role and team training coordinates the methods for the whole team.

Individual training includes things such as anthropometry training, training on the software tools that will be used, training on research with human subjects, training on data structuring for analysis, training on fit assessment, and training in the relevant statistical procedures and software tools. Individual training can take several weeks or months, depending on the skill and experience of the team members, so this needs to be factored into planning. However, it will not need to be done for every test. It will only need to be updated if some of the tools to be used are upgraded or changed. Individual training both minimizes the individual data collection error called intra-observer error and speeds up the data collection process enabling quality data with minimal time.

Team training is done with two or more team members at a time after the individuals have completed their baseline training. It is intended to ensure that they are measuring, scoring, and recording the same and to optimize the flow of the test. This training minimizes the error between measurers, called inter-observer error. It can usually be done in two to three days.

It is impossible to get the exact same measurement each time we measure, but we can minimize the error with practice. Human beings are constantly fluctuating, so they are a moving target for measurement. For example, when we breathe, our torso moves in and out, so measurements there get larger and smaller. When we get up in the morning, we are taller than when we go to bed at night because our spine compresses gradually throughout the day. When we stand, we do not stand still but we sway to maintain our posture. We lose and gain weight during the day and our blood pressure goes up and down. All these things and more can affect the measurement values and the impact can be minimized with training and measurement standardization.

The amount of error that we, as individuals, can expect to have, regardless of how good and practiced we are, differs depending on the measurement. Larger measurements, such as stature, can be expected to have larger error than smaller measurements, such as Hand Length. In addition, different measurers (also called observers) will take measurements differently for a variety of reasons such as different amounts of pressure applied or different interpretations of the measurement locations. Following good protocols and training the measuring team properly is crucial to minimizing errors.

A good anthropometry training program should include training in calibration techniques, anatomy terminology, anatomy of underlying bony landmarks, proper positioning of subjects, use of different measuring tools, proper placement of tools on the body, and assessment of inter- and intra-observer measurement reliability for live subjects. Informal training that does not include these things is not enough. The larger the error, the more subjects we need to make good decisions, so training can reduce the cost of fit testing in the long run.

Anthropometry is used in many fields with different purposes such as sports science, biological anthropology, archaeology, health, and engineering. Some measurements are unique to the field. For example, sports science has a focus on human movement and sports performance, so they are interested in measuring biceps, skinfolds, and movement angles. The health field is interested in measurements like Waist Circumference or body mass index (BMI) in relation to say, cardiovascular risk factors. What all the fields share are techniques in *how* to measure, even if the measurements themselves and applications are different. Their training programs can provide good training in anatomy, landmarks, the use of tools, body positioning, and dealing with human subjects. However, training on the measurements and tools specific to the design and engineering of a specific type of product may need to be added to the training curriculum. Sometimes this additional, product specific training can be done as part of team training if the individual training provides a good anatomical base.

Good training can be obtained from any experienced anthropometrist who has a high level of anatomical knowledge and has been involved in anthropometric surveys. The organization ISAK (International Standard for Anthropometric Assessment) offers anthropometry training, as does Anthrotech Inc. (https://anthrotech.net/).

The training program designed by ISAK includes the theory and practice for gathering anthropometric measurements for sports and health applications. ISAK courses are narrowly focusing on the sports branch of anthropometry, but they provide good descriptions and knowledge of measurement instruments, maintenance,

and calibration of manual tools; how to hold, operate, and read the instruments; how to reliably find landmarks, a good description of landmarks; and measurement quality assurance. The practice is done in each course to demonstrate inter and intra-observer is a good exercise to find bias or errors in locating landmarks or using the instruments and to implement actions to establish good and consistent measuring criteria. However, ISAK courses lack training in many of the measurements needed for product design.

Anthrotech Inc. also offers quality anthropometry training, but it is specifically aimed at product development and evaluation. Their standard training set will probably be more useful as a result. However, no basic training set can include all possible measurements, and for new types of products, additional measurements will need to be added. Either organization will provide good basic training such that you would know enough for learning or developing any additional measurements more relevant to your product.

There are some good resources for measurement descriptions to aid in creating a measurement set for your product. Studies that were funded by the U.S. Government are publicly and freely available from the DTIC. These include the following:

- Anthropometry of Women of the U.S. Army 1977 (Laubach et al., 1977)
- Anthropometry of Air Force Women 1968 (Clauser et al., 1972)
- Anthropometry of U.S. Army personnel 1988 (Gordon et al., 1989)
- Civilian American and European Anthropometry Survey 2002 (Blackwell et al., 2002)
- Mass distribution studies (McConville et al., 1980)

These studies formed the basis for many other anthropometric studies around the world and the documents contain not only measurement methods but also descriptions of data collection processes and analysis methods. There are also some standard measurement descriptions and definitions that are available for purchase such as ISO 20685 and ISO 8559.

If the product is new or has specific fit or function requirements, it may require measurements that are not listed in the above sources. In these instances, the way the measurement is taken should be documented including the instrument, the subject's posture, garments worn, landmarks, and how the measurement is taken. It is best to give it a unique name, so it is not listed as the same name but be a different measurement from an existing measurement.

There have been several studies regarding measuring errors relevant to design and engineering. For the Army ANSUR surveys, they called acceptable measurement repeatability "allowable error" (Gordon & Bradtmiller, 1992; Gordon et al., 1989, 2014). This is an umbrella term for the overall tolerance for errors from any source. These referenced documents are available for free from the DTIC (https://discover.dtic.mil). The journal article by Gordon and Bradtmiller (1992) discusses inter-observer error (the differences due to having different people do the measurements) in detail.

Sports science often requires intra-observer repeated measures during testing for various reasons, but this is usually because the repeatability of the measurement

is poor, like skin folds. If after training, the measurement is reliable, then having repeated measures during the study does not substantially improve its reliability, and so they are not worth the expense. Thus, for design and engineering purposes, repeated measures are needed during training but are not usually necessary in the test or data collection study itself.

Hotzman et al. (2011) build upon the work of Gordon et al. (1989) and present allowable error values for a list of body measurements. They used the allowable error values for measurer training as well as to track or "recalibrate" during data collection. Tracking measurement repeatability was done because data collection lasted several months and the team traveled to different geographic locations. Most individual fit tests are much shorter in duration and should not require checking repeatability after the initial training period. However, if there are many fit tests and there are gaps of weeks or months between tests, it is a good idea to have a *training refresher* before the test starts, particularly if the different tests will be compared.

The key sources for measurement error within and between measurers are as follows:

- Obtaining and maintaining the subject's posture
- Landmark definition and landmark location
- Positioning tools and measurer

Posture can also be called pose or position of the subject. Each 1D measurement is taken with the subject in a single posture or pose. The posture of the subject is one of the aspects that most influence anthropometric measurements and is usually the largest source of measuring error. It is important to understand what posture was used in a study and to copy the posture exactly if collecting new data that will be compared to a study.

When a posture is difficult to assume or maintain while measuring, it requires vigilance on the part of the measurer to ensure the measurement is taken with the subject in the correct position. For example, many people tend to slump and standing or sitting erect can be difficult to maintain. This variation in posture can have large effects on the resulting measurements and in the variation with repeated measuring. This is true regardless of the measuring tools used. For example, if a person is standing erect in one 3D scan and slumped in another, their measurements will be very different. Also, if one person is scanned standing erect but another is slumped, they might have very similar measurements, but their measurements will appear to be very different. Therefore, it is very important to include careful positioning and repeated measurements in the measuring team training.

The location of anatomical points presents repeatability errors due to the skills and training of the measurer. The difference between the measurements taken on the same subject by the same measurer is known as intra-observer error and is a random error (Kouchi, 2014). This type of error increases the STD of the sample (Kouchi et al., 1996). Measurer training should begin by learning some anatomical terminology. Human Osteology textbooks are good sources for this (Bass, 1971; White et al., 2012), as they explain the reasons behind the landmark names and anatomical directions.

At this point, one of the frequent questions is: what are the allowable intra and inter-observer errors? Especially with a new measurement. When estimating the allowable error, you can get a rough idea from a similar measurement. These can be looked up in ISO 20685. In the case of inter-observer error, the values usually referred to in the state-of-the-art are those reported by Gordon et al. (1989) who evaluated inter-observer measurement errors of four experienced anthropometrists using the marks of one technician. However, in a real situation, each measurer used its own marks and takes its own measurements. In practice, inter-observer errors are slightly higher than these reference values.

The influence of landmarking errors on manual anthropometry was quantified by Kouchi (2014) for both intra and inter-observer errors. The study was done measuring 40 subjects, which is a recommended sample size to perform a pilot study of the measuring protocols. The results of the intra-observer mean absolute difference (MAD) show that 4 out of 34 anthropometric measurements are higher than the maximum allowable error of ISO 20685. The ranges of intra-observer MAD reported are 4.4–8.5 mm for heights, 3.5–7.4 mm for breadths, 5.4–13.9 mm for circumferences (Waist Circumference being the maximum MAD value), and 1.5–4.2 mm for small measurements (e.g., hand length, head length, and foot breath). These ranges are a good reference to assess the repeatability of traditional anthropometric protocols. With good training and systematic in following the protocols, inter-observer error can be minimized until a certain point. It is important to take into consideration that we are not measuring an object. We are measuring an articulated body consequently due to the posture variations and breathing, repeated measurements always will vary.

In the case of anthropometric studies or fit testing experiments that include a large number of subjects, the participation of several measurers is very common. When different technicians are gathering the anthropometric measurements, there are also systematic biases between measurers that also affect the variance of the anthropometric measurements. Thus, the difference between the measurements taken by two different measurers, the inter-observer error, depends on the magnitude of the bias and the magnitude of the intra-observer error (Kouchi, 2014). This means that the inter-observer error will be larger than the intra-observer error because it includes the bias plus the intra-observer error. The bias between measurers is mainly due to differences in the interpretation and execution of the measurement protocol and that underscores why training is important.

In general, when a new test that includes anthropometric measurements is planned, it is recommended:

1. To define and document the measurement protocols
2. That training is led by a criterion anthropometrist
3. To train the measurement team using the documented protocols
4. To perform a reliability study to determine the repeatability and inter-observer errors
5. To update the protocol, documentations, and/or the criteria of the trained team if some deviation is detected in the reliability study before commencement of the study

Common mistakes *to avoid* when conducting your survey include the following:

1. Deciding to adapt the protocols during the design loop
2. Changing a definition of a body measurement, or
3. Changing the criteria to landmark the body

These modifications during or after data collection introduce serious difficulties when comparing data among successive studies performed during the product development loop. They create issues tracking the progress of the product design.

Allowable error should be below the tolerance of specifications related to the applications of the data, the type of product, material, and fitting requirements. These specifications should be defined at the first stage of the design process.

Facilities

For smooth and timely testing, there are important logistics of the test site, people involved, and product availability to be considered. It is a good idea to document these in a data collection plan before beginning testing to serve as a guide, to track changes, and to document for future reference. Four of the most important categories of considerations are discussed in this section. These are as follows:

- Use of human subjects
- Test site considerations
- Subject attire
- Product information record

We have found it is best to have a dry run pilot test of the data collection process prior to the first test as well. If there are any problems, it is best to find them before testing starts. This can be done as part of team training or as part of a test of the COF. It does not necessarily have to be repeated each time a new test is done unless some of the procedures have changed.

Use of Human Subjects

Good practices when using human subjects not only ensure the privacy, health, and safety of the subjects but also protect the interest of the organization doing the testing. Different countries have different laws and regulations so be sure to review the latest in any country in which data collection will occur.

Whether it is required by law or not, it is a good idea to have informed consent. Informing each subject about what they will do and what will happen in the study before beginning and getting them to sign an informed consent form ensures that we know of any potential issues an individual might have before we start, reassures the subject, and reduces the risk of litigation later. For example, while we were explaining the fit test procedures to a subject for a fit test of a coverall, we learned she had been exposed to poison ivy, so she had particularly sensitive skin that would impact her fit scores. We spared her the discomfort and avoided biased scores by postponing her participation.

Subjects will want to know that their privacy will be protected before they will consent. It is a good idea to have in place and explain how we will protect their privacy for things such as their image, their age, their health status, and their weight. One way to protect data privacy is to separate their name and identifiers from their other data. This can be done by assigning them a subject number and keeping a separate and protected file that indicates which number they were assigned.

Images can be handled either by getting consent to use and publish their image including their 3D scan or by blurring the image such that it will not be recognizable. In some countries, informed consent is not sufficient.

Test subjects need to be treated with respect and be helped to feel relaxed. Explaining everything that will happen ahead of time helps. It is also important to keep an eye on the subject during the process. Some people get nervous enough to feel faint and pass out while they are being measured. If someone seems to wobble, offer them a chair, and have them sit until they feel better.

Test Site Considerations

Data collection can occur in many different locations and not always in our laboratory. To attract people from different TPs, we often must go where they are located and find and use facilities far removed from our home office. Even at our home office, if fit testing is a new endeavor for a company, sometimes we must find a suitable place. Here are some tips for finding, creating, and using a suitable space.

First, the space must be large enough to contain all of the tools, prototypes in all sizes, furniture, and all the people who will need to be in each area of the space at one time. The furniture should include a chair for each subject being tested or waiting to be tested and each investigator. The space should be partitioned such that the test subjects are not able to see each other while they are being tested. This can lead to bias and influence in the results.

It is best if there is a meet and greet area separated from the test area for meeting new subjects as they arrive, explaining the test procedures, and having them fill out any forms. This keeps the new subjects from being influenced by subjects who are being tested and prevents disturbing the ongoing testing. It is also a good place to keep control data collection logistics. For example, Volume II of the CAESAR™ study report describes the tasks at their meet and greet station, which they referred to as the "Demographics station" (Blackwell et al., 2002). These included the following:

- Greet the volunteer and cross their name off the schedule
- In the *Subject Log,* record the date, subject number, volunteer name, and arrival time
- Brief volunteer on CAESAR project and their role as a participant
- Have volunteer sign and date the *Informed Consent* form
- Assign a volunteer number and hand out the *Demographic Questionnaire,* the *Traditional Measurement Data* form, and accompanying 3.5″ floppy disk
- Confirm that volunteer has correctly completed the paperwork (the demographic questionnaire and the top section of the measurement data form)

- Select the appropriate size of shorts (and top also for a female volunteer) based on (1) the specifications of the size selection chart and (2) the judgment of the demographic station team member
- Hand shorts (and top, if applicable) to each volunteer, along with a white laboratory coat
- Instruct each volunteer to don the laboratory coat before walking from one data collection station to another
- Provide the volunteer a rubber tub and lid for the discreet containment of personal belonging
- Escort volunteers to unoccupied changing rooms
- Input data from the completed demographic data form into the demographic database while the volunteer changes from street clothes to the CAESAR garments
- Escort volunteers from changing room to traditional anthropometry measuring station to begin the data collection process
- Collect data forms and floppy disks from volunteers who have completed the data collection process ("Departing Subjects")
- Hand appropriate rubber tubs to Departing Subjects; direct Departing Subjects to unoccupied changing rooms to change back into street clothes
- Collect used laboratory coats, scanning garments, and rubber tubs from Departing Subjects
- Record departure time for each volunteer in the *Subject Log*

There were also other duties that were sometimes required of the "greeter" such as sending out appointment reminders to potential subjects, ordering supplies, and reporting back to home base about status and issues.

If the product requires the subjects to change clothes, there should be a private changing area near the testing location. Depending on the product, it may also be necessary for the subjects to be able to walk between the changing area and the testing location without being observed by other test subjects.

If measurements are going to be taken over top apparel or equipment a clear definition of these items, how they will be worn and how the measurements will be done should be part of the COF. The apparel on the measurement should be the apparel should be considered and evaluation. For example, it may be important to have garments that are close enough to the body to reproduce the body measurement but not so tight that it changes the body contours.

For women, the design of the bra has a relevant effect on bust-related measurements. It is important to define a clear criterion for the type of bra selected according to the purpose of the study.

In the case of using standard attire provided by the measurer, it is important to select properly the correct size. For instance, a small size of a short can overpress the waist generating skin folds that are not present in the shape of the body surface if a proper size is used.

If the head is going to be measured, it is important to consider how the hair and hairstyle will be controlled. For example, will long hair be tied or allowed to hang loose? The important thing is to get consistent measurements for different people.

For 3D scanning, we might want to have the hair covered with a cap. Nylon wig caps have worked well for head scanning. If traditional head measurements will be taken, the calipers and anthropometers can measure through the hair, so a wig cap might make the measurement less accurate.

There should be storage areas within the test area for all prototypes, testing materials, and tools and for the subjects to place their belongings while they are being tested. The place for their belongings should either be within their sight or somehow secured.

Room temperature and lighting can influence subjects, so it is important that the levels are considered before starting testing. They should reflect the temperature and light levels of the environment in which the wearable will be used.

It is often necessary or useful to have several different data collection stations. Hotzman et al. (2011) provide nice descriptions of several different measuring stations used for anthropometric surveys for the U.S. Army and Marine Corps. By using different stations, many subjects can be processed at the same time and the data collection process goes much faster, as the team at each measuring station has fewer things to remember and they become proficient at their station very quickly.

There should be facilities and equipment to keep the tools and spaces clean. Measuring tools should be cleaned after each use. If apparel items are provided and will be re-used, then there should be a plan for laundering the items.

TARGET POPULATION SAMPLING AND PLANNING

A sample in statistics is a set of observations or items drawn from a population. It is usually a subset of a population taken in a limited time frame that is used to estimate the larger population. Collecting a large enough sample, with sufficient representation of important categories is one way we manage and reduce measurement uncertainty. A TP sample consists of the people collected (subjects) and the measurements or information gathered about each person.

Three different types of anthropometry and demographics samples representing the TP are collected during the design process: (1) the starting TP sample, (2) the fit test samples, and (3) the full TP sample. They are described in Table 2.11.

First, a sample is used to select and assess cases or fit models to design and mock-up or prototype the product as well as check the range of variability of the measurements. This sample is the first approximation of the TP and we refer to this as the starting TP sample.

Second, samples are collected to test the fit and performance of the product during the design loop. Many different samples are typically selected during this stage and we refer to these as fit test samples.

Third, a sample from the TP will be used to determine which sizes to produce, how many of each size to purchase, and how to help users find the correct size (size prediction). We refer to this as the full TP sample and it is typically weighted to better match the TP.

The starting TP sample and the full TP sample are similar if we have raw data and a large enough sample with all the measurements needed in the starting TP sample. This might be the case for products with similar predecessor products. For

TABLE 2.11
Three Types of Samples

Design Phase/Stage	Anthropometry and Demographics Samples
Select cases to initiate design and mock-up or prototype the product	**Starting TP sample:** The purpose is to select cases, determine the size location of cases selected, and check the variability of the main anthropometric dimensions. The sample should have at least 250 subjects. It does not typically contain fit data. It can be: • Raw data/individual data (best option) • Aggregated data/summary statistics
Fit and performance testing of the product	**Fit test samples:** The purpose is to test fit and performance of prototypes throughout development. Fit test samples are always raw data samples and include fit and performance along with anthropometry and demographics. Usually, many small test samples are collected ($n \leq 30$) with a large sample ($n \geq 100$) as the last prototype fit test.
Size cost/benefit analysis, size prediction	**Full TP sample:** The purpose is to determine how many of each size to produce and/or sell. It is the largest sample and must be raw anthropometric data that can adequately represent the target population for all important subgroups. It can also contain fit data. It is usually a combination of data from three sources: • The starting sample if it is raw data • A database created from the test samples • A new survey collected

new products that have no similar predecessors, this would be rare. In addition, as we do fit tests, we will learn about important measurements that we may not have had in the starting sample but we need for the full TP sample. There will be many tests after we start, so the starting TP sample does not need to be as accurate, as large, nor as complete as the full TP sample.

In this chapter about getting started, we discuss the starting TP sample and the planning for the full TP sample, because these are things needed before selecting cases or conducting design or sizing loop testing. Fit test samples will be discussed in Chapter 4.

Starting TP Sample

The purpose of the starting TP sample (as noted in Table 2.11) is to help select good cases around which to create our prototypes. Some companies that have similar predecessor products may have a TP sample prior to beginning the design process. However, it is more common for the starting TP sample to have been collected by someone else from a similar population and used as a rough estimate to get started. There are a variety of sources for such data. The WEAR Association (https://www.bodysizeshape.com/) provides a number of datasets from around the world to members for free and several 3D scan datasets can be purchased on storage devices that are shipped. The U.S. Centers for Disease Control (CDC) conducts regular anthropometry and demographic data surveys as part of the ongoing National Health and Nutritional Examination Survey (NHANES). Raw anthropometry and demographic

data from the U.S. population can be downloaded at (https://www.cdc.gov/nchs/nhanes/). These kinds of resources provide a variety of starting samples that can be used to select cases.

Normally it is impossible to find a perfect sample at the start of product development, particularly for completely new products. Luckily, at this stage, it does not have to be perfect, because the rest of the design process will refine the product. The starting TP sample just must be close enough to have some confidence that the starting case(s) are reasonable.

Because a sample is not the entire population, it is important to capture important aspects of population variability relevant to the product when gathering or selecting the sample. Demographics that are known to affect variability include the following:

- Gender
- Age
- Racial group
- Geographic sales region
- Occupation

For example, one might think women are just smaller than men, but this is not always true. For some measurements, such as Hip Circumference, women are larger than men on average. This is the case in the North American sample from the CAESAR™ survey (Blackwell et al., 2002) as can be seen in Table 2.12. Women are smaller for Shoulder Breadth and Stature but larger for Hip Circumference.

The combination of smaller and larger proportions means the proportioning is different, so a simple small or large scaling may not accommodate both genders. This is illustrated in Figure 2.16 with bivariate plots of Shoulder Breadth and Hip Circumference for males and females. There is very little overlap of the two genders for this measurement combination. The overall mean indicated with a square, falls between the two and does not hit the greatest concentration of people for either gender. If the overall mean is used to create the base size and scaled up and down very few people of either gender would be accommodated. For this reason, it is important to have adequate representation from people of each gender in all sample types if the product is intended for both genders and to analyze results for each gender separately.

TABLE 2.12
Means and Standard Deviations (STD) From CAESAR™ North American Sample

Name of Statistic	Males	Female	Difference
Hip Circumference maximum – Mean (mm)	1045.9	1053.7	7.8
Hip Circumference maximum – STD (mm)	97.6	124.5	26.9
Shoulder (Bideltoid) Breadth – Mean (mm)	495.8	429.8	–66.0
Shoulder (Bideltoid) Breadth – STD (mm)	36.1	35.0	–1.2
Stature maximum – Mean (mm)	1777.4	1639.8	–137.7
Stature – STD (mm)	78.9	73.3	–5.6

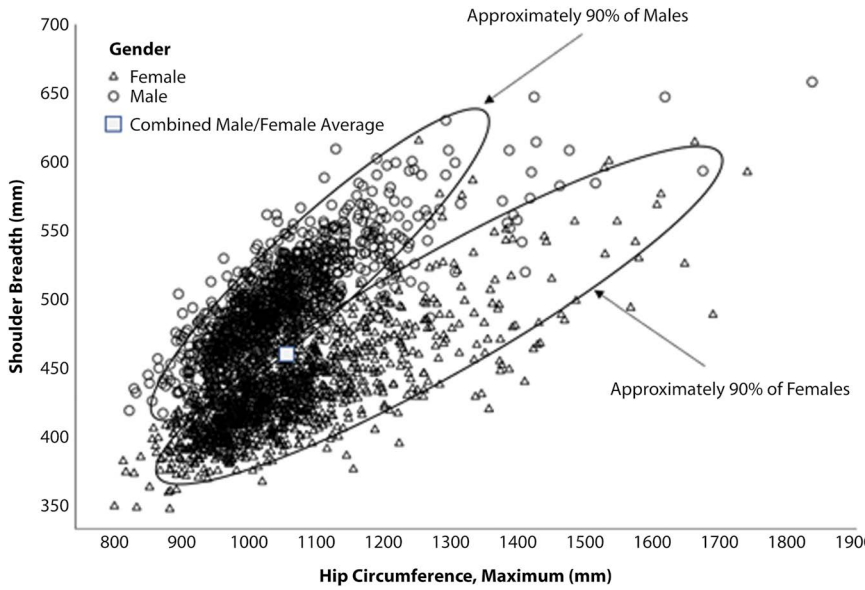


FIGURE 2.16 Shoulder Breadth by Hip Circumference, Maximum showing male and female subjects

The same may be true for different racial groups as well. To ensure all racial groups, get an equal quality or level of fit, it is necessary to evaluate the fit with sufficient samples from each racial group in the TP. This does not necessarily mean that there are separate size ranges for different racial groups. The ability to equally accommodate the different racial groups within a single size range should be determined through fit testing. For example, during the fit testing of a jacket for Navy women (Mellian et al., 1991; Robinette et al., 1991), we found that Black women frequently had a poor fit in the waist and hip area, requiring them to have major alterations (at their own expense) more often than White women. Evaluation of the data revealed that the issue was a difference in the height/width ratio of the torso versus the length of the arms. Black women tended to have shorter torsos but the same Hip Circumference as White women, while at the same time having longer arms and similar statures. The shorter torso meant the widest part of the hip was falling higher up in the jacket making it bunch up around the waist and gap at the back vent. They needed a shorter jacket in the torso, but the current short jacket had sleeves that were too short. This was resolved by lengthening the sleeves for the short jacket sizes. The sleeves were always hemmed after purchase anyway. By evaluating the fit for both racial groups we were able to achieve an equally good fit for everyone without having to have separate size ranges.

The ideal starting TP sample would be one recently collected and randomly sampled from the exact TP of interest and including all relevant measurements and demographics. However, unless the business has existed for a while making similar products and have been collecting this data, it is unlikely such a sample will be found. Until we do fit testing, we might not know what measurements are relevant,

so we cannot know what to measure exactly in advance. Therefore, we look for a sample that:

- Is drawn from a similar population or one that can be tailorable to estimate our TP
- Contains some of the most important measurements relevant to our product
- Has 250 or more people of each relevant gender
- Has at least 30 or more people from each relevant subgroup within a gender (age, ethnicity, etc.)

The sample size of 250 was arrived at from a study done by Churchill and McConville (1976) in which they evaluated sample size based on the ability to get accurate anthropometric measurement estimates. They used the data from the 1967 survey of Air Force males (Grunhofer & Kroh, 1975) that had 187 measurements and found that with just 250 subjects, 93% of the measurements would meet their accuracy criteria. Since this sample was only males, they examined the results from a survey of Air Force Women (Daniels et al., 1953) and found similar results.

The accuracy criteria Churchill and McConville used were primarily based on the amount of variation an individual human being has within the span of one day. Since, the authors note, "…any design which requires design values more precise than the variation which occurs in a human being within a normal day is, ordinarily at least, unrealistic".

The sample size of 30 for the within gender subgroups is based upon the knowledge that at n = 30 the distribution of the mean becomes closely approximated by the normal distribution, the z distribution. Estimates of the mean are less influenced by any one subject at this point. We might want more than 30 if we think we will have a focus on the subgroup for design or size evaluation at some point.

For some samples, a goal is set for each subgroup and random samples are collected from each subgroup. This is called a stratified random sample. Sampling males and females separately is a stratified random sample with the two genders being two strata. Stratum is a synonym for group and strata is a synonym for groups (the plural). Similar stratified sampling may be used for other subgroups.

For the CAESAR™ anthropometric survey, a stratified sampling plan was used with additional strata (Robinette et al., 2002). First, there were country strata. The civilian populations of three countries were sampled to characterize the population of NATO countries. The United States was chosen because it has the largest and the most diverse population in NATO. The Netherlands was chosen because it has the tallest population in NATO, and Italy was chosen because it has one of the shortest populations in NATO. The populations within these countries were sampled by age, race, and gender strata. Stratified sampling goals were used with equal sample size in each strata or cell according to the recommendations of ISO/DIS 15535. The strata (groups) are shown in Table 2.13.

Stature was used to estimate the within-strata sample sizes. A review of within-age-group STDs measured around the world indicated that 70 mm was a reasonable within-cell STD estimate for stature. The desired within-cell accuracy or tolerance

TABLE 2.13
Strata Used in the CAESAR™ Survey

North America	The Netherlands	Italy
Age groups: 18–29, 30–44, 45–65	Age groups: 18–29, 30–44, 45–65	Age groups: 18–29, 30–44, 45–65
Gender groups: Male and female	Gender groups: Male and female	Gender groups: Male and female
Ethnic groups: White, Black, and other	Ethnic groups: White and other	Ethnic groups: White and other
Total number of groups: $3 \times 2 \times 3 = 18$	Total number of groups: $3 \times 2 \times 2 = 12$	Total number of groups: $3 \times 2 \times 2 = 12$

for error was set at 10 mm, which is within the range of stature variation for an individual in one day. With a 95% confidence level, this required 188 subjects in each stratum. The resulting sample size targets are shown in Table 2.14.

This resulted in a sample that is much larger than n = 250 per gender, so the confidence in the overall statistics will be excellent and this sample will also provide within subgroup precision. However, with stratified random sampling, it is necessary to apply a statistical weight to each subject when calculating overall sample statistics because the sample has proportionately more representatives from some groups and fewer representatives from other groups than the overall population.

Sometimes samples or databases are summarized in publications using statistics such as the mean, STD, and percentiles (Gordon et al., 2014; Harrison & Robinette, 2002; Snyder et al., 1975). For example, National Institute for Occupational Safety and Health (NIOSH) provides summary statistics for Firefighters and Emergency Medical Technicians (EMTs) on a CDC website (https://www.cdc.gov/niosh/topics/anthropometry/).

TABLE 2.14
Sample Size Targets for CAESAR™ Survey

a. Target Number of Subjects for North America

	Females				Males			
Age	18–29	30–44	45–65	Sum	18–29	30–44	45–65	Sum
White	188	188	188	564	188	188	188	564
Black	188	188	188	564	188	188	188	564
Other	188	188	188	564	188	188	188	564
Sum	564	564	564	1692	564	564	564	1692

b. Target Number of Subjects for Each of Italy and The Netherlands

	18–29	30–44	45–65	Sum	18–29	30–44	45–65	Sum
White	188	188	188	564	188	188	188	564
Other	188	188	188	564	188	188	188	564
Sum	376	376	376	1128	376	376	376	1128

TABLE 2.15
Example of Summary Statistics

CAESAR Variable Name:	Waist Front Length (mm)	
Mean	462.44	377.12
SEM	1.58	1.31
STD	52.85	46.52
Sample size	1119	1261
Percentiles (mm)	**Men**	**Women**
1	366.98	303.22
2	377.63	309.56
3	383.03	314.36
5	393.24	322.29
10	407.08	331.08
20	424.32	342.97
25	430.39	348.01
50	456.12	369.90
75	488.12	397.95
80	496.53	407.34
90	522.39	430.11
95	547.72	451.11
97	567.55	466.70
98	582.50	479.73
99	616.04	505.78

While these are called datasets and samples by some people, they are **not** samples. They are statistical descriptions of samples. These descriptions are correctly referred to as either aggregate data or summary statistics. This is an important distinction because a lot more can be done with raw data than with summary statistics, particularly for tailoring a sample to better match a TP.

Table 2.15 shows an example from the CAESAR™ survey (Harrison & Robinette, 2002). In this example, summary statistics for the Waist Front Length measurement are shown including the mean, the SEM, STD, and selected percentiles.

The mean is an estimate of the middle of the measurement distribution. In other words, it is a statistic that approximates the point at which half of the population is smaller and half is larger. This mean is calculated by adding all the observations and dividing by the sample size. If the distribution follows the normal distribution, this will be the center.

The 50th percentile is also an estimate of the middle. In this example, the estimated 50th percentile is the value at which the sample is divided into two equal parts. This is also called the median. Note that the mean and the 50th percentile, while both estimates of the middle, appear to be different. The mean and 50th percentile for men are 462.44 mm and 456.12 mm, respectively. The mean and 50th percentile for women are 377.12 mm and 369.9 mm, respectively. The magnitude of the difference between these two estimates is an indicator of how close the distribution of the variable matches the normal distribution.

A third measure of the middle is the mode. The mode is the point at which the most observations are clustered. It is the highest peak on a histogram or bar chart. For finding the starting size and the case or person to represent it, the mode can be the best measure of the middle. It will give us a size where most people are located. The mode is missing from the Harrison and Robinette summary statistics document and it is rarely included in summary statistics documents. This is one reason why raw data are preferable for the TP.

The percentiles in Table 2.15 are estimates of the percentages below the percentile value. In other words, the first percentile is an estimate of the size at which 1% are smaller, the second percentile is an estimate of the size at which 2% are smaller, and so on. The estimates in this example were calculated by counting the subjects whose measurements were smaller.

Another way to estimate percentiles is to assume the distribution is normal and use the mean and the STD. In other words, the 50th percentile is assumed to be the same as the mean and the mean is used as the estimate for the 50th percentile. The STD is a measure of the spread of the sample around the mean and is calculated as the square root of the variance. If the true population distribution is normal, then the mean plus one STD would be the 84th percentile. The mean plus 1.645 STDs would be the 95th percentile. If raw data are not available, this is a method for estimating the percentile value for a particular individual and we will provide some examples and instructions on how to do this.

Summary statistics (or aggregate data) can be acceptable to use for the starting sample but raw data are better. Summary statistics can be useful to get an idea about where a person falls with respect to others in the population. However, they are 1D and do not capture the relationship between measurements. For wearables, these relationships are very important. In addition, aggregate data only describe the sample in one way, the way the person calculating them wanted. So, it might be the wrong age group or the wrong mix of subjects. For example, in the CAESAR report on U.S. population (Harrison & Robinette, 2002), the summary statistics from the CAESAR survey were weighted to match the stature and weight distribution of the U.S. population from the 2000 census. This summary of the data may be out of date. According to the National Centers for Health Statistics (NCHS) since the year 2000, the population has become more obese (Fryar et al., 2018). Also, our TP may be younger than this sample which includes people ages 18–64.

On the other hand, with raw data, it is possible to reanalyze and match a more recent population. Raw data (like that in Appendix A) can be edited, re-analyzed, relationships can be explored, and the sample can be weighted or reconfigured to better reflect the composition of the new product's TP. Some raw data is available from the internet such as the WEAR Association (www.bodysizeshape.com) or you can collect your own. We will describe how both aggregate data and raw data can be used to select or evaluate cases and get the design started in Chapter 3.

Planning Full TP Sample

The primary purpose of the full TP sample is to serve as a representative sample for anthropometry and demographics, against which we will map the fit ranges, evaluate population accommodation, and evaluate the cost versus benefit of size assortments.

We will also use it to determine how many of each size to produce. Therefore, the full TP sample must be raw data that can be segmented (have some subjects dropped), weighted (give some subjects more representation and some subjects less representation), and re-analyzed and visualized (graphed).

When compiling a full TP sample, we want our results and sizing estimates to be close to the truth, but we don't want to spend too much money or too much time. More is not always better. What we want is a happy medium. To achieve this, we need to examine two things: (1) size range coverage and (2) precision of mean values.

We don't know the size range or the range of fit in a size before we start, but we can see the impact of sample size with hypothetical size ranges. An example is shown in Figure 2.17. Here we show stature and weight for males with three different sample sizes, 1122, 250, and 100 subjects. Overlaid on the plot of their measurements are 11 sizes, three height and six weight sizes and these cover ranges from 1600 to 1900 mm in stature and 50 to 100 kg in weight.

We see that with 250 subjects, we have a substantial number of subjects in each of the sizes but with only 100 subjects, some of the sizes are nearly empty. Clearly, 100 subjects are not enough to represent the spread of male subjects but 250 subjects might be if 250 is also enough to ensure we get a sufficient precision of the important and relevant measurement mean values. The mean values represent where the size ranges will begin and be centered. Evaluating the distribution of the mean is the most common method for estimating sample size.

We showed two historical examples of sample size estimates above, one that estimated 250 for each gender with 500 total and another with more subgroups and a total of more than 2500. Both used estimates of measurement means and both are good estimates. The larger sample cost more to produce but had less risk that one or more of the subgroups would be underrepresented because subgroup means were evaluated. The best way to approach planning for the full TP sample is to evaluate the risk for important and relevant measurement means given different sample sizes and use this to make an informed decision about cost versus benefit.

For evaluating anthropometric mean accuracy versus sample size, the SEM is the statistic used. The STD of a statistic, such as the mean, is referred to as the statistic's SE. The SEM estimates the distribution of the mean and the STD of the measurement estimates the distribution of the measurement. These are two different distributions, with the distribution of the mean having a smaller range, see Figure 2.18.

We know from the central limit theorem (CLT) that the distribution of the sample mean is normal with the mean (i.e., the mean of the means) being the true mean and the STD being the SEM. This is true no matter what the distribution of the original variable is. Therefore, we can use the SEM to estimate how close we are likely to be to the true mean. This is referred to as the confidence we have regarding our sample mean.

We denote the SEM as $\sigma_{\underline{X}}$ and it is calculated as follows:

$$\sigma_{\underline{X}} = \frac{\sigma}{\sqrt{n}}$$

FIGURE 2.17 Sample sizes versus hypothetical size range coverage.

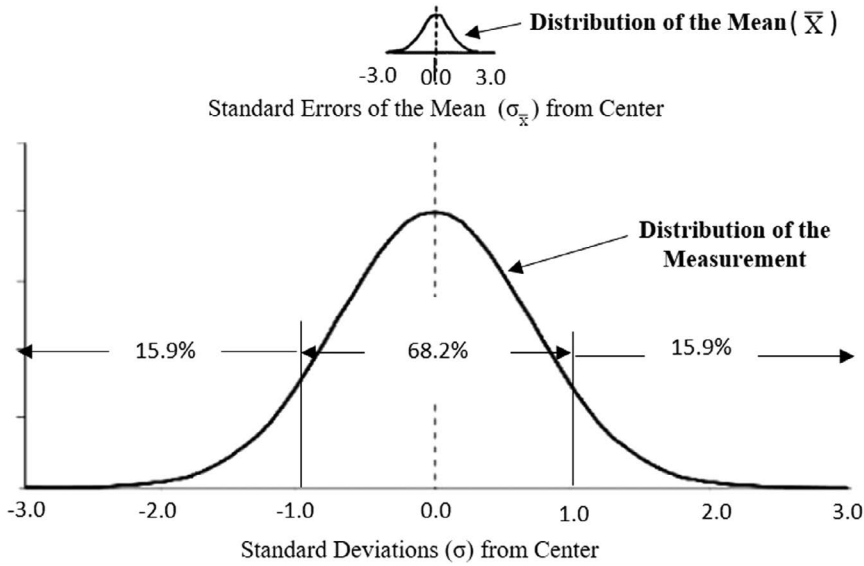


FIGURE 2.18 Distribution of the mean versus distribution of the measurement.

where σ is the STD of the measurement, and *n* is the sample size. Since *n* is in the denominator as the sample size gets larger σ_X gets smaller. Figure 2.18 is an illustration showing the difference between the distribution of the mean and the distribution of a measurement.

It is standard convention in statistics that italicized lower case terms indicate hypothetical or unknown true values, whereas upper case non-italicized terms indicate values estimated from a sample. For example, *n* is the hypothetical sample size, whereas N is actual sample size, σ is the true standard deviation which is unknown, and STD is the estimated standard deviation from the sample.

Churchill and McConville (1976) used the size of $\sigma_{\underline{X}}$ for 187 different measurements to arrive at the sample size of 250. They found that most of the 187 measurements would have a $\sigma_{\underline{X}}$ with a size that was within their tolerable error with 250 subjects. They did not estimate the confidence for the means within ethnic or age subgroups. This is where they assumed some representation and accuracy risk because the within-group values would be much larger given the sample size within the groups would be much smaller than 250.

In contrast, for the CAESAR™ project (Robinette et al., 2002) confidence ranges were estimated for just one measurement, stature, but the confidence levels within racial and age subgroups were also examined. This required more subjects, time, and cost overall, but had less risk that any of the subgroups would be under-represented or inaccurate. The two methods also had different ways to set the amount of tolerable error. Churchill and McConville set the error as a percentage of the mean values and arrived at n = 250 because more than 90% of the 187 variables were within the tolerable error. CAESAR™ set the tolerable error at 10 mm for stature and used a 95%

confidence interval estimate of the stature mean to indicate the true stature mean would fall within 10 mm. This sample had less risk but at a higher cost.

Both methods are valid and reasonable, but for our product, we need to assess sample size using measurements and tolerance ranges that relate to the amount of manufacturing or sizing error our product can tolerate. *The best way to estimate sample size is to use a variety of estimates and weigh benefit, risk, and cost to make an informed decision.*

To estimate the number of subjects, we need to know (or estimate) three things: (1) the measurements that need to be accurate, (2) the amount of accuracy they need to have, and (3) the σ, of the measurements in our TP. We need some prior data to estimate the σ values. Usually, we do not have prior data on the exact variable we think is most important and sometimes we do not know which variable is most important either. Therefore, looking at a selection of variables can help.

The most used sample size estimation formula is the one for a simple random sample of one continuous variable. This formula is:

$$n = z^2\sigma^2/e^2$$

where n is the sample size, e is the amount of error we are willing to accept (called tolerable error), σ is the estimated standard deviation for the measurement we are using to estimate the sample size, and z is the normal deviate or z score for confidence level we want to use. The z score is the distance from the mean expressed in $\sigma_{\bar{X}}$ units and the confidence level or confidence percentage is the area under the normal curve (the z distribution) at that z score distance. For example, for 95% confidence that the true mean will be within our tolerance of the sample mean we want the z score to be 1.96 or greater. These values can be found in normal distribution tables or using online calculators.

To use the sample size formula, we must choose the measurement we want to use, the amount of error we are willing to accept, and the level of confidence we want to have, for example, 90% confidence or 95% confidence, and we must also estimate the standard deviation (σ). This is what was done for the CAESAR project (Robinette et al., 2002). Stature was chosen as the measurement, 10 mm was selected as the amount of error we would tolerate within a subgroup, 95% was selected as our confidence level, and a within-group STD for stature was estimated at 70 mm as shown in Table 2.16.

The equation becomes:

$$n = \frac{1.96^2 70^2}{10^2} = 188$$

TABLE 2.16
Values for Estimating Within Subgroup Sample Size

	σ	σ^2	$\sigma_{\bar{X}}$	Tolerable Error	*z*	Confidence Level	*n*
Stature (mm)	70	4900	4.95	10	1.96	95%	188

The issue is that we do not always know what measurements or confidence levels we want to use or what tolerances we are willing to accept. So, it is helpful to look at some options like multiple variables, and multiple tolerances, and to decide how much risk is acceptable. An example of this is shown in Table 2.17 a and b. This example is essentially a combination of the Churchill and McConville method and the CAESAR method. Here we compare results for four measurements, two levels of tolerable error and two sample sizes, n = 200 and n = 50 for the male sample. We see that at n = 200 with the larger tolerance, our confidence levels are more than 99% and with n = 50 the confidence levels are all over 93%. If these are our most important measurements and the larger tolerances are acceptable adding 150 subjects only improved our confidence in the accuracy of the means for the two genders by 6%, so a sample size of 50 is reasonable.

In contrast, at the small tolerance level, there is a more dramatic difference in our confidence levels with the two sample sizes. With n = 200, our confidence levels are all over 90% and all but one over 95%; however, with n = 50, confidence levels drop to less than 70% in some cases. Note that in this example the small tolerance confidence levels with 200 subjects are equivalent to the large tolerance confidence levels with 50 subjects. This illustrates that increasing the sample size increases the precision of the estimate.

Based on this evaluation, we might decide to start with n = 50 for each gender group and check with our actual data to see how confident we can be that our group is well represented. If not, we could decide to add more subjects at that time. Or we might decide to plan for some value between 50 and 250 as a compromise.

A sample size of 30 is often used in statistics as a minimum number per subgroup. The reason for this is that this is the number at which the distribution of the mean can be estimated using the standard normal distribution (z) rather than the student's T distribution. It is thought of as the point at which the mean estimate become stable or minimally reliable.

As noted in Table 2.11, there are three sources for the final sample: (1) the starting sample if it is raw data, (2) fit test samples, and (3) a new large sample collected. The full TP sample is often a combination of the three and it is important to have a plan for the sample before data collection begins to make the best of the resources available.

Planning consists of:

1. Defining subgroups to be sampled
2. Estimating sample size desired in each subgroup
3. Determining the sources for the subjects (new sample vs fit test samples, etc.)
4. Planning for combining samples, if needed

Sometimes we start with a small sample, and after an assessment, we determine we need to add more subjects. This might be subjects of a particular size, or from some group that we feel needs more representation given the initial results. It is important to plan for that possibility. Also, we can build up a large sample over time by combining data from our various tests if we manage the data from our samples well. This is discussed further under database management and maintenance.

TABLE 2.17

Effect of Sample Size and Tolerance on Error and Confidence Level Estimates: (a) Males n = 200, (b) Males n = 50

(a) Error and Confidence Estimates for Males When n = 200

	Prior Estimates				Large Tolerable Error			Small Tolerable Error		
Variable	**Mean**	**σ**	**σ²**	**σ_x with n = 200**	**Tolerable Error**	***z***	**Confidence Level**	**Tolerable Error**	***z***	**Confidence Level**
IPD (mm)	65.0	3.19	10.2	0.226	1	4.43	99.9%	0.5	2.215	97.3%
Head Breadth (cm)	15.8	0.72	0.5	0.051	0.2	3.91	99.9%	0.1	1.956	95.0%
Head Length (cm)	19.1	0.77	0.6	0.054	0.2	3.68	99.9%	0.1	1.838	93.2%
Head Circ. (cm)	56.4	1.79	3.2	0.127	0.5	3.95	99.9%	0.25	1.973	95.2%

(b) Error and Confidence Estimates for Males When n = 50

	Prior Estimates				Large Tolerable Error			Small Tolerable Error		
Variable	**Mean**	**σ**	**σ²**	**σ_x with n = 50**	**Tolerable Error**	***z***	**Confidence Level**	**Tolerable Error**	***z***	**Confidence Level**
IPD (mm)	65.0	3.19	10.2	0.451	1	2.22	97.3%	0.5	1.108	73.2%
Head Breadth (cm)	15.8	0.72	0.5	0.102	0.2	1.96	95.0%	0.1	0.978	67.2%
Head Length (cm)	19.1	0.77	0.6	0.109	0.2	1.84	93.2%	0.1	0.919	64.2%
Head Circ. (cm)	56.4	1.79	3.2	0.253	0.5	1.97	95.2%	0.25	0.986	67.6%

Weighting Samples

Weighting is a technique in survey research where the observations (or subjects) are adjusted to reflect a projected TP more accurately. This is used if the sample has more representatives from some groups and fewer representatives from other groups than the population. Weighting can also be used to evaluate the cost versus benefit of creating sizes for different populations.

Companies with marketing departments have researched to understand their customers and they characterize their customers' demographics. These can also be estimated using data from national statistics such as the U.S. Census. With these numbers, the weights can be calculated as shown in Table 2.18. The estimated TP demographics proportions are divided by the sample demographics proportions to produce the weight values for each category. These can then be applied to create a weighted sample. For example, each Asian male in our example would be given the weight value of .85 and each Asian female, the value of .31. This means when counted in the weighted sample, they would represent .85 people and .31 people, respectively. They represent less than one person in the weighted because our original sample had a greater proportion of Asian males and females than the TP.

To create a weighted sample, we first create a variable that contains the weight. In SPSS®, this can be done in the Syntax Editor as shown in Figure 2.19. Then this weight is applied to the sample using the "Weight Cases" function shown in Figure 2.20. The

TABLE 2.18
Calculating Weights

Target Population (TP)	Male	Female	Total
Asian (A)	0.06	0.04	0.10
Black (B)	0.06	0.04	0.10
White (C)	0.40	0.27	0.67
Other (O)	0.01	0.01	0.02
Hispanic	0.07	0.04	0.11
Total	0.60	0.40	1.00
Sample			
Asian (A)	0.07	0.13	0.20
Black (B)	0.18	0.17	0.35
White (C)	0.07	0.12	0.19
Other (O)	0.07	0.02	0.09
Hispanic	0.11	0.06	0.17
Total	0.50	0.50	1.00
Weights (TP/Sample)			
Asian (A)	0.85	0.31	
Black (B)	0.33	0.24	
White (C)	5.71	2.25	
Other (O)	0.14	0.5	
Hispanic	0.63	0.67	
Total	0.85	0.31	

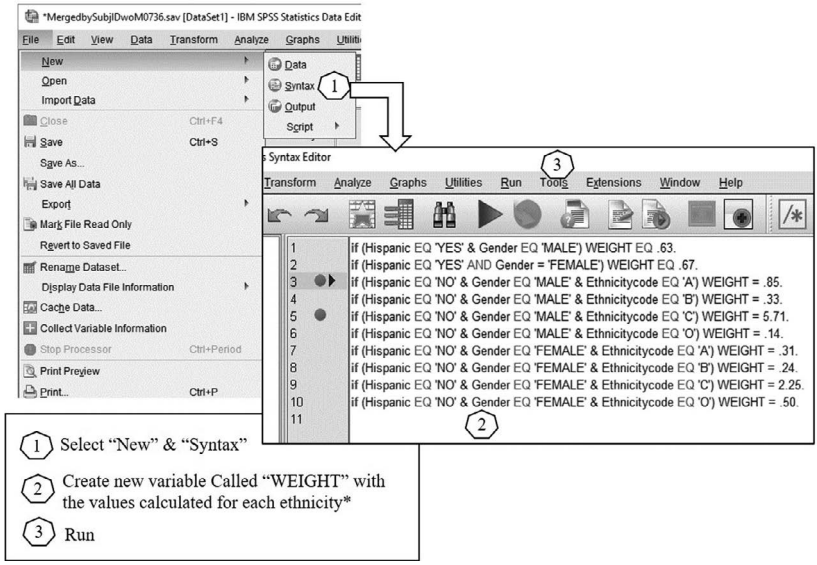


FIGURE 2.19 Creating a demographic weight variable in SPSS®.

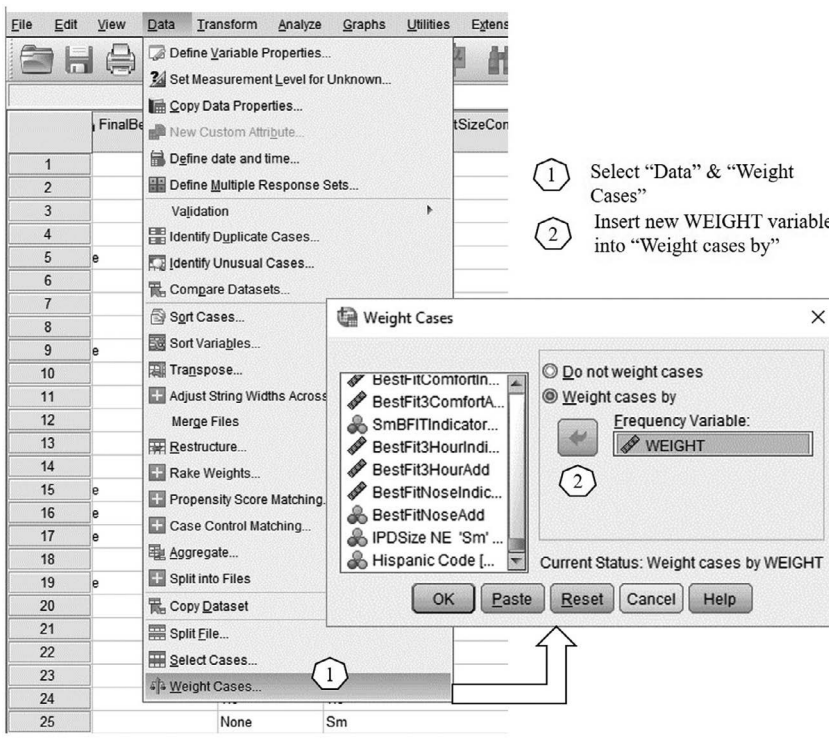


FIGURE 2.20 Applying the weights to the sample in SPSS®.

example below shows how code is written in Syntax Editor to match the weights calculated in Table 2.18.

Note that in SPSS®, the weight function can be turned off and on enabling an evaluation of the impact of using the weighted sample versus the unweighted sample. Also, several different weight variables can be created and applied to the sample. This is very useful for evaluating sizing for different TPs. So, for example, we might have one weight variable for the United States and another one for Australia. Or we might have one weight variable for a predominately younger sample and another for a predominantly older sample. This is useful for cost versus benefit analysis (CBA).

It is important to note that weighting *should not be used* if any of the strata or categories have no one in them. Zero times anything is zero, so the weights do not work. It is also not advisable if there are less than five people in one or more of the categories.

WHAT CAN GO WRONG?

No matter how well we plan, things can still go wrong. Therefore, it is best to plan for things to go wrong and have some alternative back up plans. Some examples are included in Table 2.19.

It is important that once data collection begins, the procedures are not changed. Changing in the middle can render data analysis impossible. Some examples of what *not* to do include the following:

- Do not stop recording an issue or a measurement because everyone has the issue or no one has the issue
- Do not change how an issue is recorded or a measurement is taken once the test has started

TABLE 2.19
Examples of Issues and Backup Plans

Issue	Backup Plans
Liquid (coffee?) spilled on laptop/tablet	Have two data records for all data: • Paper and electronic • On-line and local computer
Lost internet connection/website access	Save data locally either on laptop or paper (so have paper forms already printed out)
Fire alarm evacuation	Have a plan to secure data, prototypes, and protect subjects
Hurricane evacuation	Have plan to notify all subjects and personnel
Subject or investigator injured	Be prepared with first aid supplies
Sizes are mis-labeled and discovered after data collection started	Note the error, the date identified, and the last subject recorded with the error. Keep the old label variable and add a new label variable with the correct label
Prototype is broken	Plan for backup systems or repairs

- Do not stop collecting data on some subgroups because they all have fit issues
- Do not change the sampling plan

If there are serious issues that would make the test ineffective, then stop the test. Fix the issues. Then start the test over. We recommend pilot testing all procedures before starting testing so that there will not be a need to start over. All important issues should be resolved during the pilot testing.

REFERENCES

ASTM D5219-15, (2015). *Standard Terminology Relating to Body Dimensions for Apparel Sizing,* ASTM International.

Ballester, A., Wright, W., Valero, J., Scott, E., Devlin, T., Bullas, A., & McDonald, C. (2022). White Paper-IEEE SA 3D Body Processing Industry Connections--Comparative Analysis of Anthropometric Methods: Past, Present, and Future. *Comparative Analysis of Anthropometric Methods: Past, Present, and Future*, 1-52., ISBN:9781504486828.

Bass, W. M. (1971). *Human Osteology: A Laboratory and Field Manual of the Human Skeleton*. Missouri Archaeological Society.

Blackwell, S. U., & Robinette, K. M. (1993). *Human Integration Evaluation of Three Helmet Systems* (Technical Report AL-TR-1993-0028). Air Force Material Command. https://apps.dtic.mil/sti/citations/ADA271320

Blackwell, S., Robinette, K. M., Boehmer, M., Fleming, S., Kelly, S., Brill, T., Hoeferlin, D., & Burnsides, D. (2002). *Civilian American and European Surface Anthropometry Resource (CAESAR).* Volume 2: Descriptions. United States Air Force Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA408374.pdf

Blank, S. (2020). *The Four Steps to the Epiphany: Successful Strategies for Products That Win* (5th ed.). Wiley Publishing Inc.

Bragança, S., Arezes, P., Carvalho, M., Ashdown, S. P., Castellucci, I., & Leão, C. (2018). A Comparison of Manual Anthropometric Measurements With Kinect-Based Scanned Measurements in Terms of Precision and Reliability. *Work*, *59*(3), Article 3. https://doi.org/10.3233/WOR-182684

Burnsides, D. B., Files, P. M., & Whitestone, J. J. (1996). Integrate 1.25: A Prototype for Evaluating Three-Dimensional Visualization, Analysis, and Manipulation Functionality (AL/CF-TR-1996-0095). Crew Systems Directorate, Human Engineering Division, Armstrong Laboratory. https://apps.dtic.mil/sti/citations/ADA330986

Case, H., Ervin, C., & Robinette, K. M. (1989). *Anthropometry of a Fit Test Sample Used in Evaluating the Current and Improved MCU-2/P Masks* (Technical Report AAMRL-TR-89-009). Armstrong Aerospace Medical Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/a215173.pdf

Choi, B. C. K., & Pak, A. W. P. (2005). A Catalog of Biases in Questionnaires. *Preventing Chronic Disease*, *2*(1). https://stacks.cdc.gov/view/cdc/19899

Choi, H. J., Zehner, G. F., & Hudson, J. A. (2009). *A Manual for the Performance of Protective Equipment Fit Mapping* (Technical Report AFRL-RH-WP-SR-2010-0005). Air Force Research Laboratory, Human Effectiveness Directorate. https://apps.dtic.mil/sti/citations/ADA519894

Churchill, E., & McConville, J. T. (1976). *Sampling and Data Gathering Strategies for Future USAF Anthropometry* (Technical Report ADA025240). Air Force Aerospace Medical Research Lab. https://apps.dtic.mil/sti/pdfs/ADA025240.pdf

Clauser, C. E., Tucker, P. E., McConville, J. T., Churchill, E., Laubach, L. L., & Reardon, J. A. (1972). *Anthropometry of Air Force Women* (Technical Report AMRL-TR-70-5). Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command. https://apps.dtic.mil/sti/pdfs/AD0743113.pdf

Daanen, H. A., Brunsman, M. A., & Robinette, K. M. (1997). *Reducing Movement Artifacts in Whole Body Scanning*. Proceedings. International Conference on Recent Advances in 3-D Digital Imaging and Modeling (Cat. No. 97TB100134), 262–265.

Daniels, G. S., Meyers, H. C., & Worrell, S. H. (1953). *Anthropometry of WAF Basic Trainees* (Technical Report WADC Technical Report 53-12). Aero Medical Laboratory. https://apps.dtic.mil/sti/tr/pdf/AD0020542.pdf

Fryar, C. D., Carroll, M. D., & Ogden, C. L. (2018). *Prevalence of Overweight, Obesity, and Severe Obesity Among Adults Aged 20 and Over: United States, 1960–1962 Through 2015–2016 (NCHS Health E-Stats)* [Technical Report]. National Center for Health Statistics. https://www.cdc.gov/nchs/data/hestat/obesity_adult_15_16/obesity_adult_15_16.pdf

Garplid, D. (2013, November). *Top 10 Startup Mistakes*. http://100firsthits.com/2013/11/15/top-10-startup-mistakes-infographics/

Gill, S., & Parker, C. J. (2017). Scan Posture Definition and Hip Girth Measurement: The Impact on Clothing Design and Body Scanning. *Ergonomics*, *60*(8), Article 8.

Goodwin, K. (2009). *Designing for the Digital Age: How to Create Human-Centered Products and Services*. Wiley Publishing Inc.

Gordon, C. C., Blackwell, C. L., Bradtmiller, B., Parham, J. L., Barrientos, P., Paquette, S. P., Corner, B. D., Carson, J. M., Venezia, J. C., Rockwell, B. M., Mucher, M., & Kristensen, S. (2014). *2012 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics* (Technical Report NATICK/TR-15/007). U.S. Army Natick Soldier Research, Development and Engineering Center. https://apps.dtic.mil/sti/pdfs/ADA611869.pdf

Gordon, C. C., & Bradtmiller, B. (1992). Interobserver Error in a Large Scale Anthropometric Survey. *American Journal of Human Biology*, *4*(2), Article 2. http://onlinelibrary.wiley.com/doi/10.1002/ajhb.1310040210/full

Gordon, C. C., Churchill, T., Clauser, C. E., Bradtmiller, B., & McConville, J. T. (1989). *1988 Anthropometric Survey of US Army Personnel: Methods and Summary Statistics* (Technical Report Natick/TR-89/044;). U.S. Army Natick Soldier Research, Development and Engineering Center. https://apps.dtic.mil/sti/pdfs/ADA225094.pdf

Grunhofer, H. J., & Kroh, G. (1975). *A Review of Anthropometric Data of German Air Force and United States Air Force Flying Personnel 1967–1968* (Technical Report AGARDograph No. 205). Advisory Group for Aerospace Research and Development, North Atlantic Treaty Organization. https://apps.dtic.mil/dtic/tr/fulltext/u2/a010674.pdf

Han, H., Nam, Y., & Choi, K. (2010). Comparative Analysis of 3D Body Scan Measurements and Manual Measurements of Size Korea Adult Females. *International Journal of Industrial Ergonomics*, *40*(5), 530–540. https://doi.org/10.1016/j.ergon.2010.06.002

Harrison, C. R., & Robinette, K. (2002). *CAESAR: Summary Statistics for the Adult Population (Ages 18-65) of the United States of America* (Technical Report AFRL-HE-WP-TR-2002-0170). United States Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division. https://apps.dtic.mil/sti/pdfs/ADA406674.pdf

Heilmeier, G. H. (2021). *The Heilmeier Catechism*. Defense Advanced Research Project Agency. https://www.darpa.mil/work-with-us/heilmeier-catechism

Herron, R. E., Cuzzi, J. R., & Hugg, J. (1976). *Mass Distribution of the Human Body Using Biostereometrics* (Technical Report AMRL-TR-75-18). Aerospace Medical Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA029402.pdf

Hotzman, J., Gordon, C. C., Bradtmiller, B., Corner, B. D., Mucher, M., Kristensen, S., Paquette, S., & Blackwell, C. L. (2011). *Measurer's Handbook: US Army and Marine Corps Anthropometric Surveys, 2010–2011* (Technical Report Natick/TR-11-017). https://apps.dtic.mil/sti/pdfs/ADA548497.pdf

Hsiao, H., Whisler, R., & Bradtmiller, B. (2021). Needs and Procedures for a National Anthropometry Study of Law Enforcement Officers. *Human Factors*. https://doi.org/10.1177/00187208211019157

Hsiao, H., Whitestone, J., Kau, T.-Y., Whisler, R., Routley, J. G., & Wilbur, M. (2014). Sizing Firefighters: Methods and Implications. *Human Factors*, *56*(5), 873–910. https://doi.org/10.1177/0018720813516359

ISO 18825-1:2016. (2016). *Clothing – Digital fittings – Part 1: Vocabulary and Terminology Used for the Virtual Human Body*. https://www.iso.org/standard/61643.html

ISO 18825-2:2016. (2016). *Clothing – Digital fittings – Part 2: Vocabulary and Terminology Used for Attributes of the Virtual Human Body*. https://www.iso.org/standard/63494.html

ISO 20685-1:2018. (2018). *3-D Scanning Methodologies for Internationally Compatible Anthropometric Databases—Part 1: Evaluation Protocol for Body Dimensions Extracted from 3-D Body Scans*. http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/32/63260.html

ISO 7250-1:2017. (2017). *Basic Human Body Measurements for Technological Design – Part 1: Body Measurement Definitions and Landmarks*. https://www.iso.org/standard/65246.html

ISO 8559-1. (2017). *Size Designation of Clothes – Part 1: Anthropometric Definitions for Body Measurement, ISO/TC 133*. http://www.iso.org/iso/catalogue_detail.htm?csnumber=61686

Kouchi, M. (2014). Anthropometric Methods for Apparel Design: Body Measurement Devices and Techniques. In *Anthropometry, Apparel Sizing and Design* (pp. 67–94). Elsevier. https://doi.org/10.1533/9780857096890.1.67

Kouchi, M., & Mochimaru, M. (2005). Causes of the Measurement Errors in Body Dimensions Derived from 3D Body Scanners: Differences in Measurement Posture. *Journal of Anthropological Science (Japanese Series)*, *113*, 63–75.

Kouchi, M., Mochimaru, M., Tsuzuki, K., & Yokoi, T. (1996). Random Errors in Anthropometry. *Journal of Human Ergology*, *25*, Article 25.

Laubach, L. L., McConville, J. T., Churchill, E., & White, R. M. (1977). *Anthropometry of Women of the U. S. Army—1977. Report Number 1. Methodology and Survey Plan* (Technical Report NATICK/TR-77/021; p. 202). U.S. Army Natick Research and Development Command. https://apps.dtic.mil/sti/pdfs/ADA043715.pdf

LaBat, K. L., & Ryan, K. s. (2019). *Human Body: A Wearable Product Designer's Guide*. CRC Press.

Lu, J. M., & Wang, M. J. J. (2010). The Evaluation of Scan-Derived Anthropometric Measurements. *IEEE Transactions on Instrumentation and Measurement*, *59*(8), Article 8. https://doi.org/10.1109/TIM.2009.2031847

McConville, J. T., Churchill, T. D., Kaleps, I., Clauser, C. E., & Cuzzi, J. (1980). *Anthropometric Relationships of Body and Body Segment Moments of Inertia* (Technical Report AFAMRL-TR-80-119). Air Force Aerospace Medical Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA097238.pdf

McDonald, C., Wu, Y., & Ballester, A. (2018). *IEEE Industry Connections (IEEE-IC) Landmarks and Measurement Standards Comparison in 3D Body-model Processing*. 34.

Mckinnon, L., & Istook, C. L. (2002). Body Scanning: The Effects of Subject Respiration and Foot Positioning on the Data Integrity of Scanned Measurements. *Journal of Fashion Marketing and Management: An International Journal*, *6*(2), Article 2.

Mellian, S. A., Ervin, C., & Robinette, K. M. (1991). *Sizing Evaluation of Navy Women's Uniforms* (Technical Report AL-TR-1991-0116). Air Force Systems Command. https://apps.dtic.mil/sti/citations/ADA249782

National Science Foundation. (2020). *NSF Innovation Corps*. https://www.nsf.gov/news/special_reports/i-corps/index.jsp

Perkins, T., Burnsides, D. B., Robinette, K. M., & Naishadham, D. (2000). *Comparative Consistency of Univariate Measures from Traditional and 3-D Scan Anthropometry*. SAE Technical Paper.

Pheasant, S. (2014). *Bodyspace: Anthropometry, Ergonomics and the Design of Work: Anthropometry, Ergonomics and the Design of Work*. CRC Press.

Robinette, K. M. (1986). Three-Dimensional Anthropometry-Shaping the Future. Human Factors Society Inc., Santa Monica, CA. Proceedings of the Human Factors Society-30th Annual Meeting, Vol. 1, 205.

Robinette, K. M., Blackwell, S., Daanen, H. A. M., Boehmer, M., Fleming, S., Brill, T., Hoeferlin, D., & Burnsides, D. (2002). *Civilian American and European Surface Anthropometry Resource (CAESAR) Final Reports, Volume I: Summary* [Technical Report]. Air Force Research Laboratory, Human Effectiveness Directorate. https://apps.dtic.mil/sti/citations/ADA406704

Robinette, K. M., & Daanen, H. A. M. (2006). Precision of the CAESAR Scan-Extracted Measurements. *Applied Ergonomics*, *37*(3), Article 3. https://doi.org/10.1016/j.apergo.2005.07.009

Robinette, K. M., Daanen, H., & Paquet, E. (1999). *The CAESAR Project: A 3-D Surface Anthropometry Survey*. Second International Conference on 3-D Digital Imaging and Modeling (Cat. No. PR00062), 380–386.

Robinette, K. M., Ervin, C. A., & Zehner, G. F. (1986). *Dexterity Testing of Chemical Defense Gloves (U)* (Technical Report AAMRL-TR-86-021). Air Force Aerospace Medical Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA173545.pdf

Robinette, K. M., Mellian, S. A., & Ervin, C. A. (1991). *Development of Sizing Systems for Navy Women's Uniforms* (Technical Report AL-TR-1991-0117; Issue AL-TR-1991-0117). Armstrong Laboratory, Air Force Systems Command. https://apps.dtic.mil/sti/pdfs/ADA250071.pdf

Robinette, K. M., & Veitch, D. (2018). Use of Anthropometry and Fit Databases to Improve the Bottom-Line. *Advances in Intelligent Systems and Computing*, *826*, 442–452. https://doi.org/10.1007/978-3-319-96065-4_49

Snyder, R. G., Spencer, M. L., Owings, C. L., & Schneider, L. W. (1975). *Anthropometry of U.S. Infants and Children*. Society of Automotive Engineers, Inc.

White, T. D., Black, M. T., & Folkens, P. A. (2012). *Human Osteology* (3rd ed.). Elsevier Academic Press.

Whitestone, J., & Robinette, K. M. (1997). Fitting to Maximize Performance of HMD Systems. In *Head Mounted Displays: Designing for the User*. McGraw-Hill.

Zehner, G. F., Ervin, C., Robinette, K. M., & Daziens, P. (1987). *Fit Evaluation of Female Body Armor*. Armstrong Aerospace Medical Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA188721.pdf

3 Cases and Fit Models

Kathleen M. Robinette, Daisy Veitch, and Sandra Alemany

ABSTRACT

After we have all the tools and resources we need, the next step is to create our first prototype or mock-up. To do this we need some representation of an individual or individuals around which to design and build it. We call these individuals "cases". A case can have three forms: (1) a list of measurements of an individual, (2) a three-dimensional (3D) or four-dimensional (4D) model of an individual, or (3) the actual individual. The actual individual is sometimes called a fit model or a live model. This begins with the selection of a case to represent the first size called the base size. Selecting this case effectively gets the design centered on the area of the Target Population (TP) that has the greatest concentration of people. Additional cases are selected to evaluate the potential range of fit. Cases are portrayed in two ways for prototyping products: (1) physical manikins and people or (2) digital manikins. This chapter discusses how to select, model, and use a small number of cases to design, modify, or redesign a product and produce a mock-up or prototype.

This chapter deals with the first three steps in the design loop of the Sustainable Product Evaluation, Engineering, and Design (SPEED) process design loop (shown again in Figure 3.1): (1) selecting and using a case or a set of cases, (2) designing/modifying/redesigning the product, and (3) production of prototypes or mock-ups. Some projects have larger budgets and more resources available than others. We organized this chapter to start with the best methods, then we add additional methods to accommodate reduced resources or budgets. We include explanations about the risks and limitations of the different methods so the reader can decide which options work best for their situation.

While a random sample from our Target Population (TP) is a sample of cases for design purposes, it can be difficult, time-consuming, and expensive to use the whole sample. Instead, we use a small number of individuals, relevant to the product, selected based upon their body size and shape. For wearables, that have multiple sizes, case selection begins with someone in the middle, where the population is concentrated, for two or three of our most important variables. The mid-size case represents the first size, called the base size. The base size is the foundation of a product's design and sizing, and a well-selected base size is essential for a good sustainable fit standard. If the base size is in the wrong place the entire sizing system will suffer.

The base size case is used to create the first prototype and to evaluate design issues and options. It is best if this case is represented by a live model who can don the prototype and provide feedback in the design loop. It can also be represented by other models, such as a manikin or a Computer-Aided Design (CAD) model, but only the live model can provide meaningful feedback. Because the model will provide

DOI: 10.1201/9781003397533-3

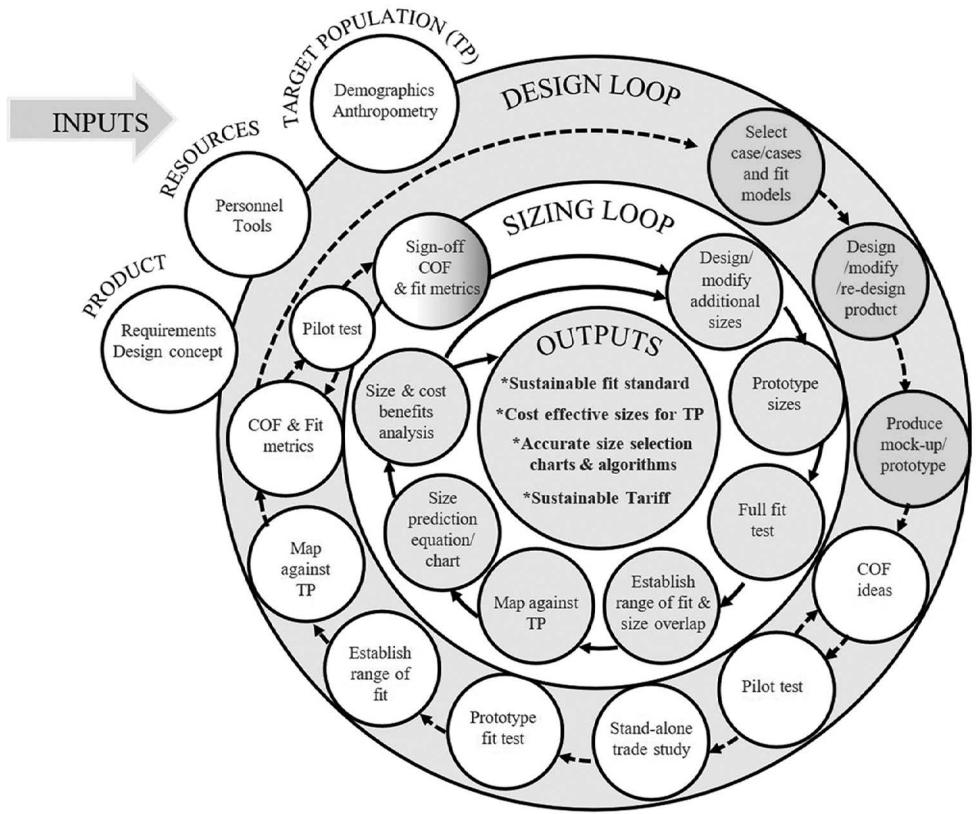


FIGURE 3.1 SPEED process.

feedback, it can be important that the live model be drawn from the user population and be very knowledgeable about how the product should fit. This is particularly true for protective equipment or wearables, where use case experience is an important factor. For example, someone who has experience wearing a helmet will give better feedback about a helmet fit than someone who has never worn one. The experienced wearer will understand how uncomfortable something can become after it is worn for a length of time and will appreciate the benefit of a snug helmet fit over a loose one, even if it is a bit more uncomfortable. Similarly, an experienced police officer who has worn body armor on the job will give better feedback regarding body armor for police officers, than someone in another occupation or who has never worn body armor.

Most products will require multiple sizes, so the base size is just the start. Design loop testing reveals design issues, the range of fit within each size and the measurements that are the best predictors of good fit. These are unknowns or theories before fit testing. When developing a new wearable (as opposed to evaluating an existing one) it is usually most cost-effective to start with one case and one size, then work out all the design and fit issues for the first case and size before adding cases and/or sizes. It is cheaper and faster to fix one prototype than it is to fix many. Also, after testing the first size it may be determined that fewer sizes are needed than originally hypothesized so some of the sizes built are wasted time and material. Even if it is determined more sizes are needed than originally hypothesized it is likely that some of the prototypes built, fall between the sizes needed so they are wasted as well.

However, if we only have one size for testing, we have limited ability to discern how much size overlap we need so we need to test additional sizes in another design loop. If we have a good idea when we start about what additional sizes we need, then testing multiple sizes at the start might save us some time. The good news is that the process for selecting the base size case can also be used to select multiple cases and sizes before fit testing if we choose. It often depends on the maturity of the product and the existing knowledge about how it should fit.

If we have predecessor products that we can use for fit testing or already have a fit model that we have been using, the case selection process can still be informative. It helps us evaluate our fit model and our test subjects to verify they are well-placed and to confirm a good representation of the TP. This is part of what we refer to as a fit audit and helps establish and sustain an effective fit standard.

SELECTING CASES

It is common for people to assume that if we design for the average person, we will fit most people. This is not true. Averages and percentiles can be very helpful in selecting the individuals to use as cases because they give us a way to understand where a person falls in size compared to everyone else. However, if we build prototypes from averages or percentiles, or 3D models built on averages and percentiles, rather than individual people, it has been illustrated by many scientists that we might not fit anyone (Dainoff et al., 2004; Daniels, 1952; Robinette & Hudson, 2006). It isn't that averages are bad. It is that averages are only relevant to one measurement. When we try to use the average of all measurements, we create an entity that doesn't exist. An individual person might be average in one measurement, above average for another, and below average for a third. A simple example illustrating the issue is shown in Figure 3.2.

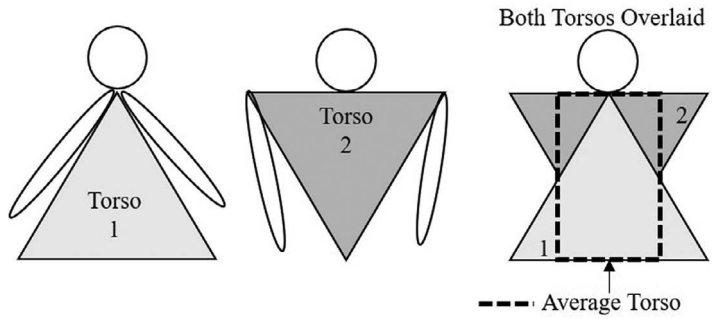


FIGURE 3.2 Average is too small for both.

In this figure, we see two individuals. Subject 1 has narrow shoulders and wide hips and subject 2 has wide shoulders and narrow hips. Suppose we are building a vest, such as a compression vest or a ballistic vest, and we must cover the torso and fit snuggly. If we use the average of these two individuals for all their torso measurements, we create the average torso, and we see that the average of these two torsos is too small for both. For subject 1 it is too small in the hips and for subject 2 it is too small in the shoulders. Depending on our Concept-of-Fit (COF) it may also be too large for both, but in the opposite areas.

In Figure 3.3, we show a real example of measurements for individuals and their relative size expressed as percentiles. Here we see three different subjects from the CAESAR™ survey (Harrison & Robinette, 2002), subjects 56, 87, and 89. The percentile value for ten of their measurements is shown. The percentile scale is shown along the vertical axis (y) and the measurement number is shown along the horizontal axis (x). The average in this diagram is represented by the 50th percentile, which is the point at which 50% fall above and 50% fall below. The first percentile indicates the point at which 1% fall below and 99% fall above.

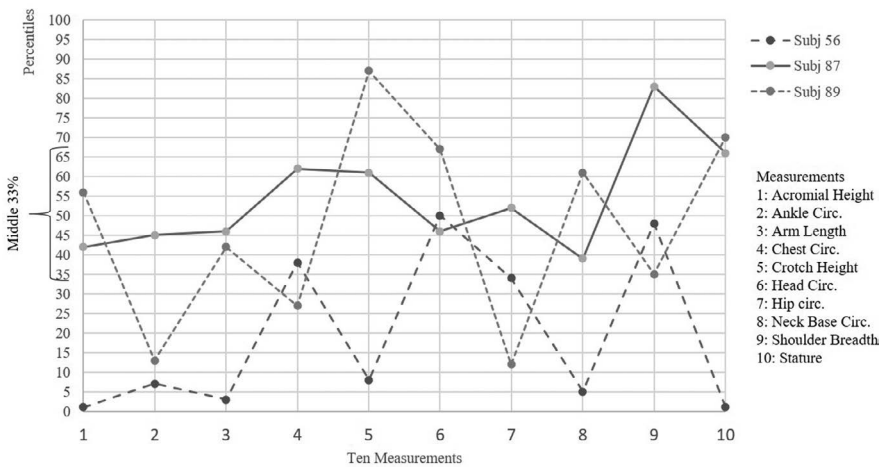


FIGURE 3.3 Three subjects' percentile values for ten measurements.

If we follow the 50th percentile across the graph we see that while all three subjects have values on both sides of this line, only one person, subject 56, falls directly on this line and only for one measurement, Head Circumference. The percentile range that encloses the middle one-third or 33% is also shown along the y axis. If we follow this range horizontally, we see that all three subjects fall in this middle range for at least one measurement and fall outside this range for at least one measurement. They are all average for some things and not for others. All real people are a mixture of small, medium, and large measurements. No one is average or any given percentile for all measurements.

While we show just three people here, Daniels (1952) demonstrated that this is literally true for everyone if there are more than a few measurements. Daniels demonstrated that out of 4,000 men, none were within the middle one-third for all 15 measurements. Thus, something built around all average values could be either too large or too small for everyone.

Instead of "average people" or "percentile people" that don't exist we use cases. A *case* is a single individual to be represented in a product design or evaluation (Dainoff et al., 2004; Zehner et al., 1993). A case can have three forms: (1) a list of measurements of an individual, (2) a three-dimensional (3D) or four-dimensional (4D) model of an individual, or (3) the actual individual. The actual individual is sometimes called a fit model or a live model. If we fit the individual, we know we will accommodate at least one actual person, the case, and probably other people who are similar in size and shape. If we choose our first case wisely, we can accommodate the greatest number of people from our TP in our first size.

In Chapter 2, we described how to find or collect a suitable sample to represent the TP with relevant variables and demographics. We also described how to tailor the sample if raw data were available. This is the starting TP sample. We use this sample to identify and select cases. The process for selecting cases has three steps:

1. Identify and select two to three key variables
2. Evaluate and select cases for the base size
3. Evaluate and select cases for size range estimation

Since people are not average for all measurements or large for all measurements, we must decide which measurements are the most important for the product and select two to three of them to guide our case selection and control our sizing. We call these variables *key variables*. These are variables that are both important to the product and, when used in combination, should control most of the important size variability. They are important for paring down our original set of variables to something manageable and easy to understand. Key variables are used in several ways including:

1. To select cases
2. To design a product
3. To evaluate fit
4. To predict fit for cost versus benefit analysis
5. To help the users find the best fitting size

In new product designs, one set of key variables can be identified initially, and after fit testing, these key variables can be changed if necessary, based on learnings from the fit testing. For products that have predecessors or a history of sizing for similar products, such as clothing, the manufacturer will typically have a good idea about what the top three key variables are. Things to consider when deciding on key variable combinations are:

- How relevant and important are they to the design?
- How easy are they for non-experts to measure or use to select a size?
- Is there an existing size standard for similar products?
- How closely related (or correlated) are they to other measurements?
- How much variability in all measurements will be controlled by the key variables?
- How well do they predict fit?

We begin by examining candidate variables and their relationships. As we progress through fit testing, we may change the key variables based on a better understanding of which variables affect fit, fit issues, and sizing.

After we have selected the key variables, we must decide how big or small the candidates should be for the key variables and how close to the desired size is close enough. There are a variety of tools that can help.

For wearables, it is most cost-effective to start the design with a case in the center for the key variables. Then move out from the center if needed, to create additional sizes. There are more people close together in the center than at the edges or boundaries of a measurement, so a product designed for a center case will accommodate more people than one designed for a large or a small case. It will also be easier to find other people near the case for fit testing.

In addition, in the commercial world, it is often not cost-effective to accommodate the extremes. There are not enough sales to make it feasible to produce and stock extreme sizes. The most important area to accommodate is the middle where most people are clustered. One size may not fit all, but it might fit most if it is in the middle. We refer to that center size as the base size.

To estimate the relative size of an observation compared to all other observations, and to find the center, we use the *frequency distribution*. When we count the number of times an observation occurs, we are counting its *frequency*. When we count how often all observations occur, we are estimating the *frequency distribution*. A frequency distribution describes the number of observations for each possible value of a variable. A histogram or bar chart is the visual representation of the frequency distribution.

When we estimate the relative size of a measurement, we are estimating how many observations are smaller and how many are larger. When we have many observations, this counting can be tedious and time-consuming. Luckily, mathematicians and statisticians have provided us with estimates of common frequency distributions that enable us to avoid having to count, if we think it is reasonable to assume that our data follows one of the common distributions. The most common frequency distribution is the Gaussian distribution which is also called the Normal distribution. (Normal is the name of the distribution. It does not mean the opposite of abnormal nor that the

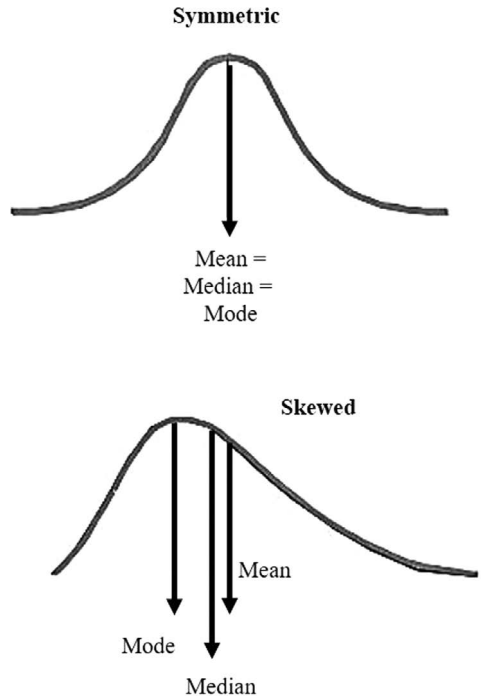


FIGURE 3.4 Mean vs. median vs. mode.

distribution is common.) The Normal distribution is a symmetric distribution that is shaped like a bell with a peak in the middle, as shown at the top of Figure 3.4.

Three different statistics indicate a middle point: the mean, the median, and the mode. Usually, when someone refers to the average, they are referring to the mean that is estimated by the sum of all observations divided by the number of observations. The median is the 50th percentile. It is the point at which half of the observations (50%) are smaller and half are larger. The mode is the peak of the observations' frequency distribution and is the point around which most are clustered.

In the true Normal distribution, all three of these middle points (mean, median, and mode), are the same value. In the real world, these three values are usually different. For example, if some observations are very large and are not balanced by very small ones, it can pull the mean away from the mode and the median to the right or to the large side. This type of distribution is said to be skewed to the right and an example is shown at the bottom in Figure 3.4. This is common for Weight and related variables, such as Waist Circumference, but less common for Stature and height variables. A distribution can also be skewed to the left, but that is less common for anthropometric variables.

The median is technically the actual middle, since half are below, and half are above. However, there can be fewer people clustered around the median than around the mode. The mode is the point where there is the greatest concentration of people. The mode is the easiest to visualize from a distribution plot. The mode, being the greatest concentration of people, is the best point for the base size. If raw data are available this can be

found by visually inspecting the frequency plots. The median is usually close to the mode for anthropometric variables, so it is a good second choice when raw data and the frequency plot are not available. The mean is good as the third choice.

The distribution for one variable is referred to as the *univariate* distribution. The distribution for two variables is called the *bivariate* distribution. The distribution for many variables is called the *multivariate* distribution. The two variable frequency can be visualized in a two-dimensional scatterplot we call the bivariate distribution plot or more simply, the bivariate plot. If we have two key variables their bivariate plot helps us visualize the bivariate mean, median, and mode and the relationship between the variables. If there are three key variables it can be difficult to visualize the trivariate plot because it is 3D. However, we can view the three bivariate plot combinations, variable 1 by variable 2, variable 1 by variable 3, and variable 2 by variable 3, provided we have the raw data to create the plots.

Higher dimensional multivariate distributions are more complicated to visualize. Fortunately, because many anthropometric variables are correlated, we typically don't need more than three to select cases.

In Chapter 2, we discussed some differences between having raw data versus aggregate data from a sample. Raw data is the sample itself, while aggregate data is only a description of the sample. Therefore, aggregate data is also referred to as descriptive statistics. If we have raw data, we can calculate aggregate data but with aggregate data, we cannot completely reconstruct the raw data. In addition, with raw data, we can:

- Tailor or adjust the sample to better represent our TP
  - create new groupings
  - apply statistical weights to the groupings
- Review the actual distributions of the measurements (rather than relying on the assumption of normal distribution)
- Calculate new variables from existing variables
- Examine relationships between variables
  - Correlations
  - PCA
  - Regression Analysis
- Use an individual from the Starting Sample as Case

With aggregate data we are limited to what is published, and the groupings that were summarized. This might not include our preferred statistics, such as the mode, or our preferred groupings, such as an age group or fitness level. Therefore, the preferred method of identifying the case candidates is by using raw anthropometric data. However, if raw data is not an option there are some tools for using aggregate data that can help get our first cases in the general ballpark. We will describe both starting with the preferred raw data option.

Selecting Cases with Raw Data

Selecting cases begins with selecting variables or measurements that are both relevant to the product and important to control. We also want to prioritize the variables based on their importance to the design. That will ensure that the most important ones are considered first. Some people like to throw every measurement available

into the mix, but this will just make the process of elimination take longer and be more confusing. More is not always better. A good rule of thumb is, "When in doubt leave it out". For example, Stature and Weight are nice variables for tailoring the TP sample, but they are not directly relevant or important for headgear or footwear case selection. Leave them out during case selection for those products. Use measurements that will directly relate to the size of the product such as Head Circumference for headgear or Foot Length for footwear.

The process of selecting key variables for case selection is a variable reduction process. The goal is to pare down the list of variables to two or three that control most of the important size variability.

If there are measurements that are directly relevant, but maybe not very important to control, they can be kept for the case selection analysis, but be given low priority for key variables. This might include things that we know will be accommodated with some adjustable features like a strap, but we want to keep an eye on them.

If, after eliminating all the measurements from our sample that are not directly relevant or are not important to control, we have just two or three left, those become our key variables. No need to pare down further because it is relatively easy to examine their combined relationships to select cases.

If, however, we have more than three relevant and important variables we need to study their relationships to pare them down further. The best way to do this is to select two or three that, when combined, will have a strong relationship with all the other important variables. A strong relationship between two variables indicates that one would be a good predictor of the other and if we control one, we will control a large part of the other. A weak relationship between two variables means they represent and control different aspects of size or shape variability.

We get a stronger combined control over all the variables if we select two weakly related variables to be key variables, rather than two variables that are strongly related. If the key variables are strongly related, they are representing the same source of size variability, so they are redundant. If, instead, they are weakly related to each other, but each is strongly correlated with other variables, we get better overall control. Each of the two variables represents or controls a different group of variables so we can control many aspects of size with just two variables. For example, Hip Circumference and Inseam Length are a better combination than Hip Circumference and Waist Circumference. The former set is not correlated with each other, but each is strongly correlated with other measurements. The latter are strongly correlated with each other and do not relate to any height or length variables. Therefore, we get more control with the first combination.

Companies that have a long history of product development or many existing products may have a good idea about what good key variables might be. Even for these companies it can be helpful to go through the process of examining relationships between variables to increase the understanding of just what is controlled by these variables and what is not. They may find that while the key variables they have been using are good, there are even better ones out there.

When we have raw data, we can calculate new statistics and create new plots to understand and represent relationships between variables including the anthropometric measurements. To understand relationships, we use *covariance* and *correlation*. The *covariance* is the measure of how much the deviation of one variable from its mean matches the deviation of another variable from its mean in its original

measurement units. In other words, it indicates whether if one variable gets larger the other variable gets larger as well. The deviation is measured using the variance for each variable, which is the squared standard deviation (Std. Dev.).

Since covariance is the measure of relationship in the original units of measure, a larger measurement, such as Stature, and a smaller measurement, such as Hand Length, would both have variances and covariances in millimeters squared. With covariance, a 10 mm change in Hand Length is treated the same as a 10 mm change in Stature when in reality a 10 mm change in Hand Length is a much bigger part of the total range of variability in the Hand than a 10 mm change in Stature. As a result, it can be difficult to understand the relationship by just looking at the covariance. Therefore, we use standardized variables and a standardized covariance which is called the *correlation* and is denoted as r or R.

Standardized variables are created from the original variables by first subtracting the mean from each value and then dividing by the Std. Dev. This removes the magnitude of the original variable so we can examine the relative change in one variable given the change in the other on a variability scale rather than an original unit's scale. The new value for each observation is a standardized value expressed as a Std. Dev. unit rather than the original units of measure. The mean of the standardized version of each variable becomes 0 and the Std. Dev. becomes 1. The covariance of the standardized variables is the correlation.

The correlation has a value between –1 and 1. Negative correlations indicate that as one gets larger the other gets smaller. Positive correlations mean as one gets larger the other gets larger as well. A zero correlation indicates one variable has no effect on the other, and this is called *independence*.

The correlation between two variables is called a *bivariate correlation*. Bivariate correlations between multiple variables are usually expressed in a table called a *correlation matrix*. A matrix in mathematics is a set of numbers arranged in rows and columns.

For example, we chose eight measurements we felt were relevant to a headgear product that will have a display and an attached mask or mouthpiece. These are shown in Figure 3.5.

Table 3.1 shows bivariate correlations between these eight measurements in a typical correlation matrix. This table shows the correlations for males from a North American

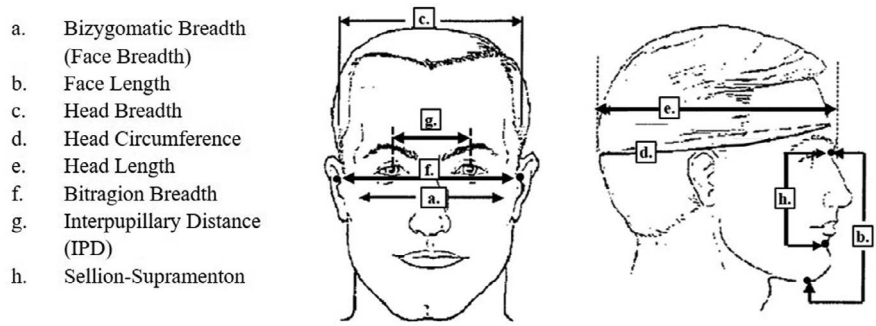


FIGURE 3.5 Eight relevant measurements for headgear example.

TABLE 3.1
Bivariate Correlations (R) for Eight Measurements for North American Males

	IPD	Head Breadth	Head Circumference	Head Length	Bitragion Breadth	Face Breadth	Face Length	Sellion-Supramenton
IPD	1	0.201**	0.234**	0.125**	0.226**	0.301**	0.133**	0.237**
Head Breadth	0.201**	1	0.479**	0.005	0.666**	0.708**	0.108**	–0.004
Head Circumference	0.234**	0.479**	1	0.788**	0.435**	0.455**	0.349**	0.167**
Head Length	0.125**	0.005	0.788**	1	0.072*	0.063*	0.338**	0.186**
Bitragion Breadth	0.226**	0.666**	0.435**	0.072*	1	0.778**	0.144**	–0.023
Face Breadth	0.301**	0.708**	0.455**	0.063*	0.778**	1	0.183**	0.032
Face Length	0.133**	0.108**	0.349**	0.338**	0.144**	0.183**	1	0.718**
Sellion-Supramenton Length	0.237**	–0.004	0.167**	0.186**	0.023	0.032	0.718**	1

* Significant at the 0.05 level (2 tailed).

** Significant at the 0.01 level (2 tailed).

raw data set (Robinette et al., 2002). The numbers down the center diagonal are all 1 because this is each variable's correlation with itself. The numbers to the lower left of this diagonal are the same values as those to the upper right. For example, the correlation between Interpupillary Distance (IPD) and Head Breadth (R = 0.201) appears in row one column two and in row two column one. The first is above and right of the diagonal and the second is below and left of the diagonal.

We created this correlation matrix using SPSS®, but it is also easy to do using Excel™ provided you have installed the "Analysis Toolpak". With this add-on data analysis is added to the data tab and correlation can be found in the data analysis pop-up window.

A significant test of each correlation is done to determine if it is likely to be a real correlation or not. In Table 3.1, four correlations are not significant at the $\alpha = 0.05$ level. We cannot be confident that these are correlated. These are the correlations between:

1. Head Breadth and Head Length (R = 0.005)
2. Head Breadth and Sellion-Supramenton (R = –0.004)
3. Sellion-Supramenton and Bitragion Breadth (R = –0.023)
4. Sellion-Supramenton and Face Breadth (R = 0.032)

The rest of the measurements appear to be correlated but that does not mean they are strong relationships. To better understand the strength of the relationship we use R^2. R^2 tells us the percentage of one variable controlled by the other.

For example, if we look at Head Length correlations, we see Bitragion Breadth and Face Breadth are significant (meaning they are likely to be correlated), but they are less than 0.1 at 0.072 and 0.063, respectively. The R^2 values for these two are 0.005 (or 0.5%) and 0.004 (0.4%). This means that even though correlated the correlation is too weak to be meaningful.

In contrast, the correlation between Head Length and Head Circumference is 0.788 and this R^2 is 0.621 (62.1%). This means 62% of Head Circumference is controlled by Head Length and vice versa. In other words, as Head Length gets larger, Head Circumference does too, so the amount of variability remaining for Head Circumference at any given Head Length is only 37.9% of the full range. Figure 3.6 is a plot of all the subjects in the raw data set (the dots). The line down the center is the best-fit linear regression line for predicting Head Circumference from Head Length. It represents the most likely value for Head Circumference at each Head Length and each point on the line is a mean value called the *regression mean*.

The other two lines are the 95% confidence lines showing the estimated range that 95% of the time will contain the Head Circumference at a given Head Length. The spread of Head Circumference points at any given Head Length value is smaller than the full Head Circumference range. The full range of Head Circumferences goes from 525 mm at the bottom left to 638 mm at the upper right. This is a range of 113 mm. If we look at the Head Circumference variability when Head Length is 200 mm, we see the range of Head Circumferences at this point is from 558 at the bottom 95% confidence line to 598 at the top 95% confidence line. This is a range of 40 mm. In other words, if we know the Head Length, we can better estimate Head Circumference than if we do not know Head Length, and vice versa.

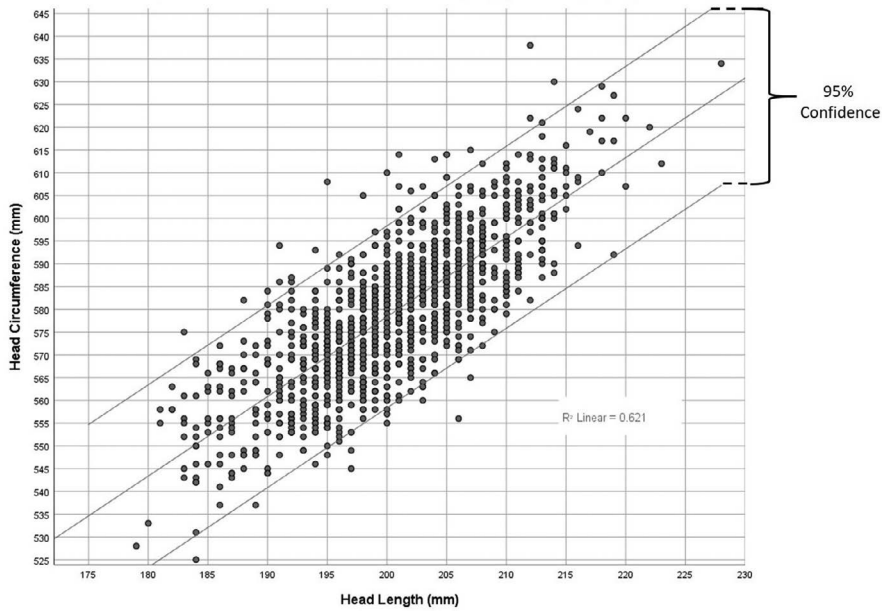


FIGURE 3.6 Bivariate scatterplot of Head Circumference by Head Length.

By comparison, the correlation between IPD and Head Circumference of 0.234, while significant at $\alpha = 0.01$, is small. The R^2 value is just 0.055 indicating only 5.5% of IPD variance is controlled by Head Circumference. The bivariate plot of these two variables is shown in Figure 3.7. In this plot, we can see that the scatter of subjects is almost circular with nearly no reduction in the spread of IPD as Head Circumference gets larger. Therefore, knowing Head Circumference does not allow us to substantially improve our estimate of IPD.

There is no consensus opinion about what is a strong correlation versus a weak one. We consider a correlation of R = 0.71 (R^2 = 0.50) to be a strong relationship since 50% (half) of the variability of one variable is controlled by the other. When we consider the fact that some of the variability for a measurement is always random or measuring error that will not be correlated no matter how strongly related, we know that 50% is substantial. We consider values between R = 0.45 (R^2 = 0.20) and R = 0.71 to be moderate. Anything below R = 0.45 is considered weak. These values are shown in Table 3.2.

We can use the bivariate correlation matrix to help us choose our key variables. As we stated earlier, when choosing key variables, we want each key variable to be strongly related to some variables, but weakly related or uncorrelated with each other. That will provide the most size variability control with just two variables. We also want them to be the high priority or most valuable measurements for our product.

For our headgear product, the cranial region is more important than the face, and the face length-related variables are only relevant to the location of the mask and this

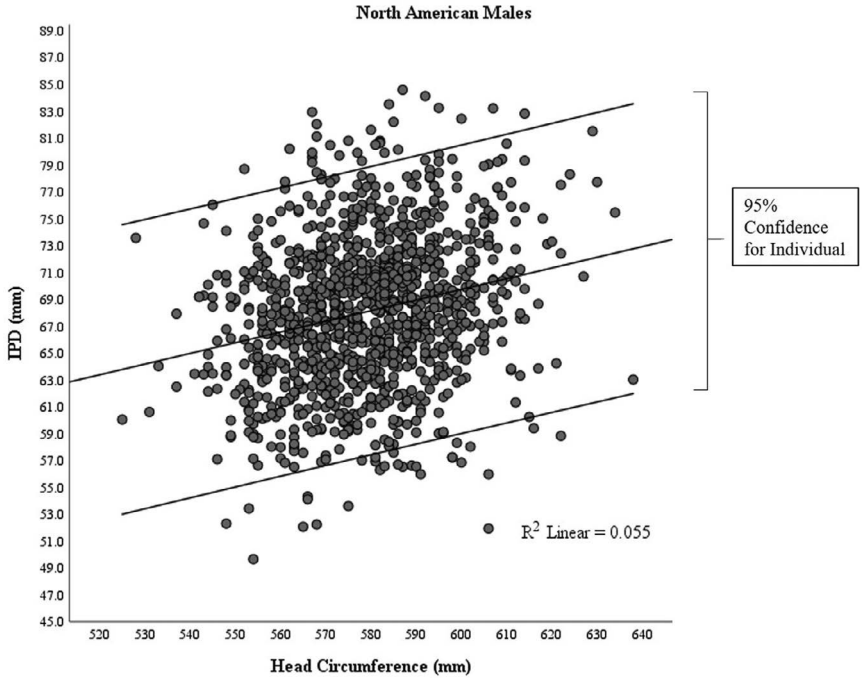


FIGURE 3.7 Bivariate scatterplot of Head Circumference by Interpupillary Distance (IPD).

will be adjustable. That means we need accuracy in the cranial region in our case, but more inaccuracy can be tolerated in the face lengths. So, while we will not use face length variables as key variables, we keep them in our set just to make sure the cases are not too extreme for these variables.

We noted earlier that Head Length was strongly correlated with Head Circumference, not correlated with Head Breadth and weakly correlated with Face Breadth and Bitragion Breadth. Of these three weakly correlated variables, Face Breadth is strongly correlated with two variables, Head Breadth (R = 0.708) and Bitragion Breadth (R = 0.778) and has the strongest (although still weak) correlation with IPD. If we select Head Length and Face Breadth for our key variables, we can represent most of the size variability of five of the eight variables.

TABLE 3.2
Strength of Correlation Values[a]

Statistic	Not Significant	Weak	Moderate	Strong	Very Strong
R	Any value	<0.45	≥0.45	≥0.71	≥0.90
R^2	Any value	<0.20	≥0.20	≥0.50	≥0.81

[a] Strength is zero if it is not significant.

In other words, we get a good idea about size variability for five variables with one bivariate plot, instead of the ten we would have needed to look at every two-variable combination of the five variables. In addition, we have slight control of IPD, although we should treat it as a third key variable since it is also deemed very important.

The two variables remaining that are weakly correlated with each of our first two key dimensions are the face length variables. Since they are not critical to our design, we can treat them as if they are independent variables that are not affected by our key variable sizes. In other words, we assume that no matter what the key variable sizes most Face Length values are equally likely. That means we can just examine the univariate frequency distribution histograms separately to influence our case selections, just to ensure they are not too extreme.

Base Size and Base Size Case

Once we have our key variables, we begin by finding the target key variable values for our base size case and selecting candidates who have those values. In those instances when we have collected the starting TP sample ourselves, we might already have candidate cases available to us. We can draw them from our sample. Also, some organizations have a subject pool, which is a group of people they can call upon for testing or analysis as needed. However, finding a person to be our case, such as a fit model, often requires recruiting a new person. Whether we already have the person or are recruiting a new one, the process is essentially the same. We start with the key variables' values to select candidates, then we examine the rest of their measurements and other factors. Since the key variables are both the most important ones and control most of the other important ones, the candidates selected with the key variables should be good for most things. That means when we are searching for candidates, we do not have to take all their measurements unless we choose them as a candidate. This helps us narrow down the possibilities quickly.

Selecting candidates using raw data has four steps:

1. Select the key variable sizes for our desired case size
2. Find and select candidates near this size
3. Examine them for other measurements and attributes
4. Choose the best from the candidates

For our base size case, we find the middle for our first two key variables, and the best indicator of the middle for selecting cases is the mode. The mode is the center of the peak of the distribution, which is the spot that has the most people. If we use this location, we will accommodate the most people in the first size. The mode is better than the median or the mean, but if the data are not skewed much the mode, median, and mean will all be about the same so medians and means can also be used.

To help us locate this spot, we like to use a graph that contains both the univariate histograms for our key variables and their bivariate plots as shown in Figure 3.8. The bivariate scatter plot is in the middle (with all the dots) and the two univariate

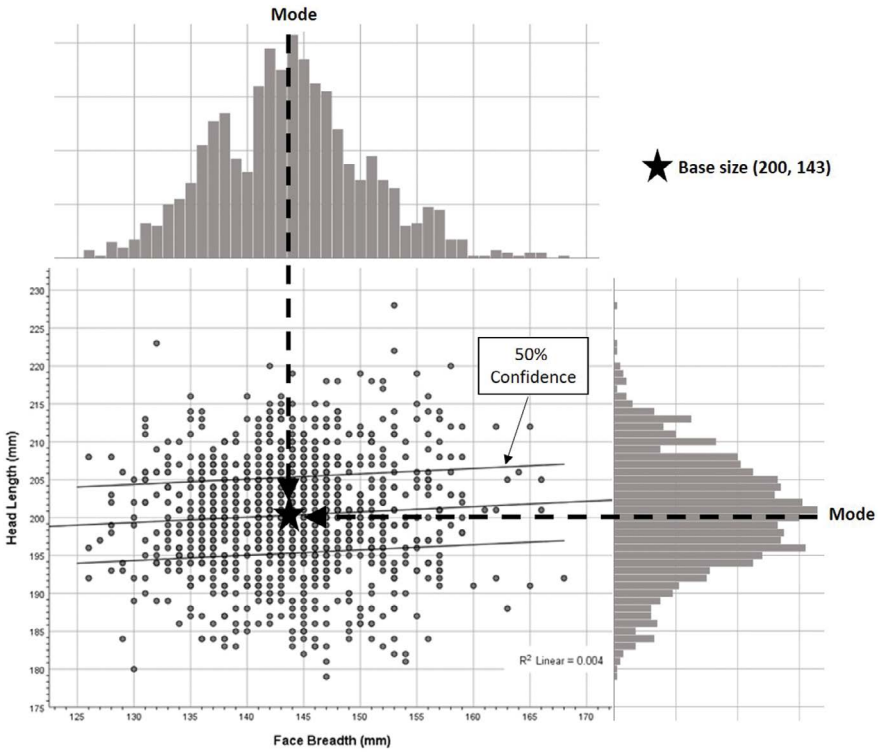


FIGURE 3.8 Bivariate and univariate plots of Head Length and Face Breadth.

frequency plots are on the right side for Head Length and above for Face Breadth. This figure was created by doing two univariate histograms and the bivariate scatterplot separately in SPSS®, and overlaying them in PowerPoint. This lets us visualize any skewness in the variables and see the mode locations.

The point at 200 mm for Head Length and 143 mm for Face Breadth is in the center of the peak for each separate variable. It also happens to be on the regression line indicating that it is the mean value for Face Breadth given the Head Length of 200 mm. This is our target for our base size case.

It can be very difficult to find a fit model who has exactly the desired measurements. Therefore, we need some way to evaluate how close is close enough. In this example, we used the 50% confidence interval (CI) for the individual prediction to help us gauge closeness. In this example, the 50% CI is approximately ±5 mm from the mean (the center line). The 10% CI is approximately 0.185 times the 50% CI (5 mm) for samples with n ≥30. This equals ±0.925 mm in our example. Since the measuring error for the Head Length measurement is about 1 mm anything smaller than that would be meaningless, so we round to the nearest 1 mm for ±1 mm. This indicates that about 10% of the population with a Face Breadth of 143 mm are expected to have a Head Length between 199 mm and 201 mm. That is a large enough percentage that we should be able to find a case in this range.

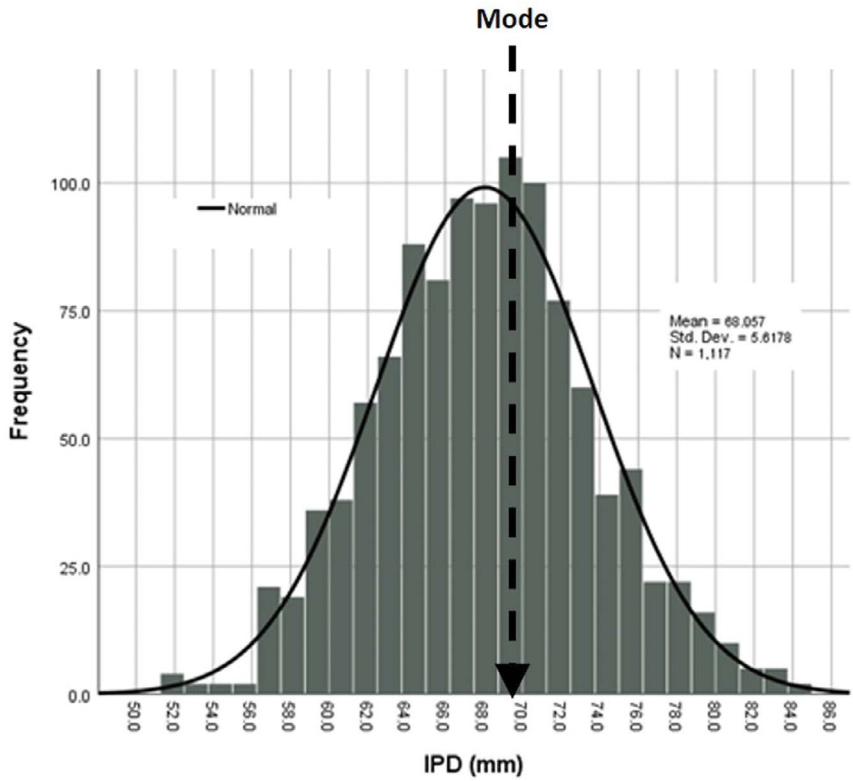


FIGURE 3.9 IPD histogram.

With two key variables, we have one bivariate chart to examine. With three key variables, it is best to start with the first two key variables to select a target size and look at the variability of the third key variable at that target size. If the third variable is not correlated or very weakly correlated, we can assume that the distribution of the third variable is the same at any size of the first two variables. Therefore, we can examine its univariate frequency distribution separately.

Our third key variable, IPD, is weakly correlated with Head Length and Face Breadth with a combined $R^2 = 0.097$. Therefore, it is reasonable to just look at its full range. The univariate frequency chart in Figure 3.9 indicates IPD is slightly skewed, so while the mean is 68 mm, the mode is 69.5 mm. We want to select a case with an IPD close to 69.5 mm if we want the best center to be represented. The Std. Dev. for IPD is 5.6178. If we choose ±1 mm as our acceptable range, the normal probability curve estimate for the percentage of people with 69.5 mm ± 1 mm (68.5 mm – 70.5 mm) will be approximately 14%. So, we should be able to find a case within this range.

If the third key variable has a moderate or strong correlation with the first two, we can use three bivariates, just as we used one for the first two. The three bivariates are variable one with variable two, variable one with variable three, and variable two with variable three.

Sometimes we don't want the target size for a key variable to be in the middle. This is dependent upon how we expect the product to fit and to be adjustable to fit. For example, for a pant we might want the circumferences to be in the middle, but we might want a slightly long leg length because the pant leg will be able to be shortened a lot, but not lengthened much. In other words, the adjustability is not symmetric. Similarly, in our headwear example, we might want the IPD to be smaller than the middle if we can widen the IPD adjustment, but not make it narrower. In that case, the accommodation of a narrow IPD in our base size will accommodate the wider ones too. The fifth percentile value for IPD is 58.8 or rounded to the nearest 1 mm equals 59 mm. The 95th percentile for IPD is 77 mm. If we have an asymmetric design accommodation issue, we might want to choose one of these. The point is, when selecting cases, it is important to consider some of the design expectations.

Once we have selected our target values for our base size case it is time to recruit someone to be our case. To find our candidates we have four steps:

1. Screen people for the key variables
2. Examine the remaining variables
3. Throw out any outliers
4. Look at other important characteristics for which we may not have measurements

If we want to start with candidates from our starting TP, we can start by finding those who are close to our target for the first two key variables and then look at the other measurements. If we need to find new candidates (which is often necessary when we are seeking live fit models), we start by measuring candidates for the key dimensions, eliminating anyone who is not within our acceptable range, take the other measurements for candidates who are in range, and examine their variability for the other measurements.

If we already have a fit model that we have been using, we should compare his or her measurements to the mode at this time. Even if we do not plan to change our fit model it is good to know where he or she falls with respect to our TP before we begin fit assessment for sizing. That will help us understand population accommodation issues and measure the cost versus benefit of keeping this model or replacing him or her with a better one.

When looking at the other measurements it is helpful to use multiple regression to predict the non-key variables from our key variables. This provides the prediction confidence range to help us learn if the candidate is an outlier or is a good representative for the other measurements as well.

Multiple regression procedures can be done in Excel™, but it is not as easy as it is in SPSS®. SPSS® uses pull-down and pop-up menus that guide the user through the process. For example, a procedure for predicting Head Circumference from our first two key variables is shown in Figure 3.10. We used the analyze menu, selected regression, selected linear, entered our variables using the arrow buttons, selected plots, and entered dependent (DEPENDNT) and adjusted predicted (ADJPRED) to

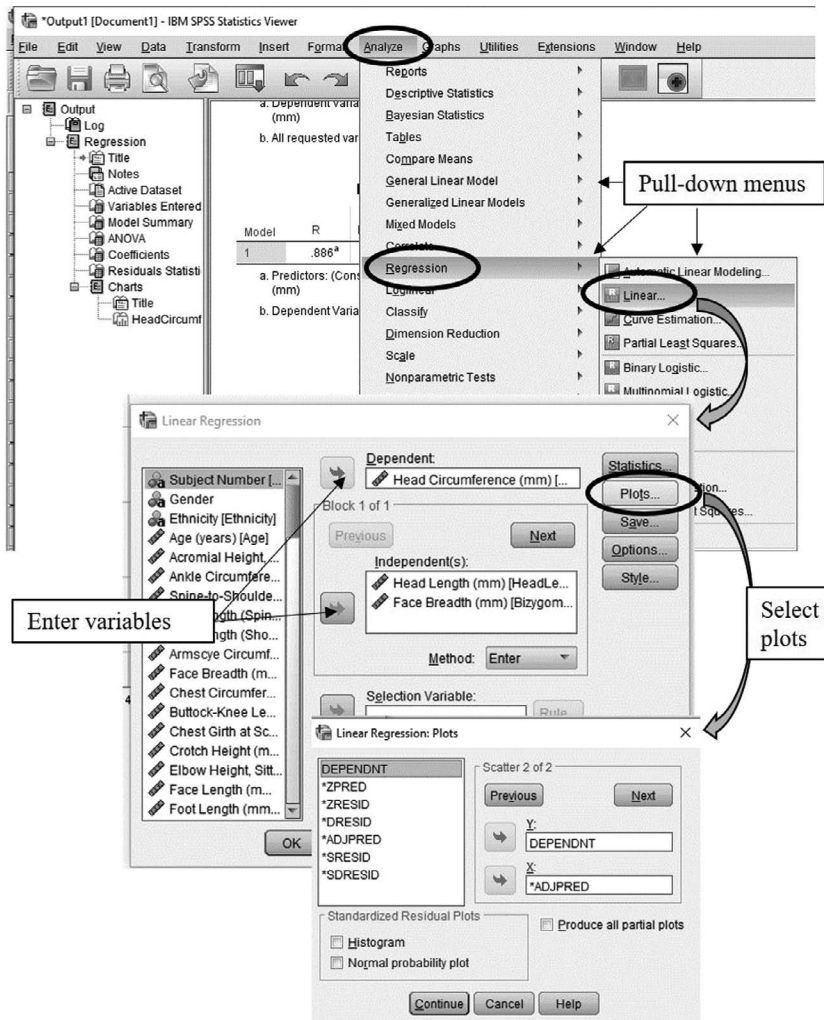


FIGURE 3.10 Example of regression procedure in SPSS®.

plot. This procedure gave us statistics about the regression equation, referred to as the model, and the equation itself. Some of these are shown in Table 3.3.

The model summary table provides the correlation (R), R^2, and the Standard Error of the Estimate. The Standard Error of the Estimate is used to estimate the CIs. The coefficients table provides the equation itself and the statistical significance of each term in the equation. In this example there are three terms, (1) a constant, (2) Head Length, and (3) Face Breadth. The equation is made up of the coefficients and the terms. In this example, the equation is 97.235 + 1.692 (Head Length) + 0.993 (Face Breadth).

TABLE 3.3
Regression Procedure Output

Part a. Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.886[a]	0.785	0.785	7.665

[a] Predictors: (Constant), Face Breadth (mm), Head Length (mm).
[b] Dependent Variable: Head Circumference (mm).

Part b. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	97.235	7.624		12.754	0.000
	Head Length (mm)	1.692	0.031	0.762	54.887	0.000
	Face Breadth (mm)	0.993	0.034	0.407	29.286	0.000

[a] Dependent Variable: Head Circumference (mm).

The procedure also gave us a bivariate plot. There are several plotting options available, but we chose the original variable values (DEPENDNT) by the predicted values (ADJPRED) because we feel it is the easiest to understand for case evaluation. It resulted in the plot shown in Figure 3.11.

The predicted values are shown along the horizontal axis and the actual values along the vertical axis. We used the plot editor in SPSS®, to include the regression line and the 95% CI for the individual prediction. The middle line is the regression mean or the most likely value given the input variables' values, and the upper and lower lines indicate the range of values likely 95% of the time. This indicates that a case's Head Circumference should be within ± 15 mm of the regression mean (the center line) 95% of the time for any combination of Head Length and Face Breadth. In our example, the Head Circumference predicted from a Head Length of 200 mm and a Face Breadth of 143 mm is 577 mm. Therefore, 95% of the time the true Head Circumference is expected to be between 562 mm and 592 mm given our target Head Length and Face Breadth values.

Continuing with the headgear example, we are selecting someone from our starting TP sample. There are tools in some software packages that allow us to quickly identify subjects in a region of interest. Figure 3.12 illustrates a target feature in SPSS® that allows us to identify the case numbers of people who fall at any point in the bivariate plot. Here we see the several case numbers at the central point we targeted including 294, 348, 490, 711, 759, 1088, 1278, etc.

We started with two of these, subjects 490 and 1088, and we took their other measurements. The values for all eight of their measurements are provided in Table 3.4. We see that both exactly match our target size for the first two key variables. And case number 1088 has an IPD within the acceptable range for the mode (68.5 mm to 70.5 mm). Case No. 490 has an IPD a bit on the small side.

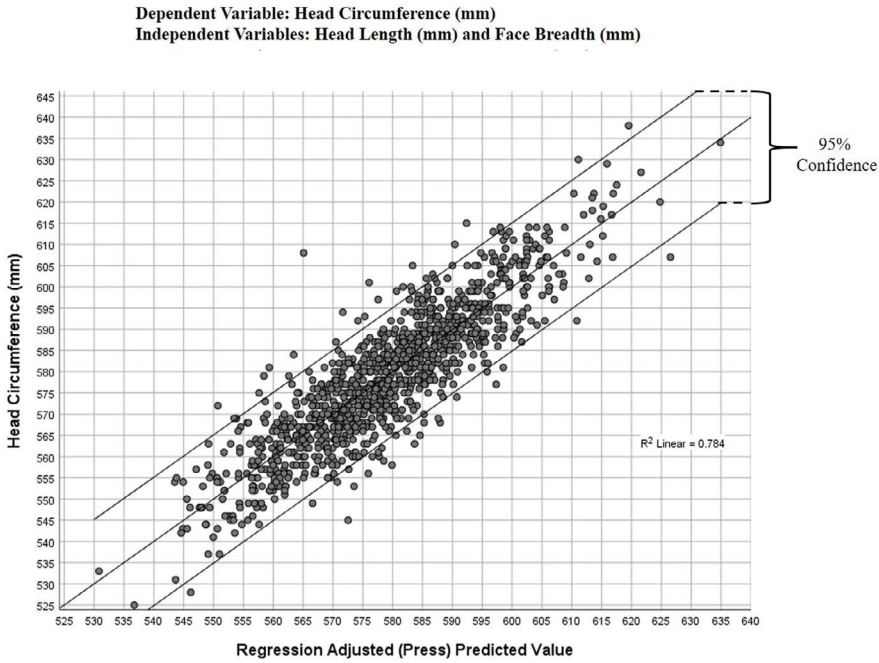


FIGURE 3.11 Example of predicted versus actual plot from linear regression.

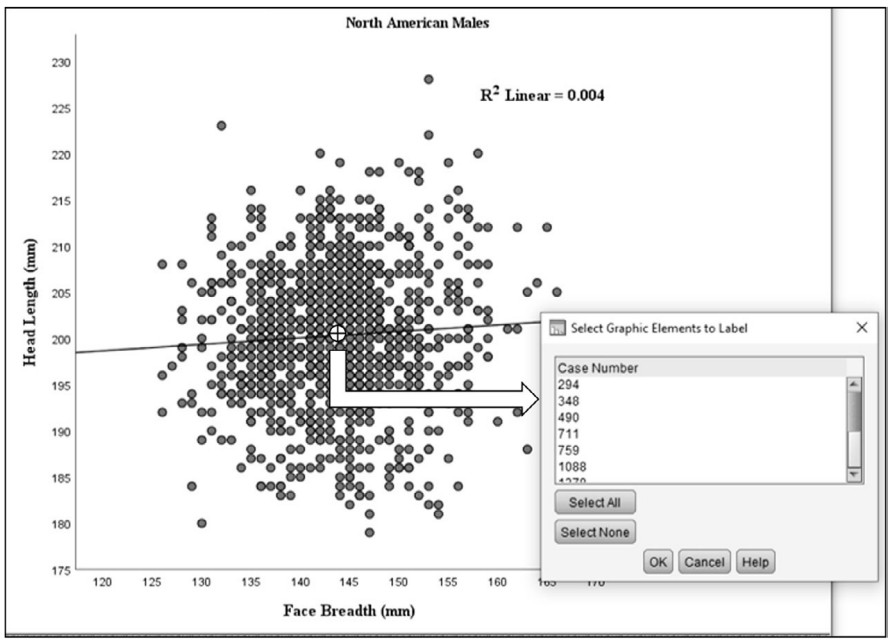


FIGURE 3.12 Finding candidates from a bivariate chart in SPSS®.

TABLE 3.4
Measurements for Two Male Case Candidates

Case No.	Interpupillary Distance (IPD)	Head Circumference	Head Breadth	Head Length	Bitragion Breadth	Face Length	Face Breadth	Sellion-Supramenton
490	65	575	158	200	149	119	143	101
1088	70.1	582	158	200	151	124	143	96

Next, we examine their size for the other measurements using multiple regression. We predict each non-key measurement from all three key variables and look at the range. This gives us five regression equations, one for each of the non-key variables. The non-key variable is the dependent variable, also called the predicted variable and the key variables are the independent variables, also called the predictor variables. The equations are:

Head Circumference = 96.253 + 0.981 * Face Breadth + 1.688 * Head Length + 0.052 * IPD
Head Breadth = 70.942 + 0.636 * Face Breadth – 0.032 * Head Length – 0.009 * IPD
Bitragion Breadth = 19.166 + 0.885 * Face Breadth + 0.027 * Head Length – 0.016 * IPD
Face Length = 30.60 + 0.161 * Face Breadth + 0.319 * Head Length + 0.062 * IPD
Sellion-Supramenton = 54.867 – 0.052 * Face Breadth + 0.159 * Head Length + 0.304 * IPD

Entering our cases' Face Breadth, Head Length, and IPD into these equations we get the predicted values shown in Table 3.5 for the other five variables. We also entered the target values which are 143, 200, and 69.5, respectively. The predicted values are different than the actual measurements for our cases, but that only indicates they are smaller or larger than the predicted mean for that variable. For Head Circumference, Head Breadth, and Bitragion Breadth we see that the predicted values for our two cases and our target input are all the same. That happened because all three had the same values for Face Breadth and Head Length and these two key measurements were strongly related to the first three non-key variables.

TABLE 3.5
Predicted Mean Values for Cases and Target

Predicted Values	Head Circumference	Head Breadth	Bitragion Breadth	Face Length	Sellion-Supramenton Length
Subject 490	578	155	150	121	99
Subject 1088	578	155	150	122	101
Target Input	578	155	150	122	100

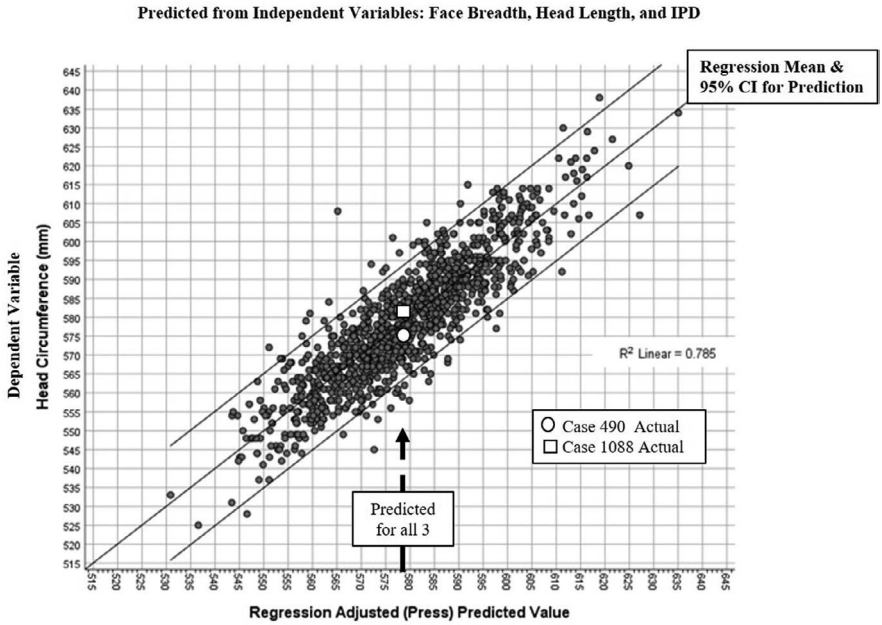


FIGURE 3.13 Candidate cases with respect to Head Circumference distribution.

To understand how where they fall with respect to the rest of the sample, we place them on plots of predicted versus actual values beginning with Head Circumference in Figure 3.13. The Head Circumference plot looks nearly identical to the previous Head Circumference plot in Figure 3.11 because although we added IPD to the predictor variables (independent variables) IPD has only a weak correlation with Head Circumference, so it had little impact on the outcome.

Figure 3.13 shows the location of the actual Head Circumference values for our two cases versus their predicted values. Also shown are the regression mean line (the center line) and the 95% CI for the prediction. The predicted values are the expected values (the regression mean values) given the values for the three input (independent) variables. The 95% CI tells us the estimated range of values around the regression mean that will include 95% of the population.

In the Head Circumference plot we see that the predicted values for both cases and the target, (indicated with an arrow saying "all three") are all the same. However, the actual values for the two cases, indicated with a circle and a square, are a bit different. Case 490 is 3 mm smaller than the regression mean and case 1088 is 4 mm larger. For a large measurement like Head Circumference, with a 115 mm range, these are small differences, and we can see that both are in the cluster of subjects in the center. Therefore, they are not unusual outliers, and both should be good cases for Head Circumference. We might prefer 1088 if being a little bit larger might be expected to accommodate the smaller people.

The plot of Head Breadth is the next and shown in Figure 3.14. Here the predicted values are again all the same, but in this instance, both cases have the same actual

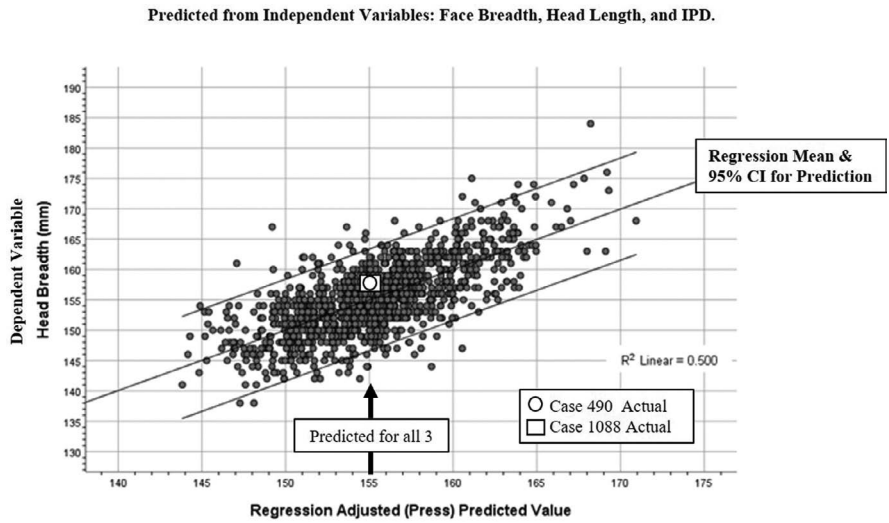


FIGURE 3.14 Candidate cases with respect to Head Breadth distribution.

values as well. This is 3 mm above the mean but also in a good cluster of subjects. Neither case is an outlier.

For Bitragion Breadth (see Figure 3.15) the actual values are different, but both are just 1 mm from the mean. Case 490 is 1 mm smaller and case 1088 is 1 mm larger. This is within measuring error given that the smallest unit on the caliper used

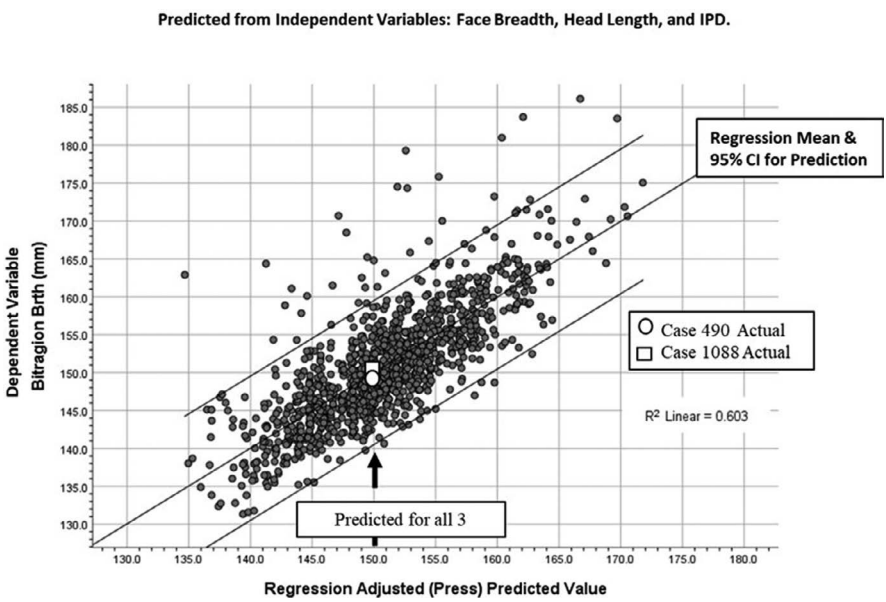


FIGURE 3.15 Candidate cases with respect to Bitragion Breadth distribution.

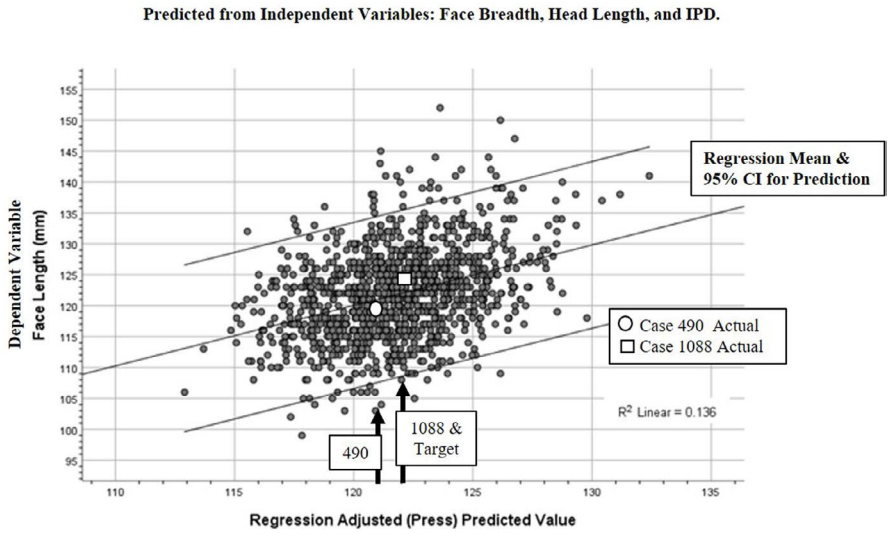


FIGURE 3.16 Candidate cases with respect to Face Length distribution.

to measure them is 1 mm. Therefore, these differences will not be significant. Both are good representatives for this measurement as well.

The last two measurements were not strongly correlated with the key variables, but we were less concerned about accommodating them. For weakly correlated variables we could have just looked at the univariate frequency plot, just as we did for IPD. However, the regression plots are also helpful and sometimes it is easiest to just do the same plots for everything. The plot for Face Length (see Figure 3.16) shows our cases are again in the sweet spot in the middle of our distribution. The actual values are just 2 mm smaller for case 490 and 2 mm larger for case 1088. Case 1088's predicted value is the same as our target input predicted value and this case is more centered in the cluster of subjects. However, both cases have reasonable Face Lengths. They are not extreme outliers.

For Sellion-Supramenton (see Figure 3.17) Case 490 is just 2 mm larger than the predicted value, but case 1088 is 5 mm smaller. While this size is not as close to the predicted as the other variables, it is not out in the fringes near the 95% CI either so we would not consider it an outlier. Statistics help us make informed decisions. However, we still must make the final decision based on a judgment call which may well be dependent on the features of the product.

In summary, both of our first two cases have good measurements. The decision about which case to use must balance measurements, demographics, and conformity to the perception of the target market or fit intention. Some of these things are not adequately captured by measurements alone. For example, the candidate might have an odd bump or oddly shaped feature that is best discovered visually. The "look" is something that is best assessed visually in person or using a 3D body scan or 2D photographs.

Both 490 and 1088 are both good candidates for all their measurements. What if we hadn't selected our key variables using correlations? How bad could it be? In

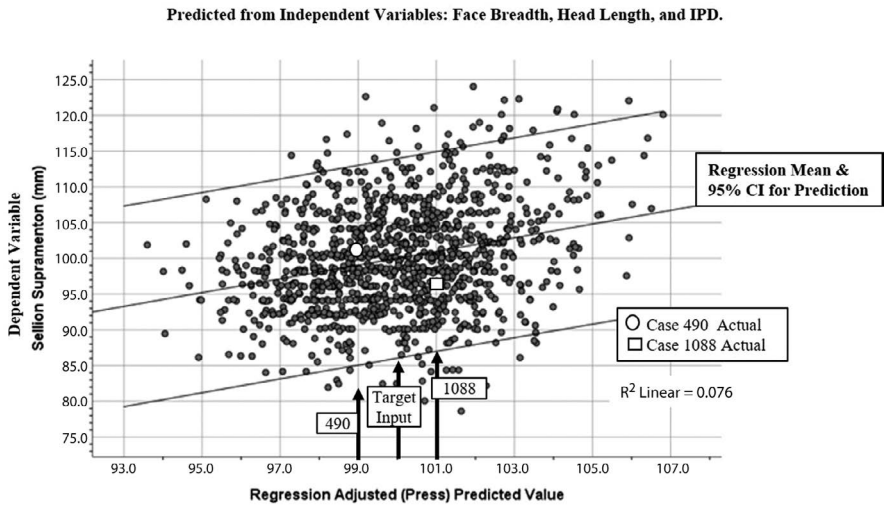


FIGURE 3.17 Candidate cases with respect to Sellion-Supramenton distribution.

Table 3.6 we see the measurements for two cases, one chosen using the Face Breadth mode (case 1M), and the other for Head Breadth mode (case 3M).

We plotted their sizes for our two key variables in Figure 3.18. We see here that both have Head Length values that are outside the 50% CI range. Case 1M is 6 mm larger than the Head Length mode (200 mm), even though he is exactly at the Face Breadth mode. Case 3M is 6 mm smaller than the Head Length mode. Similar extremes occur for some of the other measurements as well. Case 1M has an IPD that is 10.5 mm smaller than the IPD mode (69.5 mm) and 3M is 11.5 mm larger. Case 1M has a Head Circumference that is 8 mm larger than the Head Circumference mode (578 mm) and 3M is 7 mm smaller. Sellion-Supramenton values are by far the worst. 1M has a value near the 5th percentile for Sellion-Mention and 3M is near the 1st percentile!

Of course, we can always reject these candidates and keep looking. However, by selecting our key variables using relationships and capturing much of the total variability with two variables we got to our good cases quickly. It made the search much faster and more effective. Without this, we might have to go through hundreds of candidates to find one that was acceptable.

Now we must choose between them. For this, we look at additional factors that might not be captured in our measurement set. Then we must compare the candidates

TABLE 3.6
Two Cases, One Chosen for Face Breadth Alone and Another for Head Breadth Alone

Case No.	Interpupillary Distance (IPD)	Head Circumference	Head Breadth	Head Length	Bitragion Breadth	Face Length	Face Breadth	Sellion-Supramenton
1M	59	586	148	206	146	113	143	87
3M	81	571	158	194	167	97	156	82

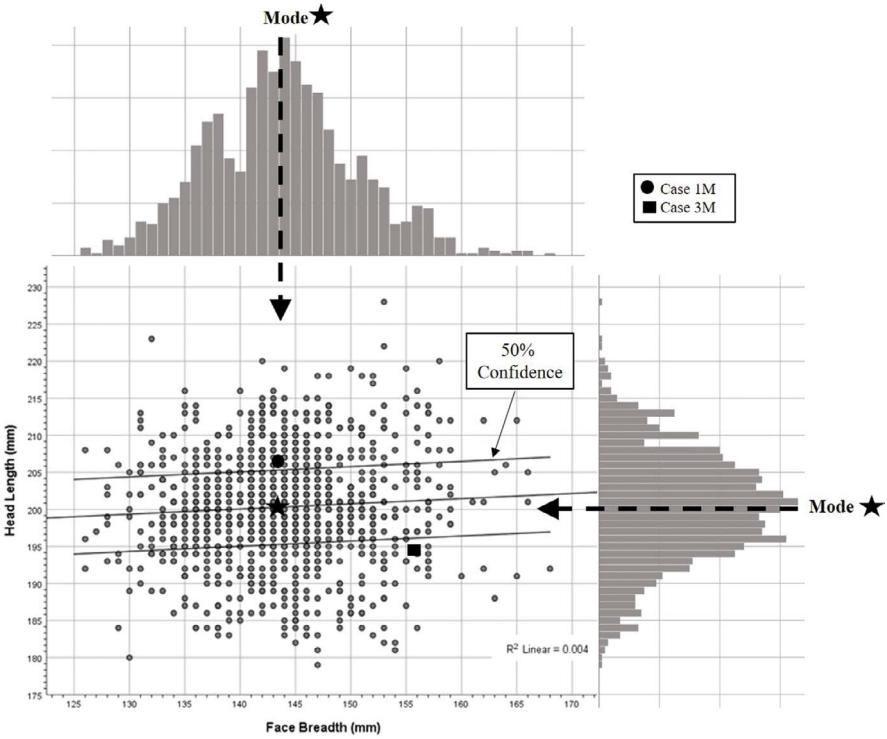


FIGURE 3.18 Plot of candidate cases chosen without the help of correlation for key variable selection.

with the expected product features in mind. The product will not perfectly match the body measurements and the methods to accommodate the body will vary depending on the way the product will be worn and what aspects are rigid, flexible, tolerant, or adjustable.

Before selecting cases, it is important to look at them visually for things that our minds can understand but can't always capture with a measurement. These are things like surface smoothness, asymmetry, unusual contours, etc.

Symmetry is important, so it is a good thing to check before selecting a case. If scanning is not an option, then taking some additional measurements on both sides of the body might help. Everyone is somewhat asymmetric. Our hearts tend to be on the left side of the body and most of us have a dominant hand, dominant eye, and dominant ear, all of which contribute to asymmetry. Therefore, asymmetry cannot be avoided, but it is good to try to select someone who has minimal asymmetry.

In the 1980s we were told a story, that we have reason to believe was true, about a man who invented a new kind of lower body anti-gravity (g) suit that was dramatically improving his ability to withstand high gravity forces. He built the first prototype for himself and tested it on himself. It performed so well that more were made in his size and additional sizes. When they were finally tested on other people it was found that they did not fit anyone else. It turned out that he had one leg substantially

shorter than the other so the openings for the knees that allowed the knees to bend and pressure to be applied correctly to the thigh and calf were severely asymmetric. As a result, no one could get both legs to fit them at the same time. All the costly, specialized suits had to be re-made. Without fit testing, errors often go into production and are never found. The only indication there is a problem is the product fails to sell well. Since there can be many reasons for a product not selling well, the manufacturer may never know they had a fixable problem.

We didn't have the subjects on hand for this example, so we looked at their scans. These were whole-body scans so the detail on the head and face is limited, but it was enough to help in case selection. The actual person and additional more detailed scans will be needed to represent the case for modeling and prototyping the product. Screenshots of the heads of cases 490 and 1088 are shown in Figure 3.19.

Comparing these two subjects visually we see that subject 1088 has a more even head shape and he is more symmetric around the nose, eyes, and chin. Subject 490 has a large somewhat crooked nose and his eyes are more deeply set and close together. Also, his head seems to have a very prominent and somewhat flat forehead that looks unusual. Given all the information we would choose subject 1088 to be our case.

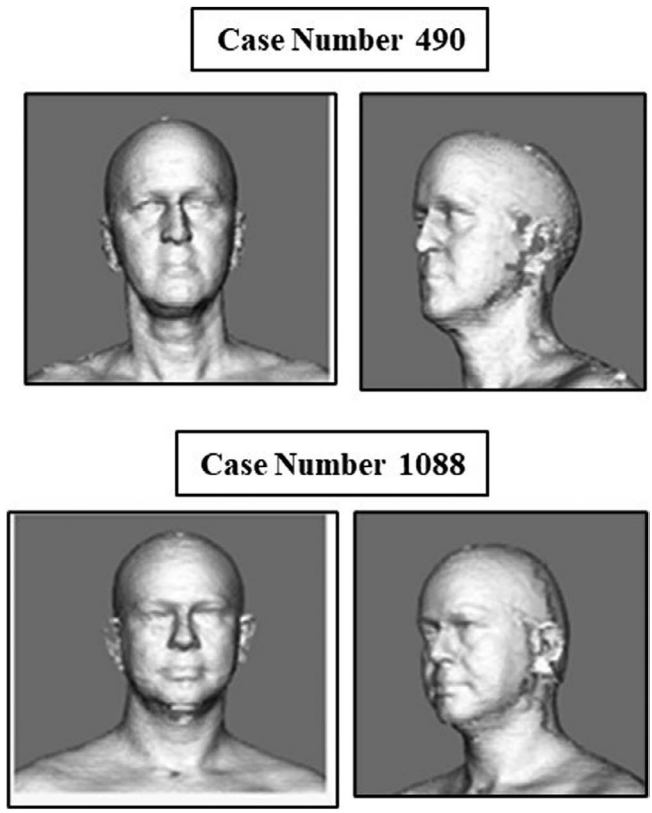


FIGURE 3.19 Screen shots of cases 490 and 1088.

At this juncture, there are still some people who propose that we average the cases rather than use an individual. Averaging is not advised and not helpful. Whether we are taking an average of a whole sample or an average of a few individuals, the result is the same. As shown in Figure 3.2, the average will represent no one, it will smooth away important details, and end up the wrong size and shape.

Multiple Cases

Thus far in this chapter, we have discussed how to get started with the base size and the base size case. This enables us to produce our first mock-ups and prototypes to begin the design loop testing. Fit testing in the design loop is using a Systems Engineering approach to product development and is the ideal way to build a new product. It allows us to evaluate the product's properties and form factors as we design it using the human as part of the system which enables us to produce the best designs for the wearers. It reveals the measurements that are the best predictors of good fit to serve as key variables (these can be different from the variables we started with), as well as the range of fit within the base size. This information indicates the additional sizes and/or cases we might need. This testing and sizing analysis is the subject of Chapter 4.

Selecting multiple cases for predetermining the sizes before fit testing is the source of most of the problems with sizing today. Unfortunately, in the real world, our customers may insist we create products in multiple sizes before doing any fit testing. There is a common misperception that if we just analyze the anthropometry and put it into a magic analysis box out will poof our sizes. We know anthropometry is not enough and people do not come in discrete sizes. The product enforces the sizes. However, we must please our customers if we want to stay in business, so it is important to have a plan for arriving at additional sizes without fit testing. Yes, it may waste time and money, but sometimes letting the customer make the mistake allows us to illustrate the improved process for the next product or iteration.

There are a few other reasons why we might want some additional cases before we move to the design loop. Sometimes, we want a few extra cases to use to examine our design concept digitally before we build our first mock-up or prototype and start the fit testing. They can help us get an idea about the fit of the base size on other body sizes before we produce our first prototypes to help us narrow down our design options. We may also want a few extra live cases for our first pilot test and want to choose them wisely.

There are two methods for selecting additional cases: (1) grading and (2) distributing. Both methods use the key variables we selected. Grading scales the base size case up and down along a line. Distributing selects new cases spread throughout the key variable distribution. Examples of each method are shown in Figure 3.20.

The grading method is the one used in the apparel industry for creating additional sizes in mass-produced products. It is implemented in the product by modifying the base size pattern from size to size rather than by creating a completely new pattern for each size. The grades are scaling amounts to move from one size to the next, and they are part of the fit standard.

The grading method is cost-effective for sewn products and maybe for other types of products as well for several reasons. Perhaps most importantly, if there is a general

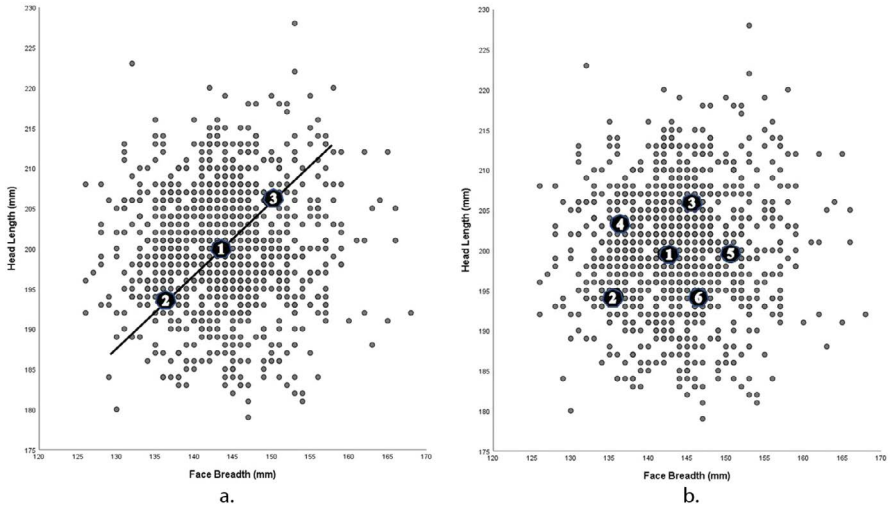


FIGURE 3.20 Graded cases versus distributed cases. (a) Graded cases fall along a scaling line. (b) Distributed cases.

fit issue with the size proportioning or fit adjustment mechanisms it can often be fixed for all sizes by fixing the base size. By contrast, if each size is produced from a different case, there can be different fit issues for each size that are not related to the sizing but simply because a different person was selected to produce it. Therefore, fixing proportioning or adjustment mechanisms must be done differently for each size which costs more and takes more time.

What has been missing in the apparel industry is the evaluation in the design and sizing loops to ensure the product is fitting who it is supposed to fit, the way it is supposed to fit, the size range is accommodating the TP, and with minimal size overlap or waste. Sizing loop testing will also indicate if more than one size range, with a separate base size and grade, is needed. Therefore, for effective and sustainable sizing, we must do design and sizing loop testing before determining the sizes or finalizing our fit standard. Before the design loop, we may use grading to select additional cases to help in the building of the first prototypes, but it is unwise to predetermine the sizes.

There are two types of grades: a body grade and a product grade. The *body grade* is a chart of scaled body measurements, starting with the measurements for the base size case. The *product grade* is a chart of scaled product measurements. These are not the same. Before we get to the design loop, we do not have a product so the grade we are using for selecting cases is the body grade.

Instead of scaling our base size, distributed cases are completely new cases, chosen in much the same way as the base size case. A distributed method is sometimes used for fit testing, because it spreads out the test subjects better than a simple random sample. There are many ways to distribute the selection of test cases. The two most common are: (1) to evenly space the cases throughout the range and (2) to use sample stratification. *Stratification* is the dividing of the population into groups and

selecting samples from each group. For some new products, for which we do not have a sizing history, selecting cases in a distributed way can help us learn how to adapt the product to accommodate the TP, whether that be by adding adjustability or by creating additional sizes.

We might also use a combination of grading and distributing cases, where we select completely new cases, but along a grade line. This might be necessary to get a good 3D digital model for use in examining the base size in CAD.

Sometimes knowing the key variable sizes is enough for selecting cases for our predesign loop evaluation. However, we often want to know the other body measurements. If, for example, the product has sensors or features that need to be in precise locations with respect to the body, it is good to check the other measurements. We use the regression equations and estimate the additional measurements from the key variables, just as we did for the base size. The only difference is key variable input values.

Continuing with our headwear example, if we use the cases identified in Figure 3.20 we get the sets of cases shown in Table 3.7. We used the cases from the Face Breadth by Head Length bivariate chart along with the IPD mode as input to the regression equations for the other variables.

The output is the regression mean for each variable. We can use these values or select subjects who are near them to be the cases we use. Selecting a subject is advisable for most situations, such as if a 3D model is desired for CAD visualization. The mean values should add together, but this is an average that works just like the overall average. As shown in Figure 3.2, the average shape may not be like anyone's and may be too small for everyone in some way. Perhaps it will not be a large amount and will still work, but we cannot know without testing. In addition, for some measurements, we may want the case to be smaller or larger than the mean, and this difference affects other measurements. If we use values other than the mean without selecting a person with those values, the case may have body measurements that do not fit together. If we select an individual who has the desired properties, we

TABLE 3.7
Example Headwear Cases

		Input Variables			Output Variables Regression Means				
	Case No.	Face Breadth	Head Length	IPD	Head Circumference	Head Breadth	Bitragion Breadth	Face Length	Sellion-Supramenton
Graded Cases	1	143	200	69.5	578	155	150	122	100
	2	137	193	69.5	560	151	145	119	100
	3	149	207	69.5	595	158	156	125	101
Distributed Cases	1	143	200	69.5	578	155	150	122	100
	2	135	194	69.5	560	150	143	119	100
	3	145	207	69.5	592	156	152	124	101
	4	137	247	69.5	651	150	146	136	108
	5	152	200	69.5	587	161	158	123	100
	6	146	193	69.5	569	157	152	120	99

TABLE 3.8
Example of Grade Rules

		Grade Rules		
	Case No.	**2**	**1 (Base Size)**	**3**
Input Variables	**Face Breadth.**	–6	143.0	6
	Head Length	–7	200.0	7
	IPD	0	69.5	0
Output Variables	**Head Circumference**	–18	578.0	18
Regression Means	**Head Breadth.**	–4	155.0	4
	Bitragion Breadth	–5	150.0	5
	Face Length	–3	122.0	3
	Sellion-Supramenton	–1	100.0	1

know we have something that will fit together and represent a real person. We select the cases at these new key variable points the same way we selected them for the base size case.

For the graded cases the cases are usually listed as grade rules, which indicate the change from size to size starting with the base size, as shown in Table 3.8. Here we see the base size case in the middle with the smaller case to the left and the larger case to the right.

How many cases should be chosen and where should we place them? We do not know. At this stage, they are guesses or theories, perhaps educated guesses, but still un-tested. This is the reason we need design and sizing loops to determine the sizes. It would be just as reasonable at this stage to select a range of Face Breadth values and keep both Head Length and IPD at their mode values. Regardless of the method used to select cases, it is best if the fit assessment is done with live people cases before any designs or sizes are finalized. A CAD or virtual case assessment is not real. Virtual models are never uncomfortable and cannot move or use the product, the lack of real user feedback that can lead one down a rabbit hole of mistakes.

This is a Systems Engineering approach and we must consider the human wearer to be part of the system. When we evaluate the cost, schedule, and quality within the trade-off space, we must use the wearer to provide quality feedback to arrive at a cost-effective, good-quality product.

The good news is, at this point we are only selecting additional cases to help us judge our design concept before we build or modify our mock-ups and prototypes. Therefore, the number and location of cases are not too critical.

Some engineering and design guides propose using boundary cases. *Boundary cases* are cases at the extreme edges (boundary) that we wish to accommodate. This is a type of distributed case. For one measurement the boundary case is represented by people who have specific percentile values, such as the 5th and 95th percentiles. In two dimensions the boundary is represented by an ellipse that contains a certain percentage of the sample and boundary cases are selected around this ellipse. A 95% ellipse indicates the area for two dimensions that contains 95% of the observations for those two dimensions. An example using our head key variables is shown in Figure 3.21. Eight

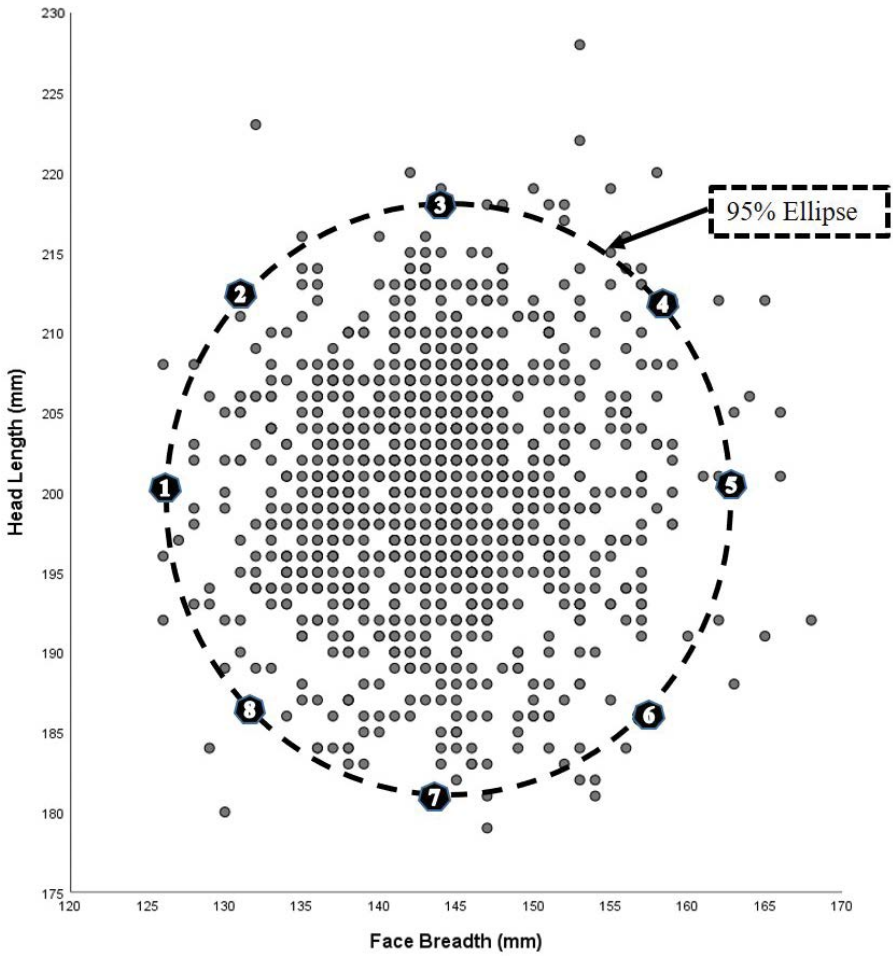


FIGURE 3.21 Boundary cases.

cases have been selected around the boundary 95% boundary ellipse. It is not possible or reasonable to use the entire boundary for two or more dimensions.

Sometimes we might use the boundary ellipses as a guide for selecting subjects for fit testing, just to ensure we have a good spread of subjects in our sample without getting too extreme. The boundary serves as a limit for sampling.

In general, however, boundary cases are not well-suited for wearable products. Their use has proven very useful for cockpits, and adjustable seating because those products have continuous movement adjustability. For example, seats move up and down continuously and can be stopped at any point between the top and bottom. In these situations, if we accommodate the boundary, we also accommodate the middle. This is not true for wearables, because the change in size is usually a completely new item. Wearables usually come in sizes and only rarely does one size accommodate the whole range from smallest to largest. Helmets, gloves, boots, pants, shirts, vests,

etc. with stringent fit requirements usually only accommodate a small range of body sizes within one size. Even very stretchy garments such as t-shirts, need multiple sizes to accommodate a population.

For wearables, the boundary cases can be accommodated without accommodating most of the population, since most of the population is in the middle and tightly clustered around the mode. We refer to this problem as the "hole in the donut", a phrase coined by our friend Keith Hendy (personal communication 1995) because the accommodation or fit range looks like a donut with a hole in the middle. We illustrate this problem in Figure 3.22. Here we show circles around each of our eight cases, each with a 10 mm diameter, simulating a range of fit for each case. The total percentage accommodated within this range of fit for the boundary cases is just 8% or about 1% per case on average. In other words, we could accommodate the range of fit for all eight cases and fit nearly no one within the middle 92%. Accommodating these extremes would not be cost-effective. We might not want to produce any sizes out there. If we instead used our base size case in the middle, we would accommodate 65% of our population with just one case.

Another issue with using boundary cases is the choice of boundaries. How far out is far enough and what is the associated cost to achieve an increase in the percentage

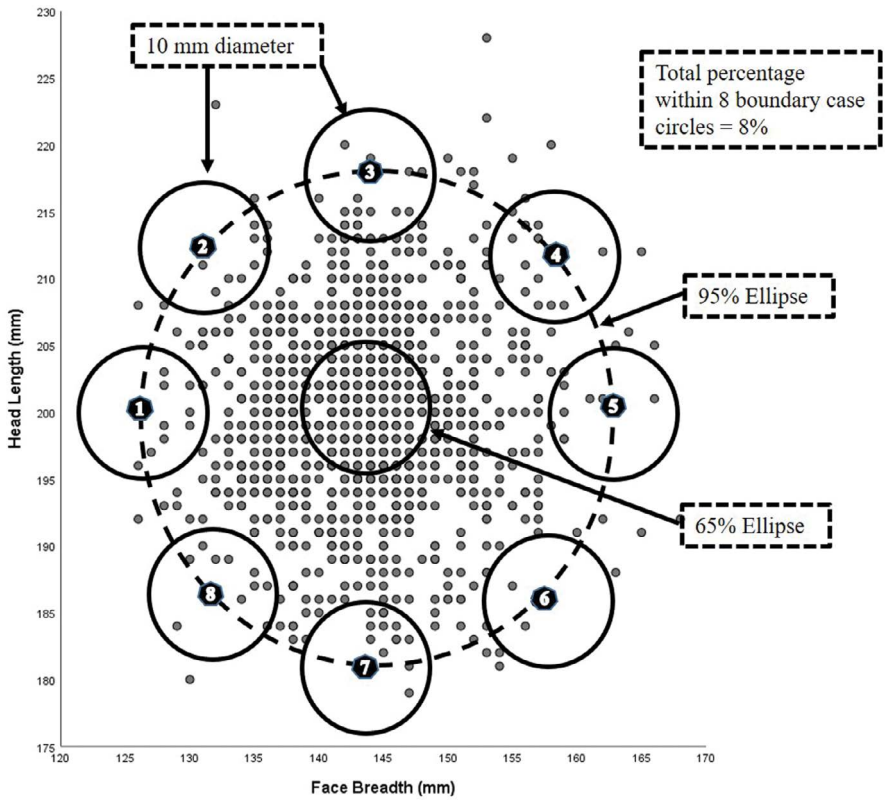


FIGURE 3.22 Boundary cases and the hole in the donut.

TABLE 3.9
Cost Versus Benefit for Increasing from 80% Accommodated

Boundary Percentiles		Percentage	Boundary Percentile Values (cm)		Adjustment Range Length (cm)	Increase in Range Length (cm)	Cost (Percentage of Length Increase)	Benefit (Added Percentage Accommodated)
Min	Max	in Range	Min	Max				
10%	90%	80%	87.2	112.8	25.6	0	0	0
5%	95%	90%	83.6	116.4	32.9	7.3	29%	10.0%
1%	99%	98%	76.7	123.3	46.5	20.9	82%	18.0%
0.10%	99.90%	99.80%	69.1	130.9	61.8	36.2	141%	19.8%

accommodated? To illustrate we use a simple one-dimensional (1D) example. We have a product that will be adjustable, and we want to know how adjustable to make it to accommodate most people in one size. To increase the adjustment range we will need to add material, adjustment mechanisms, and material strengthening structure. This will add cost, weight (mass), and complexity to the adjustment mechanism. Therefore, a smaller range provides a more effective product than a larger range and as we increase range, we degrade product quality. Therefore, we will refer to the additional range needed as the cost of accommodation.

To make an informed decision, we compared the increase in the percentage accommodated to the increase in the adjustment range needed for a given level of accommodation. We began with 80% accommodation, the 10th to the 90th percentile range, as the minimum percentage of our TP accommodated. Then, we examined 5th to 95th, 1st to 99th, and 0.1 to 99.9 percentage boundaries. This is shown in Table 3.9 and illustrated in Figure 3.23. The key measurement to be accommodated is normally distributed so we can use the mean and Std. Dev. to calculate percentiles. To simplify this example for calculation purposes we are using a mean = 100 cm and a Std. Dev. = 10 cm.

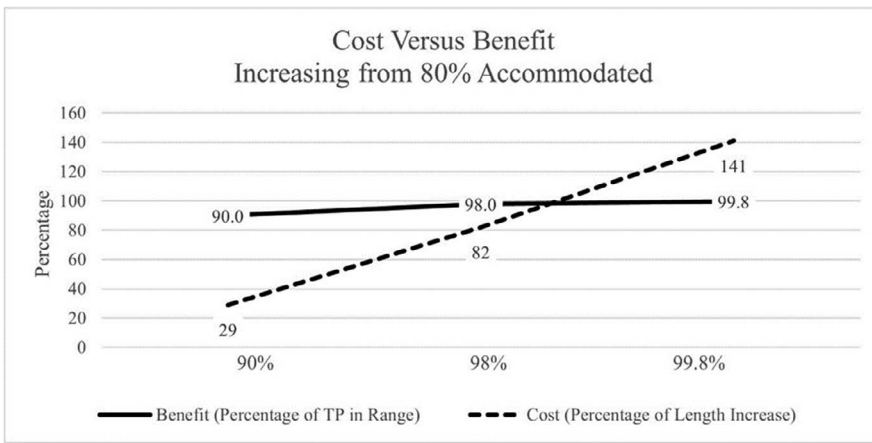


FIGURE 3.23 Cost versus benefit of increasing accommodation beyond 80%.

We see here that as we move away from the mean, the cost of accommodation increases. The adjustment range increases sharply but the accommodation range barely increases at all. When we increase our accommodation range to 90%, we must increase our adjustment range by 29% or from 25.6 cm of adjustment to 32.9 cm. If we increase our accommodation range to 98%, we must increase our adjustment range by 82%, from 25.6 to 46.5 cm. Finally, if we want to increase to 99.8%, just a 1.8% increase in accommodation from the 98% accommodation range, we must increase our adjustment range by 141%, to 61.8 cm – nearly 2.5 times the range of adjustment we needed for 80% accommodation.

If we chose the 1st to the 99th percentile (98% accommodated) we would only need a range of 46.5 cm, which is 59% less than the range needed for 99.8%. Is an additional 1.8% accommodated worth an additional 59% and 15.3 cm of adjustment range? Will the added range require us to have added pieces for telescoping the product, adding both weight and construction complexity and decrease the comfort and acceptability for the first 98% of the TP? It may be cheaper, easier, and result in a better quality product if we manufacture the product with two sizes of a component rather than trying to make one size fit everyone.

This is a simple 1D example. Most sizing issues are multi-dimensional. That means it is more difficult to assess the impact of using one accommodation range for our boundary versus another. This is why we stress doing fit testing to establish the range-of-fit in a single size before deciding on the number of sizes or the amount of adjustability needed. There is a quality and monetary cost associated with increasing the range of fit, and the cost is higher per percentage accommodated the farther from the center we go. Fit testing allows us to make informed decisions to achieve the best combination of quality, cost, and population accommodated.

In summary, we can use the boundary as a stopping point, but it is best if we start with the middle case and move out from there. This can be done as a random sample, a selection of distributed cases, or a set of grade rules. However, until we do design and sizing loop fit testing, we will not know what the range of fit is for each size so we will not know how many sizes we need or where to place them.

PCA Alternative for Key Variable Selection

Principal Component Analysis (PCA) is a multivariate method to help us understand common themes among a given set of variables and to determine the extent to which each variable in the dataset is associated with a common theme. By themes, we mean groupings of related variables and unrelated variables. Bivariate correlation helped us understand the relationships (or themes) between pairs of variables. PCA uses the correlations (or the covariances) to group the variables into related and unrelated components.

PCA can be useful for some applications if used properly and when its limitations are understood. For example, it can be a valuable tool for creating a parameterized 3D human model database, provided it is done in conjunction with standardized template models, Procrustes alignment of datasets, and uniform data point distributions. We will discuss this application in the section on software tools for prototyping.

However, PCA can be very misleading, difficult to understand, and easy to misuse, so we don't generally recommend its use for selecting key variables or cases.

We present it as an alternative method for key variable selection to demonstrate its proper use because many people use it incorrectly and have some misguided notions about its importance. Even when used properly it usually provides no additional benefit over the preferred method presented above for key variable or case selection.

PCA creates combination variables called components based on the multivariate correlations of all the input variables. PCA has the same number of components as the original variables, but it groups variables that are jointly correlated and separates them from others with which they are not correlated (independent).

The first components contain the variable groupings that explain the most variability. There is a common misperception that the first components are somehow better than the original variables. The argument is that since the first principal component (PC) is the variable combination that explains the most variability in the overall set of variables it is the most important and the best. As Brandon Walker (2019) explains, "… the first few components can explain almost all the variance in your data set. I've seen other data scientists mistakenly think that this means that … the first few components are the features that are most important. *The only way PCA is a valid method of feature selection is if the most important variables are the ones that happen to have the most variation in them. However, this is usually not true*".

In fact, if the most variability is due to measurement error or axis system and origin location inaccuracy, then error can become a large part of the first component. All subsequent components will be influenced by it since they must be independent of the first one. We saw one example of this issue in Chapter 2 when single coordinates for landmarks (x, y, or z) are treated as independent variables. When used separately these coordinates are subject to axis system and orientation definition error, the errors are correlated, and x, y, and z coordinates are dependent on each other, so they are not the independent variables PCA requires. After all, PCA is looking for dependencies (relationships), so if we put dependent variables in, we should get dependent variables out and this can hide the size and shape relationships and themes we are seeking. This renders the analysis useless, or worse. It can encourage people to use components as variables that are irrelevant to body size and shape, to the design and the fit.

This is particularly a problem for the head because a small axis system error has a large effect on the point locations. The origin and axis systems for the head can be in different locations each time a person is measured and there is no effective standard head orientation system as explained by Whitestone and Robinette (1997).

Also, Hudson et al. (2006) further note, "… since all of the dimensions put into the analysis are given equal weight, *accommodation based on the components will include some of the variability of the possibly less important dimensions at the expense of the more important ones*". In other words, the PCs can dilute the representation of the most important variables in favor of representation of less important ones.

PCA is particularly problematic if: (1) measuring errors for different measurements are correlated, (2) more of one type of variable is included than other types, (3) there are more measurements from one area of the body than another, (4) many unimportant variables are included, (5) some of the input variables are imprecise but others are very precise, and (6) the input variables are dependent. The garbage in, garbage out (GIGO) concept could have been created because of PCA. We may not know ahead of time that we have garbage going in and may not be able to discover that we have garbage coming

out. There is nothing to tell us what a component is other than a list of numbers. We can theorize about what the numbers represent, and we can make visual representations to try to understand what the components are, but we can never really know.

PCA can be used for key variables and case selection if done properly and if the purpose is to understand which of the independent variables we used as input into PCA are related and which are not. This knowledge can help us find key variable combinations from the input variables that control a lot of the size and shape variation.

Each component is a combination of the original input variables and there are an equal number of components as original variables. In other words, if we start with eight variables, we have eight components. The first component, called PC1, is the one that explains the most variability in the full set of variables. The second component, PC2, explains the next greatest amount of variability that is also independent of PC1, and so on. Each component is a combination of all input variables, so to measure someone to get their component value for any component we must measure them for all input variables. Since we are trying to reduce the number of variables we must measure, this is an undesirable property of the components themselves. So, it is often best to use PCA just to understand the relationships rather than using the components directly.

The contribution of each original variable to each component is called its loading on the component. These loadings are essentially the variables' correlation with the component. They can be interpreted to understand what variability the component represents, but the interpretations are just a matter of opinion and can be misleading. For example, the first component of a basic, unrotated PCA analysis is often referred to as overall size, but it is not actually overall size, and the unrotated PCA is also not very useful. For anthropometric measurement analysis when the purpose is to understand themes for key variables, the Varimax rotation is better. This rotation better separates the loadings of the variables, so each variable's greatest loading falls on different components. This makes the relationships between variable combinations easier to understand and use. With the Varimax rotation, PC1 is still the one that explains the most variability, and all components are still independent of each other.

To do a PCA analysis a professional statistics software package is recommended. We use SPSS® here, but most software packages have this capability. The steps to perform a PCA analysis in SPSS® are shown in Figure 3.24. It is found under the "Dimension Reduction" tab in the Analysis section. The procedure is called "Factor" which stands for Factor Analysis. This brings up a menu where we select the variables we want to include. The selection variable box underneath the variable selection box can be ignored. Under the rotation tab "Varimax" is selected. Under the scores tab "Save as Variables" and "Regression" are chosen to save the components in the data spreadsheet. This will allow analysis and plotting of the components as variables in the data set.

We performed a PCA analysis with Varimax rotation using SPSS® with the same head data used for selecting cases with raw data. The results are shown in Table 3.10. The table is divided into three parts. In part a. the loadings on the first three components for the first PCA analysis are shown. Part b. shows the loadings on the first three components for the PCA with the Varimax rotation. Part c. shows the proportion and percentage of the variance of all the variables that are explained by each component for both the original and the Varimax rotation.

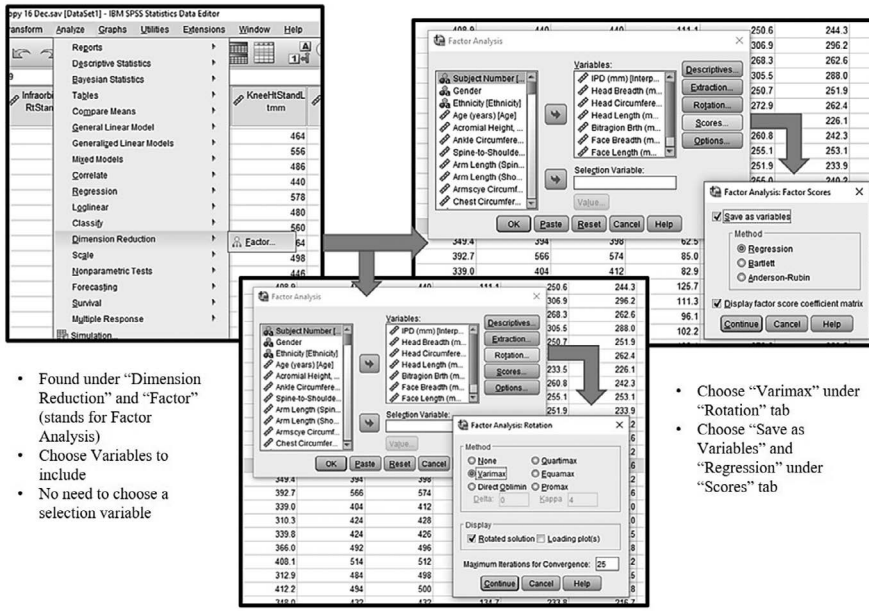


FIGURE 3.24 Performing PCA analysis in SPSS®.

Regardless of the rotation, the first component PC1, controls the greatest amount of variability in the input variables. The second component PC2, controls the next greatest amount that is independent (has zero correlation) with the first. The third controls the next greatest etc. After a certain number of components, the amount of variability controlled becomes unpredictable because all options are equally likely due to random chance. When the analysis reaches this point, it no longer matters what variable combination is chosen. In this example, this occurred after three components.

The first component from the original PCA will typically have a PC1 that has moderate to high loading for all variables. It is often referred to as "overall size", but that is an opinion about variability or error and must not be confused with a body size relevant to the wearable.

The original PCA does not help us reduce our variable set. It lumps everything together and treats all variables as if they are equally valuable. It is better to use a rotation. The Varimax rotation separates the input variables into different components while maintaining the overall variability controlled. In other words, each input variable will be a large part of only one component and will be a small part of all the others. At the same time, the overall variability represented in the meaningful components remains the same.

Part c also shows a statistic called an eigenvalue. The eigenvalue indicates what portion of the overall variance is explained by each component, expressed as the number of variables a component represents. In this instance, there are eight variables, so there are eight total components, and the total of the eigenvalues = 8.

TABLE 3.10
PCA Analysis of Eight Variables

Part a. Original PCA First Three Components

	Component		
	1	2	3
IPD (mm)	0.434	0.068	0.282
Head Breadth (mm)	0.741	–0.451	0.096
Head Circumference (mm)	0.803	0.224	–0.5
Head Length (mm)	0.467	0.546	–0.664
Bitragion Breadth (mm)	0.767	–0.444	0.089
Face Breadth (mm)	0.804	–0.416	0.155
Face Length (mm)	0.484	0.671	0.337
Sellion-Supramenton (mm)	0.308	0.717	0.525

Part b. PCA with Varimax Rotation

	Component		
	1	2	3
IPD (mm)	0.362	0.375	0.020
Head Breadth (mm)	0.868	0.004	0.091
Head Circumference (mm)	0.423	0.14	0.864
Head Length (mm)	–0.06	0.149	0.965
Bitragion Breadth (mm)	0.884	0.014	0.111
Face Breadth (mm)	0.91	0.087	0.091
Face Length (mm)	0.072	0.854	0.255
Sellion-Supramenton (mm)	–0.06	0.937	0.045

Part c. Eigenvalues and Variance

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.16	39.56	39.56	3.16	39.56	39.56	2.69	33.57	33.57
2	1.89	23.63	63.19	1.89	23.63	63.19	1.8	22.45	56.02
3	1.2	15	78.18	1.2	15	78.18	1.77	22.17	78.18
4	0.86	10.75	88.93						
5	0.37	4.6	93.53						
6	0.23	2.89	96.42						
7	0.21	2.61	99.03						
8	0.08	0.97	100						

Component one (PC1) represents the variance of 3.164 variables or 39.55 percent. When the eigenvalue is less than 1.0 it is considered unstable and it means the component represents less than one variable, therefore any component would have about the same result. In this instance, this happens after three variables and it means that only the first three are meaningful.

The difference between the original PCA and the Varimax rotation occurs within the first three components in this example, because these were the only components that had eigenvalues greater than 1. The cumulative percentage for both is the same, meaning they explain the same amount of variability together. The difference is in the loadings and the percentage of the variance explained by each component. The loadings are the correlations (R) between each variable and each component.

In our example, the Varimax rotation has more variables with strong loadings (>0.70) and the loadings are higher in magnitude with at least one variable with a loading ≥0.90 for each component. This makes the Varimax rotation best

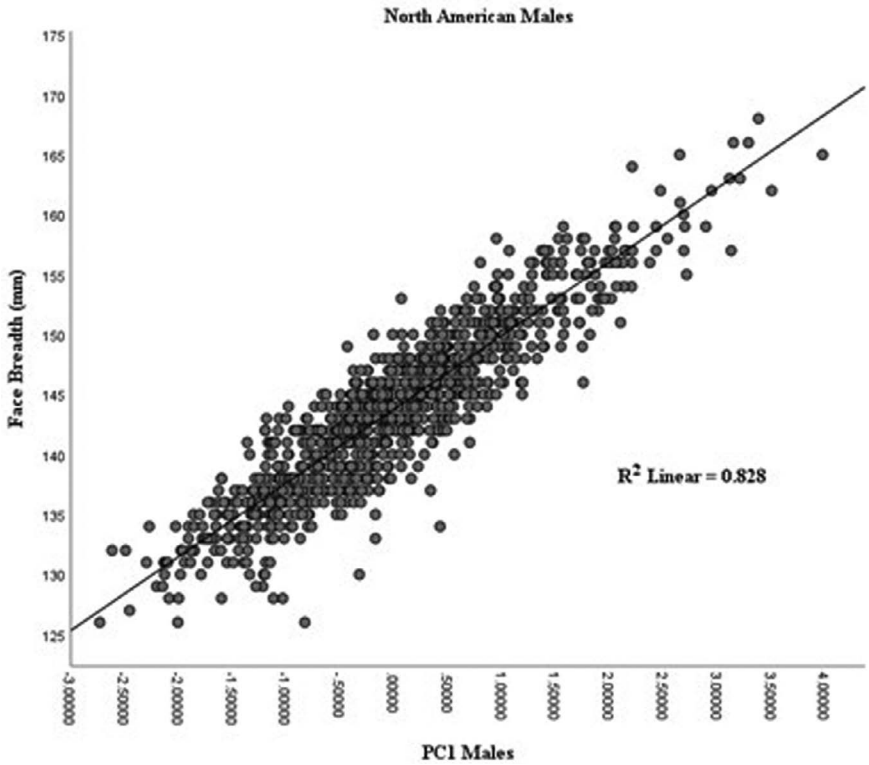


FIGURE 3.25 Bivariate scatterplot of PC1 in the Varimax Rotation by Face Breadth.

for down-selecting variables. PC1 in the Varimax rotation has high loadings for breadths and low loadings for everything else. So, we will refer to it as the breadth component. The loading for Face Breadth is the highest at 0.910. The bivariate chart of Face Breadth by PC1 is shown in Figure 3.25. The R of 0.910 give the R^2 of 0.828 or 82.8% of the component is explained by Face Breadth.

Face Breadth loads very weakly on PC2 and PC3, 0.087 and 0.091, respectively. This indicates it is nearly independent of them. Figure 3.26 is the bivariate plot of Face Breadth by PC2. We see that the points are roughly circular, the best-fit line is nearly flat, and the R^2 is nearly zero indicating the two variables are independent. Thus, PC2 explains a different component of variance than Face Breadth.

Since Face Breadth load is high on PC1 and is nearly independent of the other PCs Face Breadth is a good key variable. Note this was our first key variable in the bivariate analysis as well. This is not a coincidence. It has strong correlations with two of our eight variables, so as a single variable it explains the most variability in the eight-variable set, and it had weak correlations with everything else. This is how we selected it in the bivariate analysis and how PCA with the Varimax rotation works.

The other two components' results are also like our bivariate analysis choices, but they are in a different order. Sellion-Supramenton is the strongest loading variable

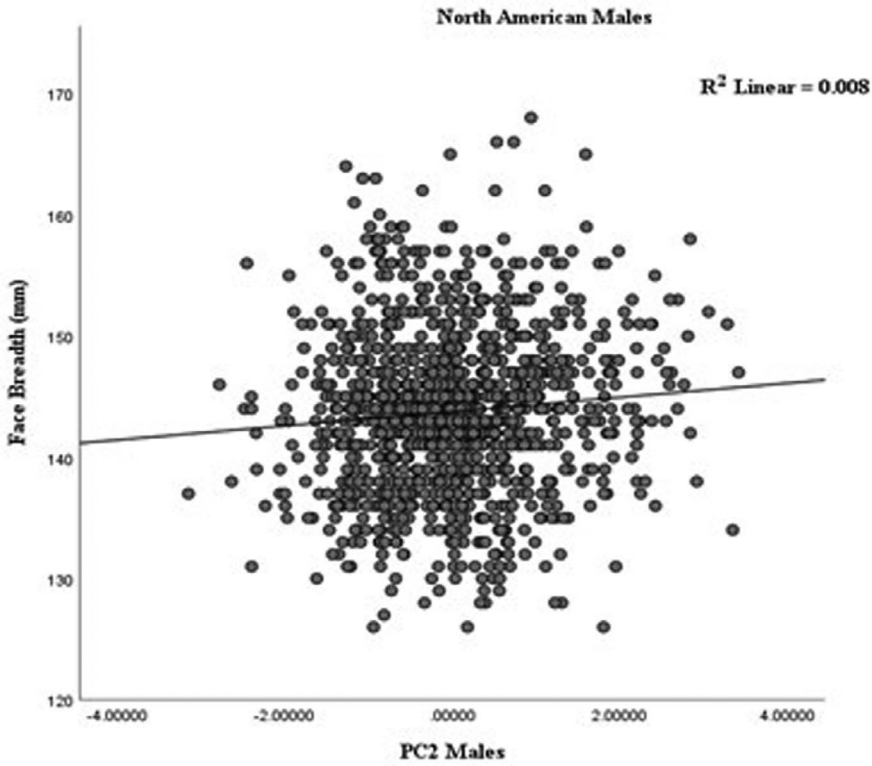


FIGURE 3.26 Bivariate scatterplot of PC2 in the Varimax rotation by Face Breadth.

for PC2 with a load (R) of 0.937. The other variable with a strong loading is Face Length. Both are measures of the length of the face and all other variables load weakly. We would interpret PC2 as a face length component. Figure 3.27 is the bivariate plot of Sellion-Supramenton by PC2. Sellion-Supramenton would be a good representative for PC2.

The strongest loading variable on PC3 is Head Length with a load of 0.965. Figure 3.28 is the bivariate plot of Head Length by PC3. Head Length controls 93% of the variance of PC3. The other strongly loading variable is Head Circumference. All other variables load weakly. PC3 can be interpreted as a head size component that is independent of the breadth component in PC1 and the face length component of PC2. This would make Head Length an excellent key variable.

One important variable, IPD, did not load strongly on any component. It also did not have a strong correlation with any of the other eight variables in the bivariate correlation analysis. However, IPD might be our most important variable if our head wearable has a visual display. IPD did not load highly on any PC because there were no other variables related to it in our eight-variable set. That does not mean it is not important. PCA analysis does not tell us what is important. It treats all input variables as equally important and ignores things that are not represented in the input variables. It also treats variability due to measuring error equally with variability

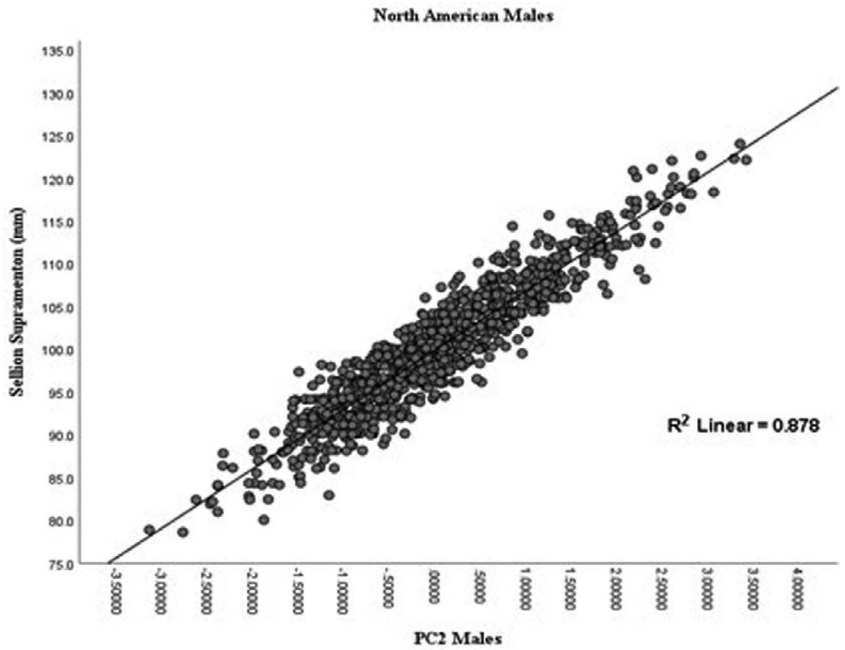


FIGURE 3.27 Bivariate scatterplot of PC2 in the Varimax rotation by Sellion-Supramenton.

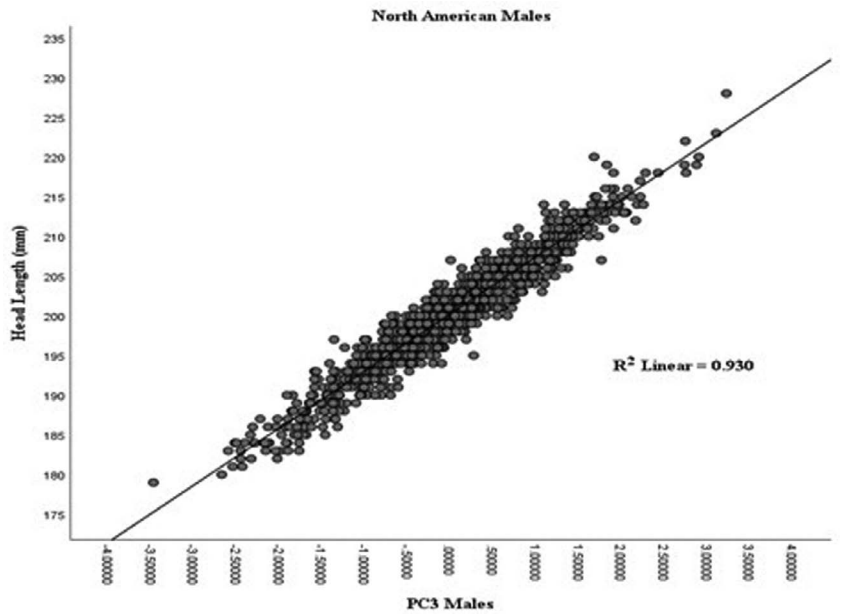


FIGURE 3.28 Bivariate scatterplot of PC3 in the Varimax rotation by Head Length.

due to actual differences in size. It does not know the source of the variability, so it is a mistake to think that the first PCs are more important than the others.

We review the PC components to understand which variables are related and which are not. This does not tell us which are more important than others. It is nice that the PCA separated the variables out for us, and tells us what measurements vary together, but we must still decide which aspects of variability are important. Just as with bivariate correlations, correlation is not the same as important.

Continuing with our discussion about down-selecting variables, we may not want to use PC2 or Sellion-Supramenton as a key variable. Let us look again at the percent of variance explained by PC2 versus PC3 in Table 3.9. It is 22.452 and 22.165, respectively. They are only slightly different. We had two length measurements of the face in our set of eight and they happened to be strongly correlated. That gave them preference to be the second component. That does not mean they are more important than the head measurements. If head variability is more important to our product than lengths of the face, we may want to use PC3 or its strongest loading variable as the second key variable.

In addition, if face length variability is not as important as IPD we should drop PC2 and its high loading variable Sellion-Supramenton from our key variables and select IPD instead. That would give us three key variables: Face Breadth, Head Length, and IPD. Or, as we did in the bivariate analysis, we can start with the two most important variables and decide if we need a third key variable later.

The components (PCs) from PCA are variables and can be used directly as key variables. There is debate about whether to use the highest loading variable or to use the PCs, themselves. We recommend using the highest loading variables rather than the components for several reasons.

First, using the PCs as our key variables is usually impractical. Using the PCs as the key variables makes it more difficult to find new cases because all input variables must be measured to calculate each PC. If we use the original variables as key variables, we only must measure 2 or 3.

The statistics software tool will typically do the calculations for you for the subjects used as input to the analysis. For additional subjects, such as when we look for a new case or recruit someone new, we need to know the equation for calculating the PC. PC equations are found in the component coefficient matrix, which is the list of numbers (coefficients) that must be multiplied by each input variable to calculate each subject's PC score. An example is shown in Table 3.11.

The full equation for calculating PC1 is:

PC1 = (0.134 * IPD) + (0.338 * Head Breadth) + (0.058 * Head Circumference) + (–0.150 * Head Length) + (0.342 * Bitragion Breadth) + (0.353 * Face Breadth) + (–0.025 * Face Length) + (–0.054 * Sellion-Supramenton)

Therefore, not only is a PC more difficult to measure but these measurements must be taken precisely as in the original study. This is not only cumbersome for finding new subjects to represent our case, but multiple measurements can introduce multiple measuring errors to the PC calculation. This is especially true if we are asking people to measure or estimate their own measurements.

TABLE 3.11
Component Coefficients

Component Coefficient Matrix

	Component		
	1	2	3
IPD (mm)	0.134	0.217	–0.100
Head Breadth (mm)	0.338	–0.036	–0.050
Head Circumference (mm)	0.058	–0.080	0.492
Head Length (mm)	–0.150	–0.080	0.618
Bitragion Breadth (mm)	0.342	–0.034	–0.040
Face Breadth (mm)	0.353	0.014	–0.070
Face Length (mm)	–0.025	0.477	0.007
Sellion-Supramenton (mm)	–0.054	0.570	–0.131

Second, using the PCs is also more difficult to communicate to consumers, designers, stakeholders, users, decision-makers, etc. about what a given PC represents. Decision-makers cannot decide what is important when they do not understand what they are. The original variables are things they understand.

Third, PCA does not tell us anything about measurements or variables we did not have available when we did the analysis, and if we use a different set of variables, we will get a different outcome. Sometimes we are missing variables in our sample that we might have liked to have had. If, for example, we had some additional eye width measurements related to IPD to use as input, such as Interocular Breadth and Biocular Breadth, perhaps IPD would have shown up as the first PC.

If instead we use the highest loading variables on what we consider to be the most important components, we arrive at good key variables that are less complicated, and easier to explain to people who are unfamiliar with advanced statistics.

In summary, PCA, if used correctly, can help us understand variable relationships and themes. However, PCA cannot tell us what is important, is easy to use incorrectly, and can be misleading so if we have a good understanding of the relationships, either from experience or from bivariate correlation analysis PCA is unnecessary and not very useful.

Selecting Cases with Aggregate Data

If all that is available about the Starting TP Sample is aggregate data (no individual data), then we have limited options for determining good key variable combinations. We cannot plot bivariate charts, calculate correlations, or do a PCA analysis. All we typically have are means, Std. Devs., and percentiles. Sometimes univariate frequency charts and the mode might be available, but this is rare. Aggregate data come in the form of a document with summary statistics rather than a spreadsheet of raw data. This means we cannot tailor the sample so our data may be biased, contain some subjects who are not in our TP and be missing others.

Finally, since we do not have individual data with aggregate data, the cases we select or evaluate are not people who were part of the sample. They must be new subjects we recruit. However, if we make some assumptions, there are still some helpful tools available to us to give us some confidence in our case selections.

With aggregate data, we have two steps for identifying or evaluating a base size case as with raw data: (1) rank order the variables and (2) select cases.

The first step for finding cases with aggregate data is to rank order the measurements available from most to least important for our product. The ordering helps us understand who is best for the most important [fit features].

First, we sort groupings of measurements, then we select one from each group to represent the group. This approximates selecting two or three key variables as we do with raw data. We are trying to have two or three highly ranked measurements that we believe are each related to a different group of variables. This should help us find more representative cases. This is done with the product in mind.

To illustrate, we used the raw data from our earlier headgear example. Working with the product developer we organized the variables we had available into ranked groups. The developer decided that the most important thing for this product is to get the product positioned well with respect to the eyes. Therefore, IPD was deemed our first most important measurement. It was the only measurement we had of the eyes, so it is the only measurement in the eye group. Next, the product is expected to be fitted to the top and back of the head, so the head measurements were deemed to be the next most important group. There will be speakers in our product that must be in the general area of the ears so the ear measurement, Bitragion Breadth was deemed next most important. This was the only ear measurement we had so it also represents the ear group.

While the product will have a mouthpiece, the location of it will not be critical because it does not have to be very close to the mouth. It will have a chin strap so face measurements are relevant, but it is expected that the strap will be adjustable so the variability of the face measurements will be accommodated with a single-size strap. We will need to know the range of the measurements but do not expect them to need sizes. As a result, we decided the face measurements are the least important.

Following this logic, we arrived at the following rank ordering of the measurement groups:

1. Eye Group: IPD
2. Head Group: Head Circumference, Head Breadth, Head Length
3. Ear Group: Bitragion Breadth
4. Face Group: Face Length, Bizygomatic Breadth (Face Breadth), Sellion-Supramenton

Next, we selected a measurement to represent the measurement groups that have more than one measurement. Since Bitragion Breadth, representing our third group, is also a type of head breadth, we selected Head Length to represent the head size group. It is less likely to be correlated with Bitragion Breadth than Head Circumference or Head Breadth, and more likely to be related to different variables than Bitragion Breadth. The first two or three variables become our key variables.

After the first three measurements, the ordering is less important, so we just move all the measurements we did not select for a group to the end. This gives us the following rank order:

1. IPD
2. Head Length
3. Bitragion Breadth
4. Face Length
5. Bizygomatic Breadth (Face Breadth)
6. Sellion-Supramenton
7. Head Circumference
8. Head Breadth

Because we do not have raw data, we do not have any individual's measurements and must find and measure new subjects to be our case(s). We first screen candidates for the first two or three measurements and if they fall too far from the middle for these two, we can eliminate them and move on to the next person. When we find one who is close to the middle, we select them as a candidate and take their other measurements. Then we examine all measurements and other characteristics, just as we did with raw data, to see if they are extreme outliers or unacceptable for other reasons.

We explained in the introduction that we are not likely to find someone who is within ±1 SD of the center for all eight measurements. If we start with the most important, then the cases we find should be near the center for at least the most important things.

When we measure new people, we must measure them in the same way and with the exact same tools as they were measured in the starting TP sample. If not, we may be comparing two different measurements. In our head measurement example, we see that we have the same tools for seven of the measurements, but IPD was taken from a scan, not with the tool we have available a pupillometer. We do not have the option of using a measurement taken from a scan so we cannot measure IPD in the same way. This may mean our measurement is not the same as that from the starting TP sample so we cannot compare it with any reliability. As a result, we must not give IPD a high ranking, or it could mislead us. We may have to evaluate the case's IPD and eyes some other way, perhaps visually or compared to other data. For this reason, we move IPD to the end of the measurement list. This gives us a new measurement ranking as follows:

1. Head Length
2. Bitragion Breadth
3. Face Length
4. Bizygomatic Breadth (Face Breadth)
5. Sellion-Supramenton
6. Head Circumference
7. Head Breadth
8. IPD

We again use a headwear example to illustrate, but this time we have a population that has equal percentages of males and females, so we want to consider cases from both subgroups. We recruited new subjects to find suitable candidates. We examine each candidate's measurements to see if they are in the center or are outliers.

Aggregate data usually includes the mean and Std. Dev., and these can be used to calculate standard scores for each measurement. The standard scores are also called z scores. The z distribution is the standard normal distribution and z scores are the measurement values in Std. Dev. units rather than the original units of measure. In other words, they are translations of a subject's measurements into new variables, the standard score variables, that tell us where each subject falls relative to everyone else in the sample. We examine the z scores to eliminate candidate cases that are outliers or abnormally small or large, or to confirm that our candidates are all reasonable. Then we examine other criteria to look for desirable attributes, such as symmetry or things we see that look odd but for which we have no measurements. Some of these are very subjective.

The z-score method assumes the measurements are normally distributed and most anthropometric measurements are well approximated by the standard normal distribution. If two measures have the same z score, then they are estimated to be the same distance away from their means in terms of each measurement's variability. The z score for the observed measurement is calculated as follows:

$$z = \frac{y - \mu}{\sigma}$$

where y is the observed measurement value, μ is the mean, and σ is the Std. Dev. z = 0 indicates the value is equal to the mean. z = 1 indicates the value is one Std. Dev. larger than the mean. z = –1 indicates the value is one Std. Dev. below the mean.

z scores can be translated into estimates of the percentage above and below the z score, the percentiles. Percentile values can be used instead of z scores, but we prefer z scores for three reasons: (1) we are particularly interested in distance from the center rather than the percentage of cases that are a given size or smaller, (2) percentile values estimates can be calculated from z scores if we want them, and (3) z scores provide smaller units of measure near the mean so it is easier to discern differences.

Univariate z scores are easy to calculate once we have the mean and Std. Dev. for a measurement. We use the standardize function in Excel™ to calculate the standardized variables. The standardize function uses the subject's value for the measurement, the mean for that variable, and the Std. Dev. to calculate z. The function as it appears in the cell entry box on the spreadsheet is, "=STANDARDIZE(B8,B2,B3)", where B8 is the row and column with the subject's measurement, B2 is the row and column with the mean, and B3 is the row and column with the Std. Dev.

To illustrate, we calculated the z scores for the four cases we viewed earlier in our headwear example, cases 490, 1088, 1M, and 3M, and graphed them by measurement number. This is shown in Figure 3.29. If we set our elimination criteria as someone who is outside the 25th to 75th percentile range, in other words, outside the middle 50%, it is the z score range from –0.67 to +0.67.

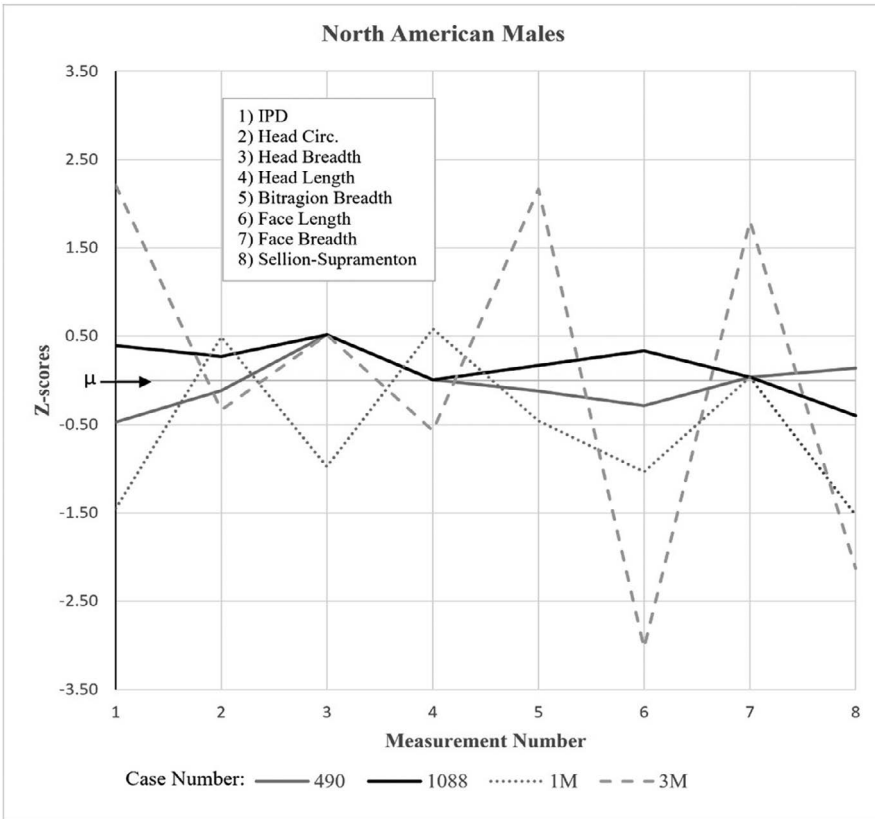


FIGURE 3.29 Z scores for eight variables for four cases.

As can be seen, the two subjects chosen based on our original three key variables when we had raw data, numbers 490 and 1088, are both close to the mean (μ) for every one of the eight measurements. Head Length and Face Breadth (numbers 4 and 7) were our first two key variables and they are right at the mean for these two. In addition, all eight variables are within ½ Std. Dev. of the mean for these two cases. The other two subjects fall outside our middle 50% range for at least four of the eight variables.

Subject 3M is extreme for five of the eight measurements and only barely within the 0.5 range for the other three. For IPD his z score is 2.3 which is near the 99th percentile. This indicates approximately 99% of the population is smaller. His Face Length and Sellion-Supramenton values are extreme in the other direction. They are very small. His Face Length z score is near –3.0 which indicates only approximately one in a thousand people would have a face that small.

We recruited subjects 1M and 3M simulating how some companies select fit models. They take measurements of candidates that they think are important and select some who are near the desired size for one or two of the measurements. They make no assumptions about key variables. Then they look at the subjects for other criteria

TABLE 3.12
Summary Statistics for Males and Females

Head & Face Statistics	Head Length (mm)	Bitragion Breadth (mm)	Face Length (mm)	Face Breadth (mm)	Sellion-Supramenton (mm)	Head Circumference (mm)	Head Breadth (mm)	Interpupillary Distance (IPD) (mm)
Male Mean	199.93	149.65	121.29	142.72	99.73	577.07	154.54	67.73
Male Std. Dev.	10.37	8.01	8.02	7.36	8.34	18.06	6.69	6.01
Female Mean	188.53	140.05	111.67	133.29	91.67	551.98	146.22	65.72
Female Std. Dev.	7.1	7.43	7.1	6.75	6.74	18.1	5.32	6.03

such as appearance, experience, symmetry, etc., and choose one. They are often unsure of their choices but do not know what else to do. In this example, each was chosen because they had at least one measurement that was near the mean. We chose 1M because he had a Face Breadth near the mean. We chose 3M because he had a Head Breadth near the mean.

It is important to use the ranges for each gender separately. If we combine them the center might be an area with few people of either gender. Since we cannot view the frequency distributions with only aggregate data we cannot see or measure where the concentrations of subjects are located and the mode is not typically reported, so we must use the center of each separate subgroup that is estimated by the mean or median (50th percentile).

The means and Std. Devs. for males and females from our starting TP aggregate data are shown in Table 3.12. When screening subjects we found two, one male and one female, who were within the 50% range for our first two variables and we labeled them 1M (the male) and 2F (the female). We also found two who were outside this range that we include for comparison purposes. They are labeled 3M (the male) and 4F (the female). (Note that 1M and 3M are the same cases we used earlier.)

We took all eight measurements for these four subjects and the measurements are shown in Table 3.13. We calculated and plotted their z scores and percentile values for each measurement using Excel™ (see Figure 3.30).

The female subjects' z scores and percentiles are from the female sample statistics, and the male values are from the male sample statistics. As a result, while a male's measurement may be larger than a female's, the percentile value and z score may be smaller. For example, subject 3M's Head Length is larger than subject 2F's Head Length, but his percentile value is 28 while her's is 58.

While the z-score plot and the percentile plot are directly related they each give a different perspective. The z score gives a better sense of closeness to the center, and this is what interests us. In the z-score plot, the mean is approximately at the zero point and a negative number indicates the subject is smaller than the mean while a positive number indicates the subject is larger. In the percentile plot, the mean is

TABLE 3.13
Candidate Cases' Measurements

Subj. No.	Head Length (mm)	Bitragion Breadth (mm)	Face Length (mm)	Face Breadth (mm)	Sellion-Supramenton (mm)	Head Circumference (mm)	Head Breadth (mm)	Interpupillary Distance (IPD) (mm)
1M	206	146	113	143	87	586	148	59
2F	190	140	109	132	95	555	146	68
3M	194	167	97	156	82	571	158	81
4F	202	148	108	142	91	567	154	83

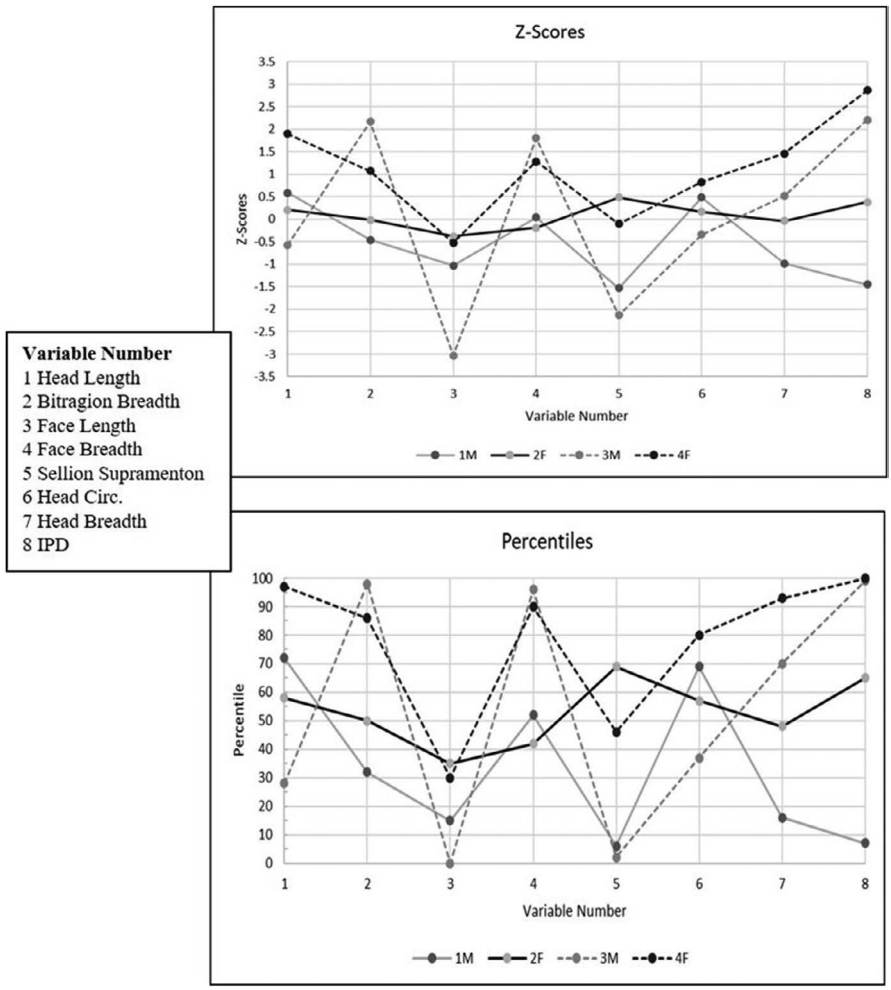


FIGURE 3.30 Z scores and percentiles from candidate cases.

approximately at the 50 percentage point and we must add and subtract from 50 to understand distance from the center.

The two subjects who were within the z score range from –0.67 to +0.67 for our two highest-ranked variables are 1M and 2F. The two who did not meet this screening criteria are 3M and 4F.

If we compare the two females, we see that the one who met our first two variables' screening criteria is closer to the middle for every measurement except number 5, Sellion-Supramenton. 2F is within the –0.67 to +0.67 range for all eight variables, while 4F is near the 90th percentile for five of the eight.

If we compare the two males, we see that 1M is closer to the middle for the first five of our eight variables and he is within the –0.67 to +0.67 range for four of the eight variables.

Comparing our two cases who were within our screening criteria, we see that 2F was much closer to the middle for our two screening variables. 1M was near the 0.5 z score for Head Length and –0.5 for Bitragion Breadth. As a result, 2F is a much better representative for the female center than 1M is for the male center. If our population was predominantly male, we might choose instead to start with the male case. This male case is borderline acceptable so we might consider continuing our search and narrowing our screening criteria to within –0.5 to +0.5 for the first two variables and try to find a better one.

When we have a population with equal percentages of males and females, we want to know if we can accommodate both in one size range or if separate gender size ranges are needed. Therefore, we want to start with a case from the middle of each group, perform a fit test on a prototype for one, and see if its size range accommodates the other. In this example, the female case seems to be a better representative for her gender than the male for his, so we examined her further.

Before deciding on a case, such as our candidate subject 2F, it is important to look at them either in person or use photographs or 3D scans to see if there are any oddities that aren't apparent in the measurements. In this instance, since they are new subjects, we looked at them in person, and we took 3D scans of 2F. We examined the 3D scan using an in-house software tool. Three views of subject 2F are shown in Figure 3.31.

Subject 2F appears to be a good candidate for a case. We can see that she does not have any unusual lumps or crooked features and she is very symmetric, both in the face and across the ears and head.

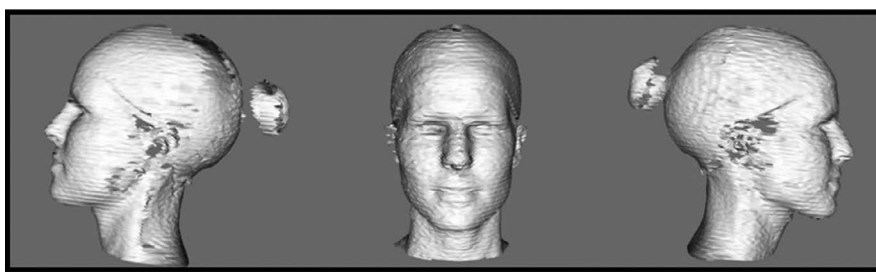


FIGURE 3.31 Screenshot of 3D scan of subject 2F.

USING CASES TO CREATE MOCK-UPS AND PROTOTYPES

Originally tailors, dressmakers, shoemakers, and blacksmiths made each wearable item to fit an individual person (a case), a process known as bespoke, haute couture or, more recently, made-to-measure, or individualized fit. Even medieval armor had to be fitted as explained in the Forge of Svan blog (Kuznets, 2018):

> After all steel elements of the future armor were shaped, forged, and hardened, the next – and most difficult – step arrived: gathering all the parts together and adjusting them to fit. This was a very important process, since completed armor should have no gaps, should be comfortable to wear, unrestrictive, and articulated. And the most essential for the knight of the Middle Ages – his armor should protect him as much as possible during the numerous medieval wars.

This custom process with fit testing is still a good way to make wearables for the first individual case, although we have some additional tools to help, such as CAD, block development, 3D scanning, parameterized 3D databases, and 3D printing. However, it is important to note that a particular wearable is usually a mixture of fitted and non-fitted sections, so the customization process is not as simple as copying the body or body segment. For example, for footwear, it is necessary to create or use a form called a last for the base size case which has areas that match the shape of the foot, such as the heel and the width across the ball of the foot, but other areas that have extra room for comfort and movement. The last is also directly used to form and produce the footwear, so it also must be constructed in such a way that it can be removed from the shoe or boot after it is completed. Some examples of shoe lasts are shown in Figure 3.32.

FIGURE 3.32 Wooden shoe lasts.

For clothing the fitted parts highlight or reveal body shape, while the non-fitted sections camouflage and conceal certain body characteristics. For example, empire waist dresses are fitted at the bust but loose below it to disguise the waist and hips, whereas A-line dresses emphasize and minimize the waist while hiding or disguising hips and thighs. With men's suits, the shoulders might be made to look broader with padding. The combination of fitted and non-fitted sections along with other constraints must be balanced during design and development. When the balance of textile, functionality, and the look combine as intended on the body, the result is a successful garment.

The desired fit appearance and intended functionality of the product are referred to as *fit intentions*. When fit intentions are expressed in a 3D form it is called the *form factor*. Computer models of the product are often used to capture fit intentions in a digital form factor. Some shoe or boot lasts can be considered physical form factors when they reflect fit intentions like the shape with a particular heel height. Hat forms, called blocks, often contain the shape of the intended product as well as the head circumference of the wearer. These would be considered form factors as well.

When the form factor is created in wearable physical form it is called a mock-up of the product. This might be a model of the product that has been carved in wood, 3D printed, or constructed from 2D fabric pieces and sewn together. The mock-up is something that can be worn and fit assessed so we can get our first feedback on the form factor in a design loop pilot test. This will guide the design process and help us alleviate issues with the design early when it is still inexpensive to do so.

The first mock-ups are usually physical versions of the product that are not fully functional and/or not made from the intended materials. A toile version of a garment or portion of a garment or a 3D printed shape for a mask are two examples. Mock-ups are inexpensive examples used to ensure things will fit together or that a shape or form is a reasonable size and attractive. Evaluating mock-ups on live subjects is a quick way to assess the basic shape or the components (sometimes called the form factor), before spending money on materials. This can reduce waste.

When we progress to fully functional products made from the intended materials, we refer to them as prototypes. Some aspects of fit cannot be assessed until we have a prototype, so it is important to iterate testing with prototypes as well as mock-ups on real people with design and sizing loop fit tests.

Until we can produce effective 3D fabrics, we must use 2D patterns for fabric products. For these items, fit intentions are captured in a 2D pattern called the block or sloper. This is the basic pattern upon which styles are based. An example is shown in Figure 3.33 (Veitch & Davis, 2009). The block is a mock-up that is typically made from a material that is easy to use and inexpensive, such as muslin rather than the final product material.

There are many approaches to getting started and the best one will depend on the product and the skill set of the producer. We can start with a case and build the mock-up on it, or we can start with a mock-up. Either way, we need to arrive at a

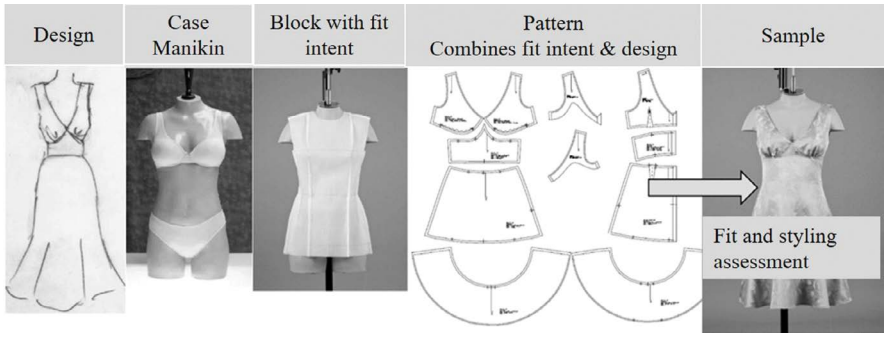


FIGURE 3.33 Development of the first apparel prototype (sample).

mock-up that we will be able to test on cases, including live subjects, in the design loop. If we use a case to build the first mock-up we might:

- Use a live person to make a mold
- Build/purchase a physical manikin representing the case and mock-up on it
- Build a 3D model of the case in software and build a mock-up on it

If we start with a mock-up we might:

- Sculpt, carve, model, or sew a mock-up
- Create a CAD software form factor, then print, cut, and sew a fabric pattern to create a mock-up
- Create a CAD software form factor and 3D print or mill a wearable mock-up
- Use/modify an existing product

We might also use combinations of these methods as we progress through the design process. Regardless of our approach, we will do multiple iterations of assessment and adjustment on live cases, which we refer to as pilot tests, stand-alone trade studies prototype fit tests in our design loop.

It is also important to keep track of the test results and any changes we make as we develop the product. There should be a procedure in place to make a digital record of the product form factors, any changes made, and the rationale for the changes to avoid repeating mistakes and to maintain the fit standard. CAD tools can be helpful because it is possible to keep a digital copy of every form factor. These should be included in the database and data management system to establish and maintain a sustainable fit standard for future products. If a CAD system is not used, some method should be in place to keep a physical or digital record of the form factors.

When we use cases to build our mock-ups and prototypes, we often use both live people and 3D copies of our cases, such as manikins or digital human models (DHMs). For ease of discussion, we define *manikin* as a 3D representation of the human body or its body segments. We will refer to all the different types of 3D copies as manikins, either *physical manikins* or *digital manikins.*

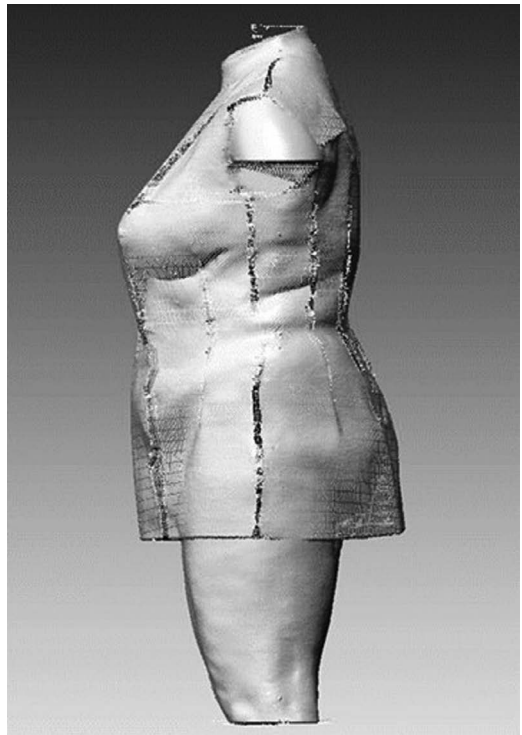


FIGURE 3.34 Superimposed garment on lifelike manikin.

Live people and manikins each have different advantages and disadvantages. For example, it is often easier and more practical to create the block or form factor to fit a physical manikin representing the base size case. Physical manikins can have pins inserted or be altered by adding clay or plaster.

Manikins remain static and unchanging, and this makes them easier to fit and easier to standardize. The manikin never gets tired or changes posture. This can make it easier to visualize the impact of product changes by superimposing 3D scans of the manikin with and without the product such as in Figure 3.34.

Also, modern rapid prototyping technologies are making it easier than ever to produce quick mock-ups for rapid trade studies. For example, we can now create digital form factors in CAD, 3D print the parts for a mock-up, assemble it, and try it out on ten people all in one day. The beauty of CAD is that once we try it and decide it needs to be changed, we can make the change, try a new version on the same group of people rapidly and keep a record of all the versions and test results. If we have a digital manikin to use in the software tool, we can visualize the product on our case and visualize the changes before we make them. In this way, we have a record of our fit model and all the form factors that worked and did not work for that model and the Concept-of-Fit used for the testing. This documents our fit standard and helps to create a fit database that we can both sustain and learn from.

Finally, one of the most important uses of a manikin is the ability to have multiple copies so they can be used as a communication tool for fit in multiple locations, say head office and the production facility. This includes sharing digital manikins over the internet for team design with remote team members.

On the other hand, manikins are never uncomfortable and can do the impossible. Manikins do not have opinions, and it is better to find out opinions about the design concept early on rather than have it fail to sell later because the users don't like something about it. For example, a head-mounted display company created a prototype that fit the CAD model well and the display performed well. However, the wearers complained that they felt like they were in a tunnel and said they would never purchase something like that because it was too claustrophobic. In other words, covering the sides of the product made the display better but made the product unsellable. This was something the CAD model couldn't tell us. Because we tested a prototype on live people, we were able to find the problem and fix it early in the design process while it was still an inexpensive fix.

Also, manikins will wear our design concepts in ways that people would never wear them and no matter how hard we try to force people to wear the product the way it was designed to be worn, often it is simply impossible. Manikins cannot reliably assess functional requirements, such as the ability to do tasks in the wearable. For example, although apparel fit is often tested in a standing position with the feet close together other postures may be important depending on the activities the wearers will undertake. Paramedics must deal with situations where the patient might be on the ground or in a difficult-to-access location and the paramedic might need to squat, kneel, bend and lift. Garments for paramedics should be tested on a live fit model doing these kinds of motions and activities.

A digital manikin can wear a product that is a perfect match to the body and look good but be misleading. It can be impossible to know how close to the body to make the product without using a live person. Skintight might be too tight to allow for movement or comfort, but for compression garments, tighter than skintight is needed. To estimate how close to the body the product measurements should be, we need testing with live subjects. For example, skintight gloves might be good or bad depending on the materials and the activities. We might need to feel what we are gripping, so tightness is good, but we might also need to have the dexterity to perform our tasks and tightness would be bad. It might be good to have tightness on the fingertips, but looseness over the back of the hand. The only way to know is to test on live people. Gloves can be tested using dexterity tests, such as those used for testing Chemical Defense gloves (Ervin & Robinette, 1987; Robinette et al., 1986). In these studies, different types of gloves were compared, but the same tests can be used to evaluate fit.

Some software tools permit us to create our mock-ups and prototypes and even do some preliminary theoretical fit evaluations. There continue to be new advances in human modeling for CAD that can help make the first mock-up or prototypes better. For example, Vital Mechanics Research has created a software tool VitalFit that has a human model with soft tissue properties, (https://www.vitalmechanics.com/?page_id=242). This enables some ability to estimate the amount of tissue compression or tightness that might be caused by a garment. The estimates are not very accurate

currently due to avatar limitations, lack of real fit data and validation of the fit predictions. However, the tool has two benefits: (1) it permits a quick and inexpensive first comparison of design concepts and (2) it permits the tracking and recording of fit assessments on live subjects. In other words, if used in conjunction with fit testing on live subjects it can help the team to visualize and track design issues. The tool can be imported into the Browzwear™ CAD tool.

CAD tools for prototyping products such as eyewear, helmets, and body armor such as AutoCAD® can be useful for the same reasons. The fit estimates are not accurate enough for the final product, but they can help with the first approximation of the product and are useful for tracking and visualizing the live subject fit test results and scans.

It is important to select or develop manikins that will have the same fit as the live model. Achieving this can require manikin evaluation since the perfect manikin will not always have the same measurements as the person. In addition to requiring close fit in some areas and loose in others, the manikins do not have the same physical properties as human beings. They do not have the same soft tissue, bone structure, bendability, etc.

Good manikins can be as complicated to obtain and use as the wearable products themselves. If using manikins, it is a good idea to document all manikins used, tested, or adapted. This is important for the fit standard and for future manikins we might need for new TPs or fit model changes. We provide some guidance on how to create, evaluate, and select manikins in the next two sections, one on physical manikins and one on digital manikins.

Physical Manikins

A physical manikin is idealized and typically has a fixed pose or set of poses. There will be some areas where its body measurements differ from the live person for good reason. For example, a real person can put their arm at their side and there is a section of body contact between the arm and the body extending from the armpit toward the ground often of at least 10 cm in depth. However, a manikin cannot be built with zero gap between arm manikin and torso, because the armhole of the garment needs to slide into the armpit of the, and there must be a gap to allow this. So, either the side of the body needs to be flattened and shaved down, the arm circumference of the manikin is reduced, or the shoulder widened to move the arm away from the body and create the desired gap. Otherwise, it can make it difficult to put the correct fitting size on the manikin.

If the waist of a manikin is the same size as the waist of the garment it may not be possible to fasten the garment at the waist unless the manikin's waist compresses like the human waist. A head manikin with rigid ears, such as that in Figure 3.35, may make it impossible to don a helmet.

Therefore, even though the product would fit the live model perfectly it may not fit the manikin. Sometimes these errors are compensated for when the garment is made by building in ease, such as on the sleeve circumference, or by shaving down parts of the manikin to allow for donning. But sometimes the ease just makes it worse or creates a different problem with the sleeves just ending up too tight or the waist too

FIGURE 3.35 Size large helmet head form made from wood.

large, or the helmet sitting in the wrong position. This is where a live fit model is essential. He or she will be able to say if the wearable is hard to don or doff, point out movement restrictions and other comfort constraints.

There are also debates about how realistic the manikins need to be. For example, there is a clear argument from industry that a symmetrical manikin is easier to work with and less confusing when dealing with remote supplier chains. However, simplifying the shape by making it symmetric may also remove a layer of realism. The goal is to have a simplified manikin that is close enough to the real fit model such that if it fits the manikin, it will also fit the real person. There is always a risk that it will not be close enough. This risk is mitigated when we test fit on the fit model and then fit test on others in the fit-testing phase.

Commercial apparel manikins have improved substantially since the 1990s, and there are companies that produce manikins to represent a specific base size case. However, they still have some limitations. For example, the commercial manikin in Figure 3.36 was built to represent a list of key measurements such as Neck Base Circumference, Bust Circumference, Waist Circumference, and Hip Circumference, which are indicated with markings placed on the manikin. However, as is true for many commercial manikins these measurements are stacked in the center vertically, but on real people, they are offset, in other words, some have more of the measurement in the back and others have more in the front. This means the shape of this manikin does not match any actual person.

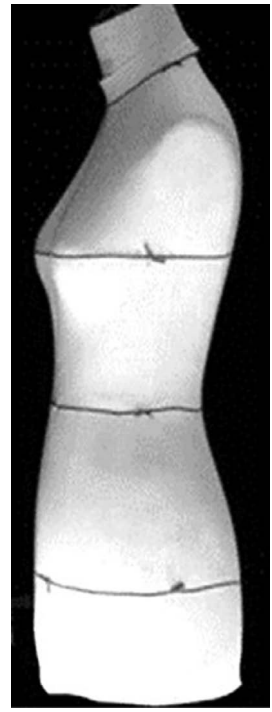


FIGURE 3.36 Commercial apparel manikin.

Figure 3.37 compares the commercial manikin to the fit model it is intended to represent. Our fit model's waist and hips fall more toward the back and sides than the manikin's hips. In addition, the fit model's waist in her clothes falls at an angle front to back, while the manikin's waist is flat or horizontal. The manikin has no buttocks or crotch shaping. These kinds of difference impact apparel fit and need to be understood to avoid fitting the manikin well but not fitting real people. The manikin might be the right size but the wrong fit or it might be good enough for some garments and not others.

It is important to consider whether you want a copy of the person or if you want a shape that will fit the person. Sometimes the shape that will fit the person incorporates some of the fit intentions. For example, a shoe last is a type of manikin that is a shape for making a shoe that will fit the foot, rather than a manikin that matches the shape of the foot. The last can have a different shape depending on the intended heel height for the shoe because the heel height affects the foot pose and changes the shape of the foot.

Hat manikins are called hat blocks in the industry. Some hat blocks are simple shapes with the desired circumference, length, and breadth, and a rounded top. Some hat blocks contain the intended hat shape along with the desired head measurements. The latter makes some assumptions about the location of the hat on the head and the intended fit of the manikin may not match the fit of the case or person. In this

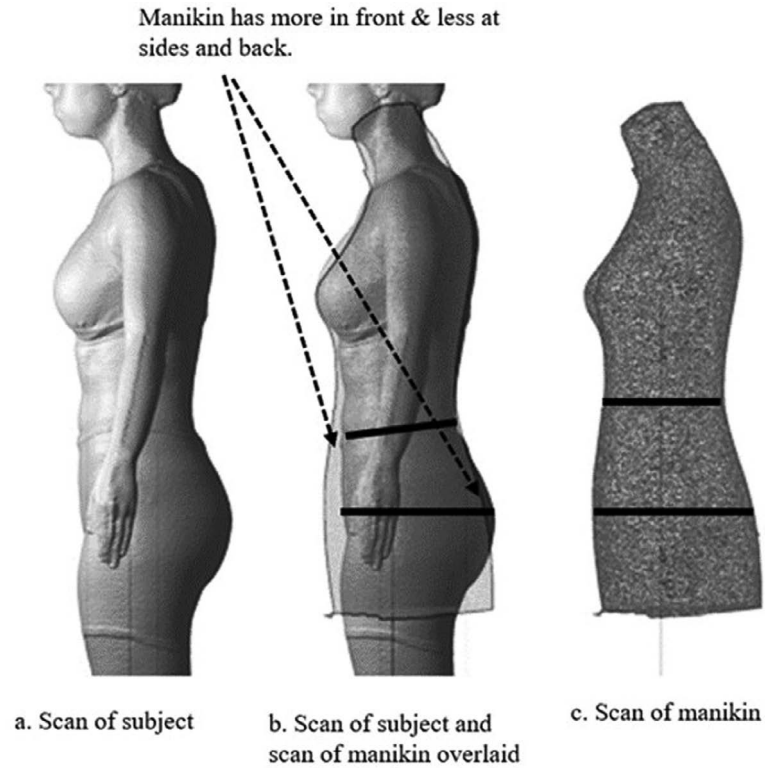


FIGURE 3.37 Commercial manikin compared to actual individual.

example, unlike the foot last, the additional fit intention can make the fit on the real person different than the manikin.

Things to consider when purchasing or producing a manikin are:

- Do you want to match the person or a shape that will fit the person?
- How idealized is acceptable?
- Are those locations where the manikin is idealized or simplified clearly shown in the manikin?
- How well do both the size and contours match the case or shape that you want?
- What areas are acceptable as not matching and what areas should match?
- Is the manikin rigid and flexible in the right places?
- Does it have removable or moveable body parts for donning?
- Does it have clearances for getting the wearable on the manikin?
- Is the manikin symmetric and do we want it to be?

Some of these issues are difficult to know without testing. Therefore, before deciding on a manikin to be used as part of the sustainable fit standard it will be

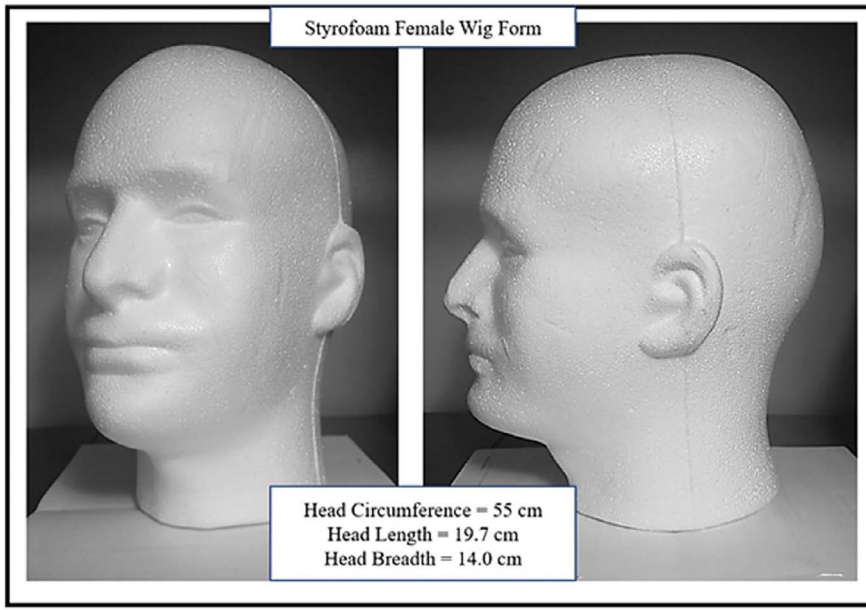


FIGURE 3.38 Styrofoam wig form.

important to try it out and test it. Even if it is not perfect and we still want to use it, the testing can reveal its limitations that can be accounted for when it is used.

Retail store visual display manikins should be avoided. The primary role of a visual display manikin is to look attractive while displaying clothes. It is not designed to be realistic. While suitable when used for display, visual display manikins have received criticism over the years for promoting an unrealistic body image. These are stylized with exaggerated features to make the product look pleasing. This means they will not have the proportions or shapes of any real person and will rarely be good enough for the creation of a block or mock-up. Typical female display manikins will be exaggerated to have abnormally long legs and small waists.

Sometimes they are proportioned to be easy to display products rather than to reflect real human proportions. For example, the head form shown in Figure 3.38 is a form for displaying wigs. Its Head Circumference of 55 cm is approximately average for American females and 10th percentile for males. So, it is too small for males for design purposes, but it would be easy to get a product on and off this small size head. In addition, even though it is about average for Head Circumference for women, the Head Length is at the 90th percentile for females and the width at the 10th percentile. This shape of the cranium is extra-long and thin. This gives it a more pleasing narrow face for display and makes it easy to get wigs on and off. The facial features are strictly an artist's concept so the placement of the features cannot be relied upon. If Head Circumference is the only important measurement, and the head shape is not a factor, this form might be acceptable for a first mock-up for women. But if Head Circumference is the only important measurement, why do we need a 3D manikin? For most products, these kinds of manikins are only good for display not for design.

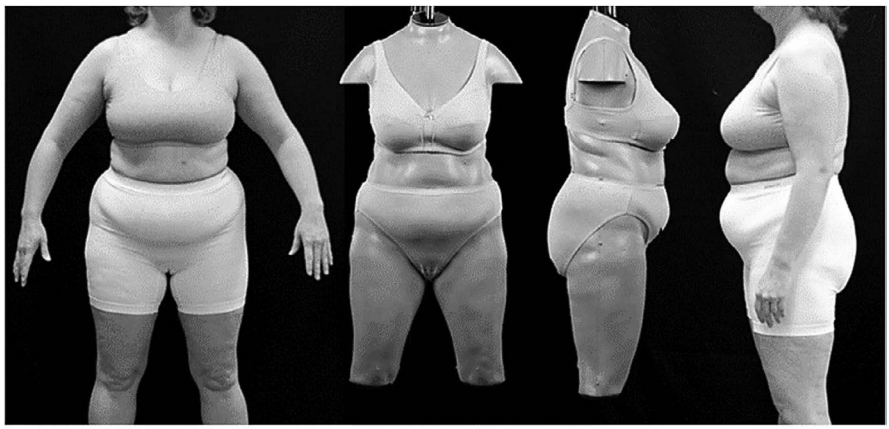


FIGURE 3.39 Example of subject with matching lifelike manikin.

There are companies that produce custom manikins. For example, SHARP Dummies® custom biofidelic manikins, such as the one shown in Figure 3.39, have a soft compressible layer that allows the body to give in a way similar to an actual person's body. Not all custom manikins have shape accuracy as good as these, so be careful to verify that the shape is accurate enough in the important places.

A specially developed manikin might also be used to represent the product and its fit intentions, rather than the body, thereby acting as a stepping-stone between a real body and the idealized garment shape. A good example of this is a shoe last, which although it has foot measurements built-in, is not the shape of the foot. Instead, the last provides the base shape of the finished shoe, which includes a predetermined heel height. In some instances, a manikin or form might be too detailed to allow manufacturing for a reasonable cost. For example, Kennedy et al. (1962) created more lifelike hand forms for dipping to produce protective gloves. During the process there was concern that the hand forms were too detailed, with sharp edges, strong finger curvatures, and fingers very close to each other, to permit the dipping process to work effectively and permit the gloves to be removed from the forms after curing. They tested the process to resolve any manufacturing issues.

In the past, manikins were made using artist tools such as plaster casting, or an artistic interpretation of a list of measurements made into a 3D clay sculpture of the body segment such as clay face forms used for creating oxygen masks (Alexander et al., 1979; Seeler, 1961). These were produced before the advent of 3D scanners, so they required an artistic interpretation of 1D measurements. Percentile values were often used instead of cases, so they often looked a bit like Frankenstein instead of a real person and their effectiveness was questionable.

Today, we have 3D human scanning technology which can be used to produce both physical and digital human models. Physical manikins can be made from properly processed 3D scans using 3D printing or computer numerical controlled (CNC)

milling. The process requires data processing to create a digital model first and instructions on the pathways for the printer or milling machine, but then can be used either to produce a manikin or to produce a mold to make the manikin.

Copies made with 3D printing can be cheaply and rapidly produced, which can make them ideal for quick prototyping. However, they are more limited in their use of materials and in the size of products created than by CNC milling. Some of the reliably accurate materials are very brittle and can break easily.

Digital Manikins

Digital manikins usually represent only the human body or body segment and do not incorporate the form factor or fit intentions into the manikin. The form factor and fit intentions are treated as separate digital entities. Therefore, digital manikins should be digital copies of the case.

Digital manikins of individual cases can be used, by a skilled and trained technician, to make digital representations of the product in CAD. These product models can then be prototyped with tools such as 3D printing or CNC milling. An example is shown in Figure 3.40. In this example, the subject's head was scanned and transformed into a CAD model. The subject was also scanned in the prototype of the noise protection headset, to locate the position of the earcup with respect to the head. The head surface under the earcup in the CAD model was used to create an individualized digital earcup seal that was produced using 3D printing in a rigid plastic.

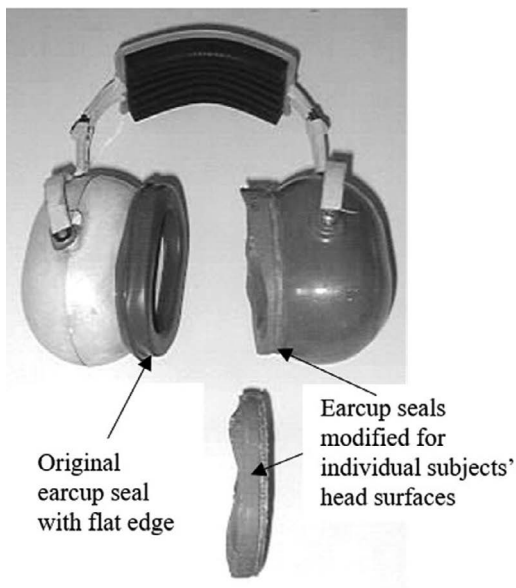


FIGURE 3.40 Combining head surface with product using CAD and 3D printing.

The 3D-printed product model was effective for testing both comfort and noise attenuation, even though it was rigid, not soft like the cushioned seal.

Digital manikins have historically been called DHMs. Not all DHMs have the same quality or usefulness for wearables. Some of the first DHMs, such as Articulated Total Body (ATB) and COMBIMAN (COMputerized Biomechanical MAN-Model), were nothing more than 3D figures made up of 3D cubes, rectangles, and ellipsoids (Evans, 1976; Leetch & Bowman, 1983). These were put together to estimate the biomechanics of the body for things such as crash research and cockpit layout and were estimated from 1D anthropometric measurements and mass distribution properties from cadavers.

Commercial DHMs with a bit more human-looking features soon followed (e.g., Jack, Ramsis, Santos). They have been used in commercial CAD systems to design and assess human-machine interaction and prevent musculoskeletal risks due to bad postures, overloads or problems to reach an important part of the machine environment (Duffy, 2007, 2008; Scataglini & Paul, 2019; Tian & Duffy, 2011).

While these models have proven very useful for workspace layout and visualization, they have many limitations that render them ineffective for wearable design. Most importantly, these models have a simplified body shape and posture without anatomical biofidelity. Figure 3.41 illustrates the difference between a commercial DHM (at left) developed for ergonomic assessment and a scan of an actual person of the same stature and weight (at right). For wearables that must follow the body contours, shape is critical and as a result, the shapes used in commercial DHMs today will not suffice. Fortunately, there are now better ways for representing the human body for wearables by developing DHMs using 3D scanning.

3D human body scanning began to be developed in the 1980s (Robinette, 1986) with some of the first small head scanners. From the second half of the 1990s

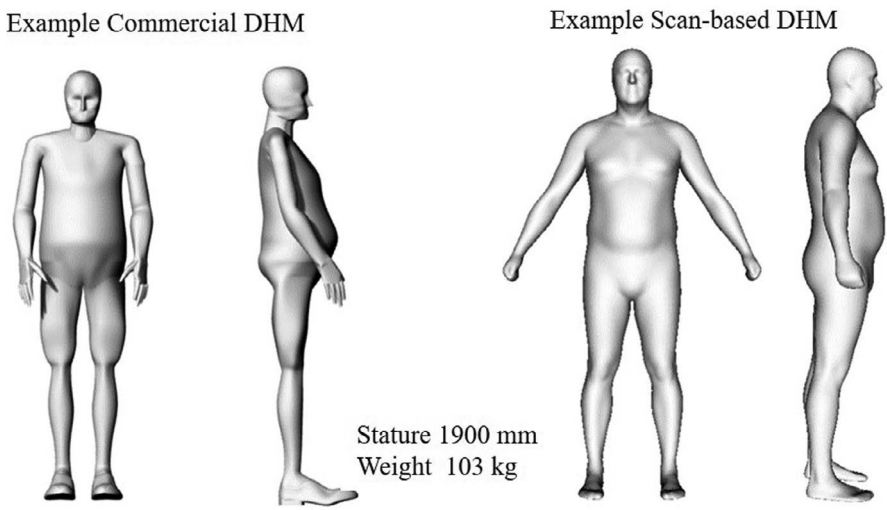


FIGURE 3.41 Left: Commercial articulated DHMs. Right: Scan-based DHM.

decade, 3D body scanning technology experienced an important development and growth (Daanen & Ter Haar, 2013; Daanen & van de Water, 1998). Body scanners delivered a digital copy of the 3D body surface, generating a 3D model of the body with the actual body shape and the possibility to calculate almost any measurement. This technology opened the possibility to create a new generation of DHM with an anatomical body shape that reproduces body features (e.g., curvatures, asymmetries, shape variations), relevant to achieve a good fitting design of a wearable. These are the scan-based DHMs.

Since the CAESAR™ project (Robinette et al., 1999), the 3D body scanning survey of reference, more than 20 similar large-scale studies have been performed around the world to characterize the body size and shape of different populations (Alemany et al., 2019). In particular, the easy accessibility of the CAESAR™ database, and the quality of the data, including 3D body scans, digital measurements, landmark coordinates as well as detailed documentation describing the data, has boosted the development of advanced 3D statistic tools to create synthetic 3D bodies. The concepts used behind these models are based on shape analysis methods that have been developed in Physical Anthropology to study the biological shape and shape change. The book "Morphometric tools for landmarks data" (Bookstein, 1992) is a useful reference for understanding the fundamentals of shape analysis methods.

Today, there are several whole-body scanning technologies and a few scanners that can capture the whole body in motion, which are referred to as 4D scanners. This has provided us with some tools for creating both physical and digital manikins. However, for wearable products, it is not always as easy as scanning and transferring a scan to a CAD software package or a 3D printer. For many CAD packages, the scans must be processed and adapted to the software.

There are several ways to generate digital manikins from 3D scans, and the best method depends on how we intend to use it and how much time and money we are willing to spend on it. Five examples, from simplest to most complex, are:

1. Use the original scan point cloud
2. Create a simple surfaced manikin
3. Create a simple watertight surfaced manikin
4. Create a standardized homologous watertight manikin
5. Use a parameterized manikin database to select a manikin

The first three methods were described and illustrated in Chapter 2, as were some of the issues to be careful about. All the methods require software tools to visualize and use the scans. Some visualization tools may come with the scanner, but some scanners do not provide this. Some scanners only provide 1D measurements that are automatically extracted and do not output the 3D point cloud itself, so they do not provide a 3D manikin. It is important to ensure that the point cloud is part of what the scanner provides. If a 3D point cloud is provided by the scanner it is also important that it be in a format that can be viewed in software tools such as the open source tool Blender™ or the commercial tools Polyworks™ or AutoCAD™.

Original Point Cloud

The output of a 3D scanner is usually a collection of 3D points called a point cloud. These can include landmark points that were pre-marked before scanning and identified and named after scanning. The original point cloud contains only points that were captured by the scanner and the points are not connected.

The point cloud can be used for visualizing or measuring landmark locations in a wearable depending on the software tool. Visualization of the body within a wearable can be done if the subject is scanned with the wearable in place and without it then have the scans overlaid. Scan overlays can be done in the open source software tool, Blender™ as well as several commercial software tools. Simple 2D overlays can even be done using screen captures and Powerpoint™. An example of the use of overlays was provided in Chapter 2, where we showed the range of pupil locations in a helmet.

With the point cloud, we do not have a surface so we cannot accurately measure between surfaces, such as the skin surface and a wearable surface, nor can we measure to any spot that does not have a point on it. For these measurements we need a surface.

Surfaced Manikin

A surfaced manikin is a very basic connection of the points to create surfaces. It can be used for visualization and can measure between surfaces to some degree. Some scanners provide surfacing options otherwise, software tools (such as Meshlab™ or Polyworks™), may be needed to do the surfacing. The two most common surfacing methods are polygonal mesh and non-uniform rational B-splines (NURBS).

A polygonal mesh connects sets of three points with flat triangles. The points connected are the original scan points and will be as precise and accurate as the scanner that obtained them. However, the surface of the triangles can be different from the true body surface, so they can be less accurate.

NURBS surfacing is commonly used in computer graphics and CAD software tools. This surfacing method connects the points with surfaces, but it can also remove some of the original points. It fits the points to curves and the original points do not always lie on the NURBS surface. In addition, it simplifies the surface so some of the detail is smoothed away. However, the curved surface is smoother than the polygonal mesh surface and can be easier to work with.

When there are large areas of missing data, such as the armpits, crotch area, top of the head, etc., these methods typically leave holes or make connections incorrectly. When these areas are important, they may need to be edited. Surfaces may not be accurate, so it is important to test the accuracy and save the original point cloud.

Watertight Surfaced Manikin

A watertight manikin is one that does not have any holes or spaces between points. It is required for insertion into many CAD software packages and sometimes for 3D printing. The first type of watertight surface manikin builds on the basic surfaced manikin but fills the remaining holes with software algorithms or tools such as Meshlab™. Watertight surfaced manikins add points that were not in the scan and can also remove some of the original points to get smooth surfaces, just as simple surfacing did.

Standardized Homologous Watertight Manikin

This is a type of manikin that is produced using standard template models. The use of template models for anthropometric modeling was introduced by Allen, Curless, and Popović (2003) when the data from the CAESAR project (Blackwell et al., 2002; Robinette et al., 1999, 2002) became available. *Standard template models* are pre-processed manikins that have: (1) a standard set of landmarks that can be used for Procrustes alignment, (2) a fixed number of evenly distributed points, and (3) the points are organized so they can be treated as homologous points.

To create the manikin, a new scan is aligned to a template model using Procrustes alignment of the landmarks. Then, the template model is deformed (or re-shaped) to match the new scan. The re-shaped matching model is said to be standardized because all points are treated as homologous to the template model, with a fixed number of evenly distributed points. When the template model is also watertight, the matching models are also watertight. The new set of points are standard points allowing the direct comparison of manikins. However, they are not the original scan points.

Several people have demonstrated that this process can enable better subject comparison (Alemany et al., 2019; Ballester et al., 2016; Trieb et al., 2013). However, the quality of the template model is important, and different template models may be needed for different genders, ethnicities, and age groups. This is still being studied.

While this method seems to work for the whole body in a standing pose, it is not clear that it will work well for body segments or other poses. For example, template models of the head do not seem to be as good as template models for the whole body. There are several reasons for this. First, there are not enough good, reliable landmarks on the back and top of the head, so the template alignments are biased toward the face. Second, the head is very spherical so there is not a strong long axis or second axis, so the alignments have a large error compared to manual 1D head measurements. Third, head scans include hair. Hair affects the shape, hides important landmarks, and dark hair for eyebrows, mustaches, and beards do not reflect light so those areas may not provide data points for some scanners.

Parameterized Database Manikin

This method uses a sample of standardized homologous watertight manikins that also have manual 1D measurements for important and relevant body dimensions. In the past ten years, several researchers have developed databases of parameterized manikins and demonstrated how they can be used for creating manikins to represent specific cases (Alemany et al., 2019; Ballester et al., 2018; Trieb et al., 2013). Since every standardized manikin has the same number of points, evenly spaced, and at standard locations, these researchers have demonstrated how PCA using all the points can be useful for creating manikins at desired case sizes.

If we have 20,000 homologous 3D data points, we have 60,000 homologous variables (3*20,000). When we do a PCA analysis of these variables we get 60,000 components from a PCA, and each subject will have an additional 60,000 variables for which they have a score. Each subject can be completely recreated from their component scores if they are known. This is a lot of extra variables, however, only

TABLE 3.14
Correlations between Manual Measurements and Standardized Manikin PC Scores

Correlation Matrix	PC1 Score	PC2 Score	PC3 Score	PC4 Score	PC5 Score
Age	–0.37	–0.54	0.06	–0.31	0.01
Weight	0.145	–0.93	–0.11	0.2	–0.01
Stature	0.98	–0.07	–0.01	–0.08	–0.02
Crotch Height	0.92	0.13	0.32	0.03	0.07
Knee Height, left	0.87	–0.15	0.37	–0.01	0.15
Knee Height, right	0.87	–0.15	0.36	–0.01	0.15
Bust Girth	–0.14	–0.92	–0.02	0.04	0.07
Waist Girth	–0.03	–0.20	–0.03	0.04	0.02
Hip girth	0.01	–0.93	–0.13	0.27	–0.06
Arm Length, left	0.81	0.03	0.36	–0.05	–0.32
Arm Length, right	0.81	0.03	0.36	–0.05	–0.31

the first few may be needed for case and manikin selection. Since all the components are independent of each other by definition, they are unrelated so we can change one without affecting another. Also, the components are ordered with those explaining the greatest amount of overall variation in the data points first and the least amount of overall variation last.

For example, we had a dataset of more than a thousand female subjects for whom we had manual 1D measurements (and age) and 3D scans. We created the standardized homologous manikin for each subject from her scan. Then we calculated the correlations between manually taken 1D measurements and the PC scores from the standardized 3D manikin data. These are shown in Table 3.14.

The first two PC variables were strongly correlated with the selected 1D manual measurements. PC1 is strongly correlated with stature (0.98) and other height variables. Stature alone explains 96% (0.98 squared) of PC1, therefore if we know Stature, we should be able to accurately predict PC1. PC2 is strongly and negatively correlated with Weight (–0.93) and weight-related variables. If we know Weight and/or Hip Girth we should be able to accurately estimate the PC2 score. (The negative value indicates that as weight gets larger PC2 gets smaller.) PCs 3 through 5 had significant but small correlations with the manual measurements and age.

To visualize these relationships, we show the range of body shapes represented by the first five PCs in Figure 3.42. The shapes go from the mean – 3 SDs (small), to the mean + 3 SDs (large) for each individual PC.

We see that the stature clearly gets taller for PC1. For PC2, we see that the horizontal body size clearly gets smaller as we move from –3SD to + 3SD. In other words, for this PC as we move from the small values to the large values of the PC the body width does the opposite, moving from large width to small width. The statures are the same for all PC2 values. This type of PC is called a contrast and it represents an inverse relationship between stature and body width. The other three are less clear,

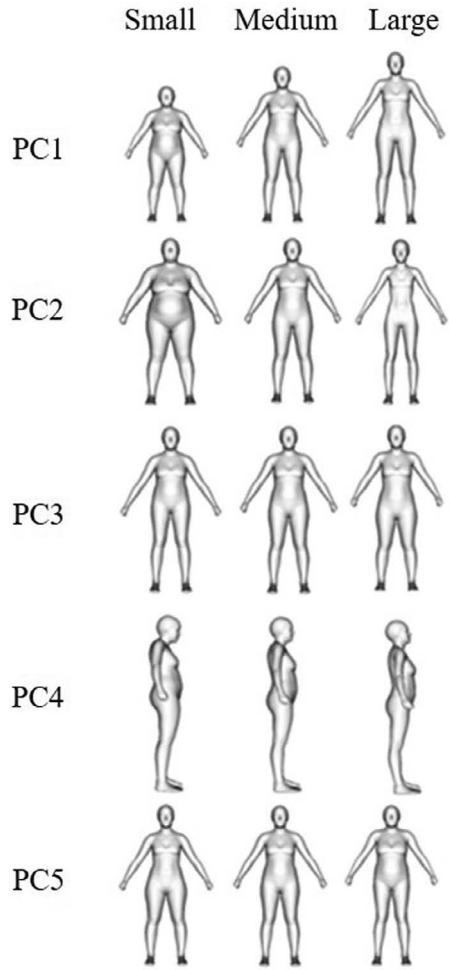


FIGURE 3.42 Body shapes for 5 PCs.

although it is interesting to note that PC4 seems to show moving from a hunched back to a more erect back. The variable with the largest correlation with PC4 is age ($r = -0.31$). This is a posture change that occurs to some of us as we approach age 60 and older but would be less evident in younger adults.

Why is this useful? If we have manual 1D measurements on the subjects in the database, we can estimate a new case without having to scan a person. We can calculate regression equations to predict the component scores from 1D measurements, and the component scores can be used to create a statistically based model of the individual. In other words, any new person's 1D measurements can be used to predict and create a standardized watertight manikin model representing them. Then, we can choose to use a statistical manikin created by PCs or a manikin of an actual person near the case location in the PC score dataspace.

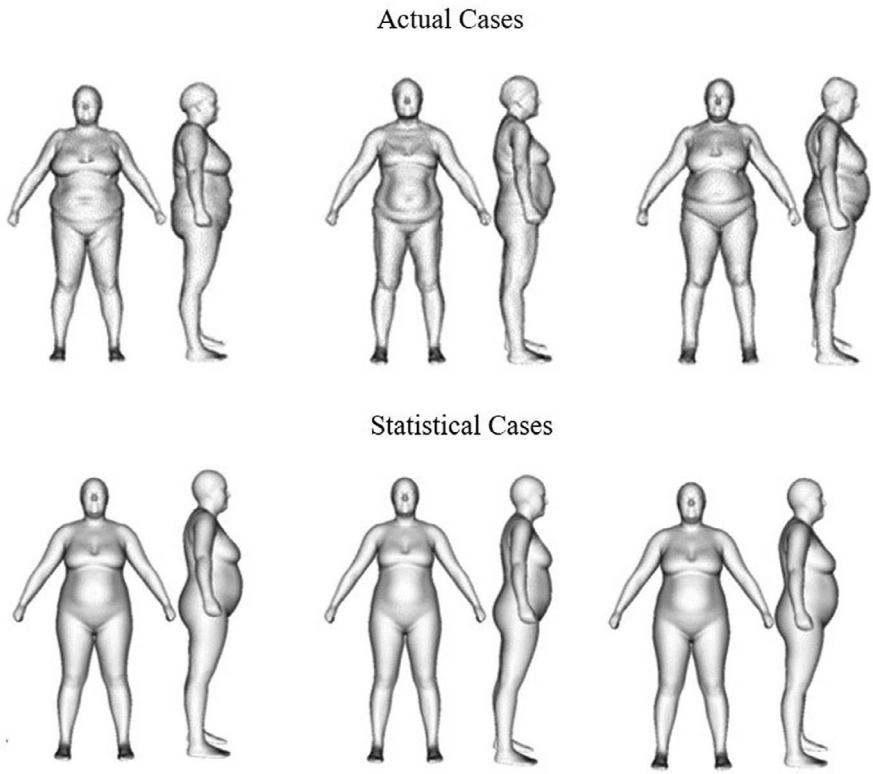


FIGURE 3.43 Actual cases and their corresponding statistical model.

Just as with physical manikins, it is important to compare the manikin to the actual person to ensure the manikin is representative of the primary or most important measurements. In Figure 3.43, we show three examples of actual people versus statistically based manikins predicted from their stature and weight.

The three subjects chosen had large weight values so they would be expected to have some extra soft tissue contours that a slim person might not have. As can be seen, the statistical cases are smoother, but they have very similar contours.

We used Stature and Weight to illustrate because they had the strongest correlations with the first two PCs. However, it might be better to use more product-specific measurements, such as Hip Circumference and Crotch Height to create a manikin for a pant. Then, at least those two measurements would be exactly what we want. These will vary depending on the product and which measurements are the most important.

This method seems to result in reasonable digital manikins for the whole body in the standing pose. However, this may not be the case for the head and may not work well for smaller body segments or other body poses. For the head segment, there are problems with the template model deformations because there are so few landmarks on the back and top of the head. Also, the up versus down direction definition is a bigger proportion of measurement variability for head and face measurements than

FIGURE 3.44 Sequence of processed body scans during a vertical jump captured with the 4D scanner Move4D.

it is for total body measurements. For head and face measurements 2 mm is a lot of error, but for total body measurements, 2 mm is in the noise.

The whole body in the standing pose has a strong long axis to establish the strongest first component and the up versus down direction. In the seated pose, the points along the thighs and those on the abdomen are close together and will complicate the template model deformation and the components.

Therefore, while parameterized database manikins show huge promise for many kinds of wearables, they may not be the best option for others. These manikins should be evaluated for their precision and reliability for the most important measurements related to the product.

New technologies are being developed to capture the body in motion, and this may provide a way to deal with pose or posture changes. We refer to this as 4D scanning technology that enables the capture of the body shape in movement such as shown in Figure 3.44. The scans have been displaced horizontally to show the body shape changes during movement.

The first research conducted with 4D addressed the modeling of the variations of the shape during movement considering different body types (Bogo et al., 2014; Pons-Moll et al., 2015). This type of dynamic model uses an internal skeleton attached to the external mesh to change and control the pose of the 3D body in combination with PCA models of shape and pose (see Figure 3.45). The body shape with the internal skeleton can be exported in the standard format FBX from Autodesk and is compatible with the CAD software used by the animation industry. At this moment, the research of dynamic DHMs is aiming to improve the accuracy of these models to reproduce any body shape in any pose with a realistic deformation.

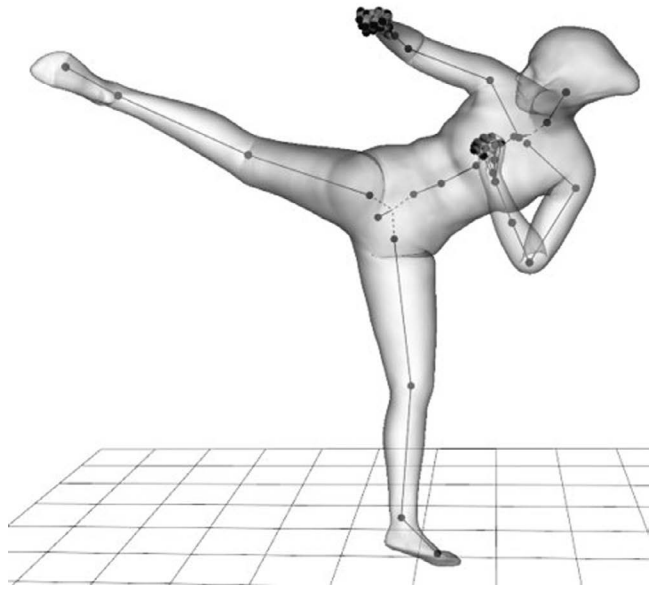


FIGURE 3.45 Dynamic 3D body model with the internal skeleton.

4D scanners are also a powerful tool to scan the user wearing a product and performing different tasks. This is a new way of analyzing the dynamic fit and checking the performance of the fitting compared to the CAD simulation done during the design process. The methodology to gather insights with this technology and transfer them to the design process is not still fully developed and is considered a future line of research.

REFERENCES

Alemany, S., Uriel, J., Ballester, A., & Parrilla, E. (2019). Three-Dimensional Body Shape Modeling and Posturography. In *DHM and Posturography* (pp. 441–457). Elsevier.

Alexander, M. W., McConville, J. T., & Tebbetts, I. (1979). *Anthropometric Sizing, Fit-Testing and Evaluation of the MBU-12/P Oral-Nasal Oxygen Mask* (Technical Report AMRL-TR-79-44). Aerospace Medical Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA074723.pdf

Allen, B., Curless, B., & Popović, Z. (2003). The Space of Human Body Shapes: Reconstruction and Parameterization from Range Scans. *ACM Transactions on Graphics (TOG)*, *22*(3), 587–594.

Ballester, A., Parrilla, E., Piérola, A., Uriel, J., Pérez, C., Piqueras, P., Nácher, B., Vivas, J. A., & Alemany, S. (2016). Data-Driven Three-Dimensional Reconstruction of Human Bodies Using a Mobile Phone App. *International Journal of the Digital Human*, *1*(4), Article 4. https://doi.org/10.1504/IJDH.2016.084581

Ballester, A., Pierola, A., Parrilla, E., Uriel, J., Ruescas, A. V., Perez, C., Dura, J. V., & Alemany, S. (2018). 3D Human Models from 1D, 2D and 3D Inputs: Reliability and Compatibility of Body Measurements. *Proceedings of 3DBODY.TECH 2018 - 9th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, Lugano, Switzerland, 16–17 October 2018*, 132–141. https://doi.org/10.15221/18.132

Blackwell, S., Robinette, K. M., Boehmer, M., Fleming, S., Kelly, S., Brill, T., Hoeferlin, D., & Burnsides, D. (2002). *Civilian American and European Surface Anthropometry Resource (CAESAR). Volume 2: Descriptions*. United States Air Force Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA408374.pdf

Bogo, F., Romero, J., Loper, M., & Black, M. J. (2014). FAUST: Dataset and evaluation for 3D mesh registration. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3794–3801.

Bookstein, F. (1992). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Daanen, H. A. M., & Ter Haar, F. B. (2013). 3D Whole Body Scanners Revisited. *Displays*, *34*(4), 270–275. https://doi.org/10.1016/j.displa.2013.08.011

Daanen, H. A. M., & van de Water, G. J. (1998). Whole Body Scanners. *Displays*, *19*(3), 111–120. https://doi.org/10.1016/S0141-9382(98)00034-1

Dainoff, M., Gordon, C., Robinette, K. M., & Strauss, M. (2004). *Guidelines for Using Anthropometric Data in Product Design*. Human Factors and Ergonomics Society.

Daniels, G. S. (1952). *The "Average Man"?* Air Force Aerospace Medical Research Lab. Wright-Patterson AFB OH. https://apps.dtic.mil/sti/tr/pdf/AD0010203.pdf.

Duffy, V. G. (2007). *Digital Human Modeling*. Springer.

Duffy, V. G. (2008). *Handbook of Digital Human Modeling: Research for Applied Ergonomics and Human Factors Engineering*. CRC Press.

Ervin, C. A., & Robinette, K. M. (1987). *A Manual for Administering a Standardized Dexterity Test Battery (U)* (Technical Report AAMRL-TR-87-036). Armstrong Aerospace Medical Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA188718.pdf

Evans, S. M. (1976). *User's Guide for the Programs of COMBIMAN* (Technical Report AMRL-TR-76-117). Aerospace Medical Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA038323.pdf

Harrison, C. R., & Robinette, K. (2002). *CAESAR: Summary Statistics for the Adult Population (Ages 18-65) of the United States of America* (Technical Report AFRL-HE-WP-TR-2002-0170). United States Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division. https://apps.dtic.mil/sti/pdfs/ADA406674.pdf

Hudson, J. A., Zehner, G. F., Parakkat, J., & Choi, H. J. (2006). *A Methodology for Evaluating Advanced Operator Workstation Accommodation* (Technical Report AFRL-HE-WP-TR-2007-0016). Air Force Research Laboratory, Human Effectiveness Directorate.

Kennedy, S. J., Woodbury, R. L., & Madnick, H. (1962). *Design and Development of Natural Hand Gloves* (Technical Report Series Report No. 33). Headquarters Quartermaster Research & Engineering Command, U.S. Army. https://apps.dtic.mil/sti/pdfs/ADA047962.pdf

Kuznets, S. (2018, August). Making the armor. How they did it in the Middle Ages? *Forge of Svan*. https://forgeofsvan.com/making-the-armor-how-they-did-it-in-the-middle-ages/

Leetch, B. D., & Bowman, W. L. (1983). *Articulated Total Body (ATB) "View" Program Software Report, Part I, Programmers Guide* (Technical Report AFAMRL-TR-81-111, volume I). Air Force Aerospace Medical Research Laboratory.

Pons-Moll, G., Romero, J., Mahmood, N., & Black, M. J. (2015). Dyna: A Model of Dynamic Human Shape in Motion. *ACM Transactions on Graphics (TOG)*, *34*(4), 1–14.

Robinette, K. M. (1986). Three-Dimensional Anthropometry-Shaping the Future. Human Factors Society Inc., Santa Monica, CA. *Proceedings of the Human Factors Society-30th Annual Meeting*, *Vol. 1*., 205.

Robinette, K. M., Blackwell, S., Daanen, H. A. M., Boehmer, M., Fleming, S., Brill, T., Hoeferlin, D., & Burnsides, D. (2002). *Civilian American and European Surface Anthropometry Resource (CAESAR) Final Reports, Volume I: Summary* [Technical Report]. Air Force Research Laboratory, Human Effectiveness Directorate. https://apps.dtic.mil/sti/citations/ADA406704

Robinette, K. M., Daanen, H., & Paquet, E. (1999). The CAESAR project: A 3-D surface anthropometry survey. *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No. PR00062)*, 380–386.

Robinette, K. M., Ervin, C. A., & Zehner, G. F. (1986). *Dexterity Testing of Chemical Defense Gloves (U)* (Technical Report AAMRL-TR-86-021). Aire Force Aerospace Medical Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA173545.pdf

Robinette, K. M., & Hudson, J. A. (2006). Anthropometry. In *Handbook of Human Factors and Ergonomics* (3rd ed., pp. 322–339). John Wiley & Sons.

Scataglini, S., & Paul, G. (2019). *DHM and Posturography*. Academic Press.

Seeler, H. W. (1961). *Development of Oral-Nasal Masks, Oxygen, MC-1 and MBU-5/P* (Technical Report ASD Technical Report 61-395). Life Support Systems Laboratory, Aerospace Medical Laboratory. https://apps.dtic.mil/sti/tr/pdf/AD0267151.pdf

Tian, R., & Duffy, V. G. (2011). Computerized Task Risk Assessment Using Digital Human Modeling Based Job Risk Classification Model. *Computers & Industrial Engineering*, *61*(4), 1044–1052. https://doi.org/10.1016/j.cie.2011.06.018

Trieb, R., Ballester, A., Kartsounis, G., Alemany, S., Uriel, J., Hansen, G., Fourlic, F., Sanguinetti, M., & Vangenabith, M. (2013). EUROFIT—Integration, Homogenisation and Extension of the Scope of Large 3D Anthropometric Data Pools for Product Development. *4th International Conference and Exhibition on 3D Body Scanning Technologies*, *Long Beach, CA, USA*, 19–20.

Veitch, D., & Davis, B. (2009). Practical Application of 3D Data for Apparel Industry Use. *Proceedings of the 7th World Congress on Ergonomics-International Ergonomics Association (IEA)*, 8.

Walker, B. (2019, December). PCA Is Not Feature Selection. *Towards Data Science*. https://towardsdatascience.com/pca-is-not-feature-selection-3344fb764ae6

Whitestone, J., & Robinette, K. M. (1997). Fitting to Maximize Performance of HMD Systems. In *Head Mounted Displays: Designing for the User*. McGraw-Hill.

Zehner, G. F., Meindl, R. S., & Hudson, J. A. (1993). *A multivariate anthropometric method for crew station design: Abridged* (Technical Report AL-TR-1993-0054). Armstrong Laboratory, Air Force Material Command. https://apps.dtic.mil/sti/pdfs/ADA270652.pdf

4 Testing and Analysis Procedures

Kathleen M. Robinette

ABSTRACT

This chapter addresses all the parts of the Sustainable Product Evaluation, Engineering, and Design (SPEED) process that involve testing and analysis during the design and sizing loops. We will be using inferential statistical methods to test, draw conclusions, and make predictions about our product, and its design features and its sizing. We begin with a review of experimental design and analysis methods applicable to wearable design. Then we present procedures with examples in the design and sizing loops starting with the simplest and moving to the most complex. Fit testing as part of the design process is iterative and, in the beginning, the tests are simple, quick, and rudimentary. As the design develops, the tests develop as well, becoming more precise and comprehensive. Changes to the design concept are easier and less costly when done in the early stages than in the later stages. Toward the end of product development changes become more expensive and there is less room for error. The general purpose of all testing is to enable us to make better decisions and choices. This involves balancing or trading off risk, cost, and benefit. Risk is the chance of making the wrong decision and the impact of the error. There is always a risk because we can never be 100% certain about any choice. It is important to try to minimize the risk, but at some point, the additional risk reduction is not worth the cost. The procedures described in this chapter can help us manage risk and make good design decisions within a cost range we can afford.

The Sustainable Product Evaluation, Engineering, and Design (SPEED) process is a systems engineering approach and that means systems testing is essential, and as Budurka (1984) stated, it is iterative and incremental. He gave us the following definition of systems engineering:

> Systems engineering is the iterative but controlled process in which user needs are understood and evolved, through incremental development of requirements specification and system design, to an operational system.

For wearables this includes fit testing with the wearer considered to be part of the system. Testing is used for designing complex systems because when one part of the system changes it can affect all or some of the other parts. In other words, we can fix one problem and inadvertently cause another. This is particularly an issue for wearables because of the complexity, the wide range of differences and the constantly changing human body. For example, fat, muscle, and bone all have different densities and the

DOI: 10.1201/9781003397533-4

locations and distribution vary from person to person, from day to day and even from minute to minute as we move and breathe. Also, different areas of the body surface have different numbers of nerve endings. Areas with more nerve endings are more sensitive to pressure. This means one spot on the nose, for example, might be comfortable for the pressure of the nosepiece to sit, but a spot just a millimeter or two away might be intolerable. Furthermore, the location that is comfortable for one person is different than the location that is comfortable for another, and some people can tolerate more pressure than others. Changes in product properties and location with respect to the body affect its quality of fit and performance and changes or differences in people can affect the performance of the product. We do not have the technology to quantify these types of things effectively other than fit testing with human subjects.

Fit testing captures information about the impact of the physical interface between the wearable and the person. This includes such things as comfort, stability, slippage, temperature, pressure, component locations with respect to the subject, and personal preference. Many aspects of a wearable can impact fit including the materials used, the adjustment concept and mechanism used, the environment in which it will be worn, the area of the body with which it comes in contact, product contours, and product mass and mass distribution. As a result, fit testing also reveals design issues and potential design solutions, making it essential to a successful design process.

In the field of Probability and Statistics these kinds of tests are referred to as experiments and designing or setting up the conditions for each test is called *experimental design*. We cannot be 100% certain about all things, but we can evaluate the probability that they are true or are within a tolerable range. If we are confident that we have the correct answer it means, we believe the chance (or probability) of being wrong is low or that the probable amount of error is tolerable. We design experiments to give us confidence in the answers to our questions with a minimal amount of cost in terms of time and expense. Our confidence is also influenced by the amount of risk we are willing to take if we are wrong. Will an error in judgment affect the quality of the product, the cost of the product, the marketability of the product, the production time (especially if we must start over), etc.? Therefore, the goal of experimental design is to provide confidence in our design decisions for an affordable cost.

Every good test begins with a clear statement of the question or questions we are trying to answer. The results and conclusions that can be drawn from a test depend on the way the data are collected (experimental design), and the type of analysis we plan to use, so we must understand both the experimental design and the analysis before we begin. Therefore, we begin this chapter with a general review of experimental design and analysis methods most likely to be needed for fit testing of wearables. The goal is to enable the reader to be able to design tests and do the analyses. However, it should also be helpful as a guide for a professional statistician to understand wearable design and the questions the manufacturers, buyers, designers, and engineers need to be answered. In this sense, it would allow meaningful communication between people with different backgrounds and expertise.

For more information, there are many good texts on experimental design such as Design and Analysis of Experiments (Montgomery, 1976) and good books on statistics for decision making such as Statistics for Modern Business Decisions

(Lapin, 1978). There are also good books that focus on specific aspects of design or analysis, such as Sampling Techniques (Cochran, 1977) or Response Surface Methodology (Myers & Montgomery, 1995).

After reviewing experimental design and analysis methods in general, we discuss experimental design and analysis specific to different points in the process. If we review the SPEED process as shown in Figure 1.1 in Chapter 1, we see that we divide wearable testing into two groups: (1) design loop tests and (2) sizing loop tests. The design loop testing helps us optimize the design quality given the cost and schedule limitations. The purpose is to either refine the design or refine the test procedures in preparation for the sizing loop tests. In this loop, questions are answered about the weight, product contours, adjustment features, materials, etc., that are comfortable enough, protect and perform the desired functions for the product, and permit the wearer to do the tasks they need or want to do. The design loop is also where we determine which variables are the best predictors of good fit and the range of good fit for those variables in the base size.

There are three types of design loop tests listed in the diagram: (1) pilot tests, (2) stand-alone trade studies, and (3) prototype fit tests. Pilot tests are pre-experiment guidance tests that evaluate our test procedures before we begin testing. They ensure that the tools, processes, and the concept-of-fit (COF) are reasonable and effective. The other two types of tests are design guidance tests. The design guidance tests may have fewer than the full set of sizes, only some of the test conditions, or only a partial or non-functioning prototype than tests done in the sizing loop. They also do not necessarily require the use of a representative sample of the target user population (TP). Nonetheless, the design loop tests of full prototypes can be quite complex.

The sizing loop focuses on the sizes after the product is designed. This testing tells us if we are missing anyone with our sizes, and what assortment of sizes is needed for a given population. It enables us to evaluate the benefit of including a particular size in terms of the proportion of the population who would wear it. For example, even if we must fit the extremes of a population, it can be more cost-effective to custom-fit some people rather than build a size that may never be needed.

The sizing loop is the final loop for a new product we are developing, but it is also the loop to use when we are evaluating a product that is already on the market. Perhaps we want to determine if the product will work for our population before we purchase it, or we may want to know if the product conforms to our fit standard. When this is the purpose of the full-fit test, we refer to it as a *fit audit*.

We do fit audits to ensure that our new products fit the same people as our previous products so that if our wearers achieved a good fit in a particular size previously, they would get a good fit in the same size of the new product. This can increase customer loyalty and confidence in online size selection. We also do fit audits if we do not know who we are fitting or missing so we can improve the size assortment and create a sustainable fit standard for future products.

Sizing loop tests examine all aspects of the fully functioning wearable system and require human subjects representing the full TP. Both design loop tests and sizing loop tests can be iterative; however, it is hoped that the design loop tests will provide enough guidance for the development of the first fully functioning whole system, such that the sizing loop test would only have to be done once. Furthermore, if we

have a sustainable fit standard and we verify that the product meets our standard during the design loop, there may be no need to do a sizing loop test. In that instance, we only need to do the sizing loop if we need to do a full fit audit because we have changed something substantial such that we may need to change the sizing.

The sizing loop is done to: (1) verify that the final product sizing is acceptable, (2) consider cost versus benefit before selecting the set of sizes to produce or purchase, (3) determine how many of each size to build or purchase (*the tariff*), and (4) help the wearer find the best size and adjustment. If done well, the final test can also be used to adjust the sizes and numbers to purchase for new populations or markets as well.

The complexity and scope of each test are also dependent on resources and constraints. In other words, what are our time, money, materials, and space limitations for the test? What is the deadline for the answer? (Sometimes, in industry, the deadline for the answer is yesterday.) What tools and product prototypes will be available? Even if time and money are very limited, there is usually a lot that can be done if we use good statistical methods. For example, we might choose a more limited sampling method that will only address a portion of our TP or we might accept a lower confidence level. Therefore, it is best to start with an idea about the range of resources and constraints and choose our risk priorities ahead of time.

EXPERIMENTAL DESIGN

The experimental design is a plan that will meet the test goals, while minimizing risk and is compatible with the available resources. It is designing what you are sampling and how you are going to measure.

Sampling is a process in which a finite number of observations or entities are taken from a larger population. When we say "population" we are not just talking about people and a sample from a population is also not just a sample of people. For wearable testing at least three populations are sampled: (1) people (test subjects), (2) products, and (3) conditions. Product populations are all the possible designs, sizes, adjustment mechanisms, materials, sized components, etc. for the product. Conditions are all the possible use conditions for the wearable such as the environments and the tasks the wearers might do. For example, in the apparel industry, a prototype garment that serves as an example is also called a sample, and it can be thought of as an entity or single example taken from the population of all possible garments. To help alleviate confusion, we call the product combinations and the conditions *treatments* in our experimental designs.

Fit testing has two parts:

1. What you are sampling:
   a. Subjects
      i. Who
      ii. How many
   b. Product Treatments
      i. What treatments to test
      ii. How to assign treatments

   c. Conditions Treatments
      i. Environment
      ii. Duration
      iii. Activities
2. How to measure:
   a. Variables to include
      i. List of measurements
      ii. Questionnaire questions
   b. Analyses to be used

Randomization is the foundation of experimental design. By *randomization*, we mean that the selection of subjects, the allocation of treatments and the order in which the trials of the experiment are run are randomly determined. It helps us minimize the effects of extraneous factors and sampling bias.

Sampling bias is a tendency to favor the selection of units with particular characteristics. This type of bias is minimized through randomization and stratification. A random sample is obtained by choosing units in a way that allows every unit an equal chance of being selected. This applies to the test subjects and the treatments. That means we should randomize the order of presentation of each condition such as the sizes to be worn during the test. This is very important for wearables because subjective opinions are critically important and, as we learn about a product, our opinion can change.

Randomization can be achieved with simple random sampling, or with a stratified random sampling. With a *simple random sample*, each item from a population of items has an exactly equal chance of being selected.

Stratification is the dividing of the population into groups and selecting samples from each group. It is a method for spreading the data out, so we get more variability with the same number of observations, as well as to ensure we get sufficient representation from important population subgroups. It can be helpful to keep the overall sample size small while getting enough variability to make good design and sizing decisions.

In Figure 4.1, we show the estimated stature distribution for males from the US population (Harrison & Robinette, 2002). The curve represents the approximate number of subjects at each stature. The stature values are shown at the top of the figure and the standard deviations are shown at the bottom.

As can be seen, there are more people (observations) in the middle than at the lower or higher extremes of stature. The central 160 mm range (people with statures from 1686 to 1846 mm) contains approximately 68.2% of the population while the lower 160 mm range (people less than 1686 mm in stature) and upper 160 mm range (people more than 1846 mm in stature) are each estimated to contain just 15.9% of the population. Even though the three ranges of stature are all 160 mm, the proportion of people in each range differs.

If we gather a simple random sample of 100 stature observations, we expect 68 of them to be within the middle 160 mm, 16 of them to be of the smallest 160 mm, and 16 to be of the largest 160 mm. If our sample size were only 30, we would expect 20 observations to be from the middle with only 5 from each of the two tails. In addition, the 5 at each end who are outside the middle range are more likely to be close to the middle than they are to be farther away from it.

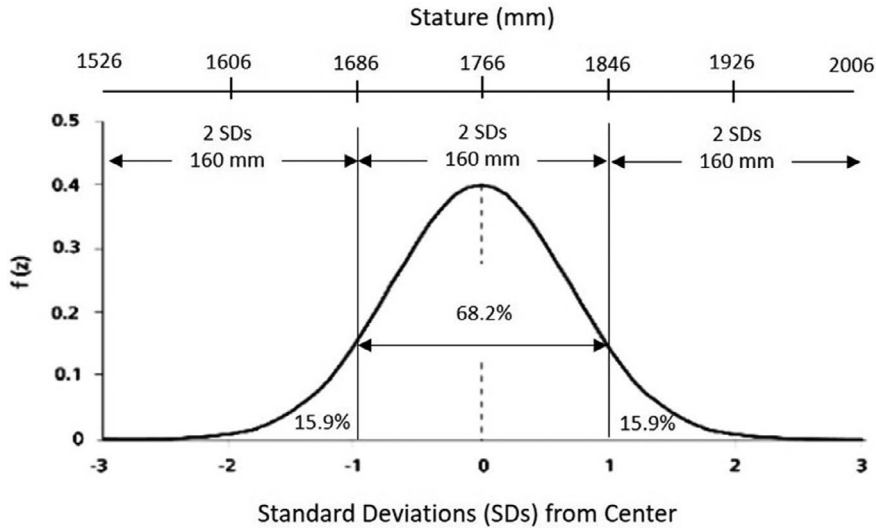


FIGURE 4.1 Normal distribution with standard deviations for stature.

If getting representatives who are small and large is important, we may want to divide our sample into small, medium, and large groups (the strata) and randomly sample from the groups. This is called a stratified random sample. For example, we might use the three stature groups shown in Figure 4.1 as our strata and sample the same number randomly from each group. This would give us 33 people who are less than 1686 mm, 33 people who are between 1686 mm and 1846 mm, and 33 people who are larger than 1846 mm. With 99 subjects we would have a greater spread of stature in our sample than we achieved with 100 samples using the non-stratified approach.

The most common groupings used in stratification are demographic characteristics, such as that used in Chapter 2 for representing the TP. With demographic groupings, we stratify our sample to ensure we have enough representation from each group. For example, our TP might consist of gamers, including both males and females in two age groups of interest: 18–30 years, and 31–45 years. Table 4.1 is an example to illustrate how stratification affects the sample. In this example, we estimated from a marketing survey of the TP that 75% are male, 25% are female, 75% are 18–30 years old, and 25% are 31–45 years old. First, we estimated the sample size in each strata if we collected a simple random sample of 120 people ($n = 120$) from the TP. The result would approximately be the numbers in the top part of Table 4.1. We see here that we would only get 7 females aged 31–45 years. That is a small sample for this group so they might not be well represented.

Next, we estimated the overall sample size if we specified that there be at least 30 subjects in each stratum. A sample size of $n = 30$ is a "rule of thumb" minimum sample size in statistics. This is the threshold above which the distribution of the mean may be estimated by the standard normal distribution (z) rather than the Student's t distribution and has to do with precision in the estimation of the mean. Typically,

TABLE 4.1
Simple Random Sample Versus Stratified Sample

Simple Random Sample, *n* = 120	**Ages 18–30**	**Ages 31–45**	**Total**
Male	*n* = 67	*n* = 23	*n* = 90
Female	*n* = 23	*n* = 7	*n* = 30
Total	*n* = 90	*n* = 30	*n* = 120
Simple Random Sample Minimum, *n* = 30			
Male	*n* = 287	*n* = 99	*n* = 386
Female	*n* = 99	*n* = 30	*n* = 129
Total	*n* = 386	*n* = 129	*n* = 515
Stratified Sample Minimum, *n* = 30			
Male	*n* = 30	*n* = 30	*n* = 60
Female	*n* = 30	*n* = 30	*n* = 60
Total	*n* = 60	*n* = 60	*n* = 120

we would want at least 30 subjects when estimating inferential statistics for a demographic group. As we see in the second part of Table 4.1, if we use a random sample and do not stop sampling until we have at least 30 in each group, that will require a sample size of $n = 515$.

The third option we illustrate is what we refer to as the stratified random sample. Here we used a stratified random sample with $n = 30$ randomly selected subjects in each stratum. We still get a total of 120 subjects, but they would be evenly distributed in the subgroups as shown in the stratified sample portion of Table 4.1.

With the stratified random sample, we can estimate the unstratified population distribution by assigning a weight to each subject to make the total distribution match the population. That way, we get sufficient representation from the subgroups and sufficient representation of the TP with about a quarter of the number of subjects that we would have otherwise needed.

While stratification can be helpful in some cases, it is not always feasible to do so. Stratification requires the ability to obtain subjects within each stratum (group) in a reasonable, cost-effective manner. Sampling by stature, for example, can be problematic if we do not have stature data on candidates or people we are trying to recruit. It is typically not cost-effective unless we have an in-house subject pool for whom we have anthropometric data, whereas stratified sampling by gender, age, or other demographic variable can be feasible if we can obtain this information while recruiting with a questionnaire.

For tests with human subjects in a free country, true randomization is nearly impossible to achieve because we may choose to have someone in our sample, but they may choose not to participate. The people who choose not to participate may be people who have specific characteristics, such as very thin people who are self-conscious about their bodies. Therefore, despite our best efforts, our sample can be

biased. To compensate, it is a good idea to plan to examine important characteristics in our sample while we are collecting data. If it seems to be substantially biased, we can target the under-represented groups and provide additional incentives to increase participation.

A convenience sample is the easiest sample to collect because it is simply choosing the units/people that are convenient to get. This is not random, so it is more likely to have bias error. It is never ideal, but it can still be useful for evaluating our COF and sometimes for making some quick decisions between two simple options. A convenience sample might be simple or stratified.

The **sampling of treatments** can be done using either an *independent design* or a *repeated measures design*. An independent measures design consists of using different participants for each treatment. This is often the case when we are comparing new treatments to treatments we tested previously. A repeated measures design consists of testing two or more treatments on the same individuals. In other words, each subject tries more than one treatment. This lets us evaluate effects within subjects to give us a greater ability to find significant effects. This is common for small trade studies when all treatments are available and can be randomly assigned because it has more statistical power for finding differences with small samples of subjects. When there are just two treatments, the repeated measures design is called a *paired design*.

Another type of paired design is to match each subject with another subject who has similar characteristics. This is used a lot in medical research, but it is rarely used for wearable design, so we will not discuss it here.

There are many ways to randomly assign treatments to subjects for testing. Testing can be done using a sampling method called complete blocks, where every treatment is tested on everyone, in random order, or it can be done as incomplete blocks, where each person is only tested in some of the treatments. For example, a commonly used incomplete block is to estimate the size that will fit best and test the subject in three sizes: (1) the estimated best fit size, (2) one smaller size, and (3) one larger size. Randomizing the order of testing the three sizes will reduce the risk of bias toward one of the sizes. Testing neighboring sizes should help resolve which is the best fit and give us an indication about the amount of size overlap or size duplication. When there are many sizes or size combinations, this can be a less costly way to achieve good results. It does make the analysis more complex, however.

There are risk, benefit, and cost trade-offs associated with each sampling method and there is no one method that is best for all tests. In addition, it is sometimes necessary to make do with what you have and live with the risks. For example, many companies, particularly start-ups with unique products, are more worried about the risk of competitors finding out about their products than they are about the risk of getting poor fit or a poorly functioning product. In those instances, they may prefer to use only test subjects who are direct employees of the company. This is a very commonly used convenience sample, at least during most of the development process. It has a high risk of sample bias, both in the people who volunteer and in the responses they give. We have found that employees can often be biased for perfection, and thus more critical, or they can be biased for love of their product, and therefore not critical enough.

The experimental design is dependent upon both the questions that need to be answered and the analysis methods we plan to use. Therefore, it is important to consider the analysis methods before deciding on the experimental design.

ANALYSIS METHODS

Analysis of any test data begins with statistical descriptive summaries of the results. These are simple tabulations that can be done in any spreadsheet software such as Excel™, or in statistical software packages. These tell us how well the system performed and what some of the issues are. The summaries include results such as: (1) the proportion of subjects who passed versus failed, (2) the proportion of poor scores on each fit question, and (3) the average body measurements for people who passed versus people who failed. This kind of information helps us hone-in on the most essential analyses to do next. There will be many options and it can save time and money to narrow them down to target the big issues. The summaries also serve as a first status report for our management or customer, who often wants the answers before the testing is complete.

It is usually best to visualize some of the statistical summaries in graphs, such as a pie chart or a bar chart. It makes it easier to comprehend. In Figure 4.2, we show a summary in the form of a bar chart of the proportion of test subjects who achieved a good fit in each of two sizes. In this example, there was some overlap in the two sizes and 14.1% of the subjects got an equally good fit in both sizes. Only 7.1% failed to get a good fit. Next, we would investigate why and how we might accommodate the ones who failed. It could be we simply need to make a strap more adjustable or offer an additional pad. We might need another size, but it is possible we could accommodate them by making the size Large larger, spreading the Large and Small sizes

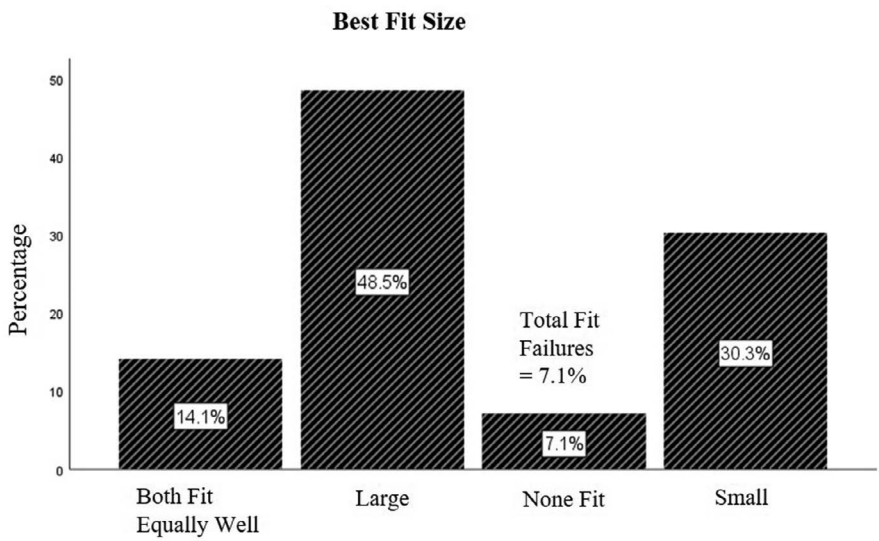


FIGURE 4.2 Bar chart summarizing pass/fail and size of best fit.

apart more to reduce the overlap where both fit. To do this we examine the range of body measurements that are accommodated in each size, which we call the *range-of-fit* assessment.

Fit mapping is an analysis comparing (or mapping) the range-of-fit of each size with the total variability in our sample. This can lead us to decide to make a change in the base size and size assortment before we transition to the sizing loop. Sometimes we re-test the new size assortment in the design loop and sometimes in the sizing loop. This will depend upon how confident we are in the effectiveness of the sizing change. If we are happy with our sizes after the prototype system test or if all our subjects received a good fit, then we won't make a change and we are ready to move to the sizing loop.

We use inferential statistical methods to determine what factors affect fit and what variables predict fit. There are several different options for alternative hypotheses, and each has a different statistical test. We will cover the most common ones for fit testing. The number of subjects or observations needed will vary depending upon the type of test and analysis. In Chapter 2, we provided some examples when we needed a sample that was representative of the variety of people in the TP. For some trade studies, we do not need to represent the variety of people and can have a good result with fewer than 30 subjects. For example, if we are testing the comfort of one type of padding versus another, it might be reasonable to assume that the set of subjects we use does not matter if we are just looking for the difference. We can use a paired-difference sample, where two different treatments are tested on every subject and the difference is evaluated. With this type of test, it is sometimes possible to detect a difference in comfort with as few as 10 subjects.

Many hypothesis tests are based upon the Central Limit Theorem (CLT) which essentially says that if many large enough random samples of a variable are taken from a population: (1) the distribution of the mean of those samples will approximate the normal distribution, (2) the mean of the means of the samples will approach the true population mean μ, and (3) the standard deviation of the means (called the standard error of the mean) will approach the population standard deviation σ divided by $\sqrt{n}$ where n is the number of samples that were drawn. This is true even if the distribution of the population itself is not normal. This theorem means the true distribution of the mean is the standard normal distribution (the z distribution).

Why do we care about the CLT? If we know the probability distribution of a mean, we can use it to determine the probability that a mean is different from another mean or another value. The CLT tells us what the probability distribution of the mean is.

Hypothesis tests for which the variable's distribution is known and whose parameters, such as the standard deviation or error, are needed are called parametric tests. Hypothesis tests for variables for which we have not specified (or do not know) the variable's distribution are called nonparametric tests.

There are two types of variables that might be evaluated: *continuous* and *discrete*. Continuous variables can take on any value along a range and they have an order to them, or in other words, a scale. They are also called *scalar* variables as a result.

Discrete variables are variables with groupings or categories that have a limited number of choices, such as yes or no, or small, medium, or large. They are sometimes called categorical variables.

There are two types of discrete variables that may be analyzed differently: *ordinal* and *nominal*. Ordinal variables are discrete variables that have an order to them, such as sizes listed as small, medium, and large. Nominal variables are discrete variables that do not have an order to them, such as the subjects' preference for the color of the product. Red is not higher or lower than Blue, for example.

Nominal variables can only be analyzed as discrete variables. Ordinal variables can also be analyzed as if they were continuous. Since they have an order, we can assign a number to the category that indicates its place in the order to simulate a continuous variable. For example, Small would be assigned the number 1, Medium the number 2, and Large the number 3. Treating ordinal variables as continuous variables can sometimes increase the power of the test.

The most common statistical procedures used are listed in Table 4.2 roughly in order from simplest to the most complex procedure. That is not necessarily the order in which they will be used. We also provide a brief description of each type of test and some examples of scenarios where they may be applied.

The procedures are grouped into two categories: (1) simple analyses and (2) General Linear Models (GLMs). There are some simple tests, both parametric and nonparametric, to use when there are only two treatments, outcomes, or variables, which are very common with trade studies. For more complicated analysis, the most common methods for wearables fall under GLMs. We list them as GLMs because that is where they may be found in some statistical software packages.

Student's *t*-test

Student's *t*-test (with the associated *t* distribution), is one of the most common tests used in research. The test is named for its creator, William Sealy Gosset, who used the pseudonym "Student" to publish his work because his employer, Guinness, would not allow its employees to publish. The letter t was used to represent the distribution, to differentiate it from the standard normal distribution that is represented by the letter z.

Thanks to Mr. Gosset, for inferential testing, it is not always necessary to have 30 observations to find reliable significant differences. The *t* distribution is used when the number of observations (n) or the number of degrees of freedom ($n - 1$) are <30. With 30 or more, the *t* distribution approaches the *z* distribution (in other words they are essentially the same), so the *z* distribution is usually used for large samples. For simplification purposes, most people refer to the test as the *t*-test regardless of which distribution is used.

The *t*-test is used in two ways: to assess if two means are different or to assess if a mean is different than a specific value (μ_0). The means can be subgroup means, regression means, overall sample means, etc. With the *t*-test, we use sample means and standard errors of the mean to estimate the probability that a mean is real (not equal to zero) or different from another mean. A probability (p) of .01 based on the *t* distribution means that only once out of 100 times would we find a mean or a difference this large that was not real.

TABLE 4.2
Commonly Used Statistical Tests

Procedure		Description	Examples of Uses
Simple Analyses	Student's *t*-Test	Parametric method for testing hypotheses about a mean or for comparing two means.	Compare two treatments when each subject only has one of the treatments. Compare one outcome mean to a specific value such as the mean score is greater than zero.
	Paired *t*-Test (Repeated Measures with two treatments)	Parametric method for testing hypotheses about the mean difference.	Compare two treatments when each subject has both treatments.
	Wilcoxon Signed-Rank Test	Nonparametric method to test consistent differences between pairs of treatments.	Compare two treatments when each subject has both treatments.
	Wilcoxon Rank-Sum Test	Nonparametric test for comparing two treatments.	Compare two treatments when each subject only has one of the treatments.
	Proportion Test	Test for estimating the proportion of observations with a certain outcome or trait.	Evaluate confidence in a proportion or percentage, meeting a pre-determined goal such as 90% are comfortable. Compare the proportion of pass versus fail for two independent products.
	Chi-square Test for Independence	Method to assess the goodness of fit between observed values and those expected theoretically or to test for independence of two variables.	Evaluate relationship (or independence) between two categorical or nominal variables.
	Correlation	Method to measure the degree of linear association or relationship between variables.	Evaluate linear relationship between continuous (scalar or ordinal) variables.

(*Continued*)

TABLE 4.2 ***(Continued)***
Commonly Used Statistical Tests

Procedure		Description	Examples of Uses
General Linear Models (GLM)	Analysis of Variance (ANOVA) with multiple comparisons tests	Methods for testing the impact of independent variables on dependent variables with more than two means or categories.	Compare anthropometric means for more than two sizes. Understand what variables including demographic variables impact size and fit scores.
	Linear Regression/Stepwise Linear Regression	Methods for estimating the relationships between a dependent variable (often called the "outcome variable") and one or more independent variables (often called "predictors" or "covariates").	Find variable combinations that are good predictors of best fit size. Establish a predictive equation for predicting an individual's best fit size.
	Discriminant Analysis	Method to find a linear combination of features that characterizes or separates two or more classes of objects or events.	Understand what characteristics are related to fit or size of best fit. Find variable combinations that are good predictors of best fit size.
	Logistic Regression	Method that uses a logistic function to model a binary (two options) dependent variable.	Predict the likelihood fit pass or fail in a size. Find the size most likely to fit.

The *t*-test assumes that each sample observation is selected randomly from a normally distributed population. Usually, the population referred to for this test is the population of the sample means. Therefore, because we believe the CLT, we assume this to be true. Since the *t*-test is based upon the *t* or the *z* distribution and their mean and standard error parameters, it is a parametric test.

The *t* distribution is different depending on the *df*. In this case, the $df = n - 1$. So, we might say there are many *t* distributions, a different one for every *df* less than 30. Once we have 30 *df* we have only one distribution, the *z* distribution, also called the Standard Normal Distribution. We specify the *t* distribution as:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \quad df = n - 1$$

where $\bar{y}$ is the sample mean, μ_0 is a specified value, *s* is the sample standard deviation, and *n* is the sample size.

Once we have calculated the value for *t* we must find where it falls in the *t* distribution for the related *df* to determine if it is large enough to indicate significance. Most statistics text books have a lookup table in an appendix for both *t* and *z* distributions (Lapin, 1978; Ott, 1977). This can also be done using a statistical analysis software. There are many free calculators available on the internet as well. We use a calculator at the Social Science Statistics website (https://www.socscistatistics.com/). This site has many calculators and even has a resource to help decide which statistics to use (Stangroom, 2021).

The *t*-test uses independent samples. Independent samples tests are those where each subject is only tested in one of multiple treatments. There can be many reasons why a person cannot or should not be tested in multiple treatments. For example, if the treatment must be worn for many hours for a given test it may not be feasible to test a person in more than one.

Data from independent samples look like the data in Table 4.3. Each subject has only one score and that is for one treatment. The assignment of the treatments is randomized to alleviate bias.

TABLE 4.3
Example of Independent Samples for Two Treatments

Subject	Score	Treatment
2	10	1
4	1	2
5	3	2
6	5	1
8	10	1
9	7	2
10	10	1
11	8	2
12	1	1
13	7	1

The biggest issue with the *t*-test is its overuse. It is meant for one comparison between two values and the probability level is based on one test. If we do many *t*-tests the probability level becomes unreliable. For example, if we set the probability level (*p)* at .01, which means that only 1 out of 100 times would we get a difference this large that is random, and we do 100 *t*-tests, we would expect one significant difference to occur that is just due to random chance. If there are many comparisons, it is best to use one of the GLM methods, such as ANOVA, or Linear Regression, along with their associated adjustments for multiple comparisons.

Paired *t*-test (Repeated Measures with two Treatments)

Paired tests are repeated measures tests that have only two treatments. Paired signifies both treatments were tried on each person or on matched pairs of people. For wearable design, we rarely use matched pairs of people. However, we often test two treatments on each subject because it allows us to look at the differences within subjects. Data from paired samples look like that in Table 4.4. Since we have both observations for each subject, we can subtract the scores for each subject and analyze the differences. By doing this, we filter out the person-to-person variability to obtain a meaningful comparison of the treatments.

When we are comparing two treatments, we are usually testing the hypothesis (H_0) that there is no difference (difference = 0) against the alternative (H_A) that one is better or worse than the other. We would sometimes want to test if one is better by at least a score of some value. In this instance, H_0 is that the mean difference, (μ_α) is greater than (or less than) some value D_0. To cover both hypotheses in one formula, when we are hypothesizing that there is no difference greater than (or less than) D_0, then the test statistic for a paired *t*-test is:

$$t = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} \quad df = n - 1$$

TABLE 4.4
Example of Paired Samples for Two Treatments

Subject	Treatment 1 Score	Treatment 2 Score	Difference
1	2	7	–5
2	1	10	–9
3	3	9	–6
4	7	8	–1
5	2	10	–8
6	5	7	–2
7	7	6	1
8	3	7	–4
9	4	8	–4
10	3	7	–4

This is the same t distribution discussed previously. It is simply applied to the mean of the differences rather than the difference of the means. In other words, the calculation of the test statistic is different than the Student's t-test, but the lookup table or distribution the result is compared to is the same. Again, it is usually best to use a statistics software tool or online calculator.

Wilcoxon Signed-Rank Test

This is a nonparametric test for comparing matched pairs, meaning there is no assumption about the underlying distribution of the data. The pairs may be two different conditions tested on subjects who have been matched for certain characteristics or they may be two conditions that have been tested on each subject. To calculate the signed rank, there are seven simple steps:

1. Calculate the differences for each sample pair: $p_i = X_{A_i} - X_{B_i}$
2. Record the sign of the differences (either positive or negative)
3. Calculate the absolute values of the differences: $|p_i|$
4. Rank order the $|p_i|$
5. Apply the sign of the differences to the ranks
6. Sum the negative ranks to get W_- and the positive ranks to get W_+
7. Select the W (the sum of the ranks from Step 6), that is, the smaller

The null hypothesis (H_0) is that there is a 50% chance of any p_i to be positive (or negative). Under H_0, half of the ranks are expected to correspond to positive differences (and half to negative differences), and the sum of these ranks, W, should be the same in value as half of the total rank sum. The alternative (H_A) is that the negative and positive ranks are different.

If the sample size is greater than 10, a normal approximation (z) for the distribution of W can also be used to test for significance. There is a normal deviate (z) equation that may be used to estimate the probability of the difference being real if the sample size >10.

$$z = \frac{W - \frac{n(n-1)}{1}}{\sqrt{\frac{n(n-1)(2n+1)}{24}}}$$

where n is the sample size and the number of pairs. This test can be useful for finding consistent differences in scalar scores from a questionnaire for two different conditions applied to each subject.

The critical values for W are determined by the probabilities for the number of paired observations and can be looked up in a critical values table or found using statistical software and calculators. We use a calculator at the Social Science Statistics website (https://www.socscistatistics.com/).

Wilcoxon Rank-Sum Test (Mann Whitney Test)

This is a nonparametric test for comparing different conditions using independent samples. It compares two samples, the control, and the experimental groups, each drawn from a different population. The null hypothesis (H_0) is that the two are drawn from the same population. The alternative (H_A) is that the populations are different. For this test, the two samples are pooled together, and the observations are ranked from lowest to highest rating. Then the sum of the rankings for each group is calculated. If the two samples are drawn from the same population (H_0), all ranks would be equally likely within each sample. If the two samples are of the same sample size, then the sum of the ranks for group A should be the same as the sum of the ranks for group B. If the two groups have different sample sizes, then the sample size difference must be taken into account. There is a normal deviate (z) equation to estimate the probability of the difference being real:

$$z = \frac{W - \frac{n(n_A + n_B + 1)}{2}}{\sqrt{\frac{n_A n_B (n_A + n_B + 1)}{12}}}$$

where W is the sum of ranks obtained by group A, and n_A and n_B are the respective sample sizes for each group.

An example of an appropriate use of the Wilcoxon rank-sum test might be for comparing two groups of ranked responses from a questionnaire. For example, it might be used if we want to compare comfort rankings for females to comfort rankings for males to determine if one group or the other feels more comfortable.

Proportion Test

A proportion (P) is the number of observations with a particular property divided by the total number of observations. For example, the proportion of people who said they were comfortable is the number of people who said they were comfortable divided by the total number of people in the sample.

Proportions are always numbers less than one and they are used to calculate proportion statistics. Many people are not as comfortable with numbers less than one as they are with percentages. Therefore, the percentage is often used to communicate results. The percentage is the proportion times 100.

The proportion is useful when there is a requirement for a certain percentage of a group or a population to be accommodated. For example, we may be required to accommodate at least 75% of all males ($P \geq 0.75$) and 75% of all females ($P \geq 0.75$) with 90% confidence ($\alpha = 0.10$). The standard normal deviate (z) for $\alpha = 0.10$ is 1.28 if the population is large and the sample size is equal to or greater than 120 (for proportion estimates, a sample size of at least 100 is expected to get a reliable estimate).

If sample observations are random and independent, P is a random variable having either a binomial distribution (sampling from a large population) or a hypergeometric

(small population) distribution. For estimating the confidence interval for P, the normal approximation is used (thanks to the CLT) with mean π, sample size n, and standard deviation:

$$\sigma_p = \sqrt{\pi(1-\pi)/n}$$

To calculate the confidence interval, we use the formula:

$$\pi = P \pm z_{\alpha/2} \sqrt{P(1-P)/n}$$

For $\alpha = 0.10$ this becomes:

$$\pi = P \pm 1.28 \sqrt{P(1-P)/n}$$

If we found that the male accommodation proportion was 88% and the female accommodation proportion was 80% with samples of 120 each, our interval estimates at $\alpha = 0.10$ would be:

$$\text{Male } \pi = 0.88 \pm 1.28 \sqrt{0.88(1-0.88)/120} = 0.88 \pm 1.28\ (0.0296) = 0.88 \pm 0.0378$$

Male π ranges from 0.8422 to 0.9178 with 90% confidence

$$\text{Female } \pi = 0.80 \pm 1.28 \sqrt{0.80(1-0.80)/120} = 0.80 \pm 1.28\ (0.0365) = 0.80 \pm 0.0467$$

Female π ranges from 0.7533 to 0.8467 with 90% confidence.

This indicates that the percentage of males accommodated is estimated to be between 84.22 and 91.78 with 90% confidence and the percentage of females accommodated is estimated to be between 75.33 and 84.67 with 90% confidence. For both males and females, the entire range is above the required 75% so we estimate we have achieved our goals.

It should be noted that the only difference in the male and female equations in this example is the sample proportion accommodated. This difference affected the interval range. As the proportion gets farther from the middle (0.50) the interval range gets smaller. So, the interval for the males with the proportion of 0.88 is only 0.0378 (3.78%) while the range for the females with the proportion of 0.80 is 0.0467 (4.67%). Increasing the number of subjects will also reduce the range.

The proportion test uses the normal distribution and associated parameters, so it is also a parametric test. The biggest limitation of the proportion test is the number of subjects required to achieve reasonable confidence in the estimates. It may have limited utility during early testing when the sample sizes will be less than 100, but it is very useful for the final test when we are trying to determine what sizes and configurations to produce, and we have large samples.

Chi-squared Test for Independence

The chi-squared (denoted X^2) distribution is used in several testing scenarios, but for fit testing its main application is in testing for independence (or conversely for a relationship) between variables. This is similar to correlation, but it is applied to qualitative or categorical variables, whereas correlation applies to continuous variables.

When variables are qualitative characteristics such as discomfort level or categorical such as gender, it is sometimes useful to test to see if the variables are independent. Independence means they are not related or dependent on each other. These are variables that have categories that can be counted and put into a table showing the count for every treatment, called a *contingency table*. The χ^2 distribution is used to test the null hypothesis (H_0) that the two variables are independent, or in other words, have no relationship. It compares the expected frequencies in each row and column to the sample frequencies to determine if they are independent.

For example, if we have two product or product component treatment options and we want to know if there is a relationship between gender and product preference, we might use a questionnaire and a X^2 test. The contingency table might look like the one in Table 4.5.

In this example, the null hypothesis (H_0) is that gender and preference are independent. In other words, there is no relationship between gender and preference. The alternative hypothesis (H_A) is that gender and preference are dependent. In other words, the preference is different, depending on the gender. The test statistic is:

$$X^2 = \sum_i \sum_j \left[\frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}} \right]$$

where n_{ij} is the number of observations in each cell with row i and column j, E_{ij} is the expected value for that cell, and with $df = (r - 1)(c - 1)$, where r = number of rows in the table and c = the number of columns in the table. The expected value E_{ij} for a cell is the row i total times the column j total divided by the total n.

If X^2 exceeds the value from the X^2 table for the level of Type I error (α) then H_0 would be rejected. In the above example, $df = (2 - 1)(3 - 1) = 2$. X^2 is 20.0842. The p-value is 0.000044. The result is significant at $\alpha < 0.05$. Therefore, we would reject H_0 and conclude that preference seems to be dependent on gender.

TABLE 4.5
Evaluation of Relationship Between Gender and Product Preference

	Product Treatment Preference (*j*)			
Gender (*i*)	**Prefer A**	**Prefer B**	**No Preference**	**Total**
Female	12	36	12	60
Male	35	15	10	60
Total	47	51	22	120

Bivariate Correlation

Like the X^2, bivariate correlations also test for a relationship between two variables, but in this instance, it is two continuous or scalar variables. When we use the term correlation, we are referring to the Pearson Correlation Coefficient, *r*, which is a measure of the linear relationship between two sets of continuous or scalar data. In this instance, linear means the straight-line relationship between two sets of data.

Correlations were introduced in Chapter 3, along with the significance test for them. We used correlations to help us narrow down our variables to a few key variables. Correlations are also incorporated into other tests and analyses such as linear regression and discriminant analysis. These analyses also use correlations between more than two variables which is called multiple correlation, and for variables that have been squared, log-transformed, or transformed in other ways. These will be discussed below.

General Linear Models

The category called General Linear Models (GLMs) is the framework for several statistical methods that all follow the same model pattern: dependent variable(s) = independent variable(s) + error. These models share a common type of denotation that serves as a kind of shorthand, so we do not have to spell every variable out every time. We use *Y* and *X* instead of the variable names to simplify expressing the model for analysis. The models are described using a formula that looks like this:

$$Y_j = \beta_0 + \beta_i X_i + \varepsilon$$

where Y_j is the dependent variable(s), β_0 is the intercept (a constant) β_i is the coefficient for the variable X_i which is the independent variable(s) and ε is the error. The coefficient β_i is the part of this model that represents the effect of the *X* variable on *Y*. The error, ε, is the term used to evaluate the likelihood that the effects are real or due to random chance.

It is helpful to write our models in this way when we are trying to formulate the experimental design to answer our questions. We would typically have one model for every question and there are typically many questions each with different outcome variables and many independent variables. It is a lot easier to list the variables as Y_1, Y_2, Y_3, … and X_1, X_2, … than to list all their long names, especially as they become more complex.

The most common type of models for fit testing are the four shown in Table 4.2: ANOVA, Linear Regression, Discriminant Analysis, and Logistic Regression. Some of the differences in these models are shown in Table 4.6. Linear regression is used to predict a continuous outcome variable (*Y*) from either continuous or discrete predictor (*X*) variables, while ANOVA is used to predict a continuous outcome variable from discrete or categorical variables. Discriminant analysis predicts a discrete variable (such as product size) from continuous variables (such as stature and weight). For logistic regression, the outcome variable (*Y*) is dichotomous, or in other words,

TABLE 4.6
Differences Between GLM Methods

	Dependent Variables		Independent Variables
GLM Method	**Type**	**Dichotomous**	**Type**
ANOVA	Continuous	No	Discrete
Linear Regression	Continuous	No	Both types
Discriminant Analysis	Discrete	Sometimes	Continuous
Logistic Regression	Discrete	Yes	Both types

has only two possible outcomes, such as in or out, fit or not fit. GLM also includes combinations of these four methods.

For fit testing, our dependent variables (*Y* variables) are usually fit scores or the best fitting size. The independent variables (*X* variables) are usually the anthropometric and demographic variables. For example, if we tested one size of product and we wanted to know if one or more of five head measurements had an impact on comfort score, comfort score would be *Y* and the five head measurements would be X_1, X_2, X_3, X_4, and X_5. Our model might look like this:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

The β values are called coefficients, and they represent the values that indicate the amount of contribution of the associated variable to the prediction of *Y*. Each item with a coefficient β is called a term. β_0 is a constant that essentially adjusts the scale and is called the intercept. When you use software packages to calculate the models, you may be asked if you want to calculate with or without the intercept. We usually want to run the models with the intercept.

It is common to have two or more independent (*X*) variables. It is also possible to have more than one *Y* (dependent) variable in one model. Using more than one *Y* variable is called multivariate analysis. Usually, we do not do a multivariate analysis unless we expect that *Y* variables do interact with each other and there is not an easier option. Multivariate analysis can be difficult to understand. As a result, it is more common, for fit testing, to evaluate each dependent variable in a separate model or test rather than combining them in one multivariate model or test.

It is also possible to have too many *X* variables. Each *X* variable takes away at least one degree of freedom (*df*) from the error term. This is important because the error term allows us to evaluate the probabilities that our estimates of *X* coefficients and *Y* are real or due to random chance. The *X* coefficients are indicators of each *X* variable's correlation with and contribution to the estimation of *Y*. The *df* for error is in the denominator for error which means the fewer the *df* the larger the error term. A zero *df* for error means it is infinitely large. If the error term is large, then we will not be able to determine the probability that the *Y* predictions are real or due to random chance.

Luckily, there are statistical methods to help us keep the number of variables to a minimum and find combinations of *X* variables that are good predictors or discriminators. Two of these used in fit testing are stepwise regression and stepwise

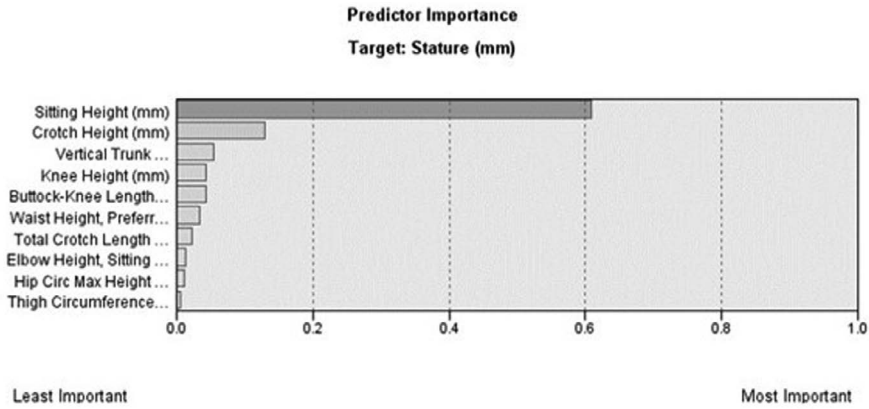


FIGURE 4.3 Example showing predictor importance from a stepwise regression.

discriminant analysis. A stepwise method creates (or fits) a series of models in steps and compares the models.

The forward stepwise regression procedure begins by creating a simple regression of Y with each possible X variable. The candidate X variable that explains the greatest amount of variation in Y is selected. Then three-dimensional regressions are performed using the remaining X variables, one at a time and these are compared. The X candidate that further explains the greatest amount of variation in Y is selected. The addition of X variables one at a time is continued until no further significant Y variation is explained. Figure 4.3 shows the estimated predictor importance from a forward stepwise regression predicting stature from other anthropometric variables.

There is also a backward procedure that is sometimes used. The backward procedure starts with all X variables in the model and drops the one that contributes the least. It is less common to start with the backward procedure because it can require a lot of observations if there are many variables. It is more common to start with the forward stepwise procedure and evaluate the removal of X variables from the smaller set of variables.

Usually, for fit testing, we are not just concerned about finding the best predictive model from a statistical point of view. We also want variables for other practical reasons, such as the ability of the users to take a measurement on themselves. Therefore, it is typically sufficient to just use the forward stepwise procedure to make decisions about which X variables to use.

The stepwise discriminant analysis process is essentially the same as stepwise regression. The difference is in the Y variable which for discriminant analysis is a discrete variable.

The exact same GLM calculations can usually be run in statistical packages in several different places or ways. If there is a GLM feature, nearly all the GLM methods can be run in that section. The good thing is we can run very complex models there. The not-so-good thing is we must specify everything every time. Luckily, some of the easiest and most common procedures will have their own section or drop-down menu. This makes it easier to run because the most common defaults are pre-selected.

In the next section, we describe how to specify the model, run the model in statistical packages and understand the results. We will not provide the details on how the models are calculated.

Analysis of Variance (ANOVA)

Analysis of variance, or ANOVA, is a method for comparing three or more categories. For example, we may want to compare material types and their impact on comfort, when there are three or more types. ANOVA uses the *F*-ratio statistic to evaluate if there are differences.

When there is only one dependent variable (*Y*), it is called a univariate analysis, although the term univariate is rarely used. If there is more than one dependent variable, it is called a multivariate ANOVA or MANOVA. We rarely run MANOVA for wearable evaluations.

If subjects are each only tested in one of the three or more treatments, then the treatments will be listed as one variable and one column in the dataset. This is called an independent measures sample.

When subjects are tested in two or more of the treatments, then a repeated measures ANOVA is appropriate. The data for a repeated measures analysis are usually arranged with a different column containing the score for each treatment. For example, the comfort score for the first material would be in one column, the comfort score for the second material in another, etc. When doing a repeated measures ANOVA, it is necessary to indicate which columns contain the repeated scores.

When there is only one independent variable, it is called a one-way ANOVA. One-way, univariate ANOVAs are the simplest and there are free calculators available on the internet such as the one on the Social Science Statistics website (https://www.socscistatistics.com/).

SPSS has both a GLM section and stand-alone ANOVA sections. The latter is much easier to use. Figure 4.4 shows an example of the pull-down menus to run one-way ANOVAs in SPSS (version 26). They are listed under the Bayesian Statistics tab and there we find both the repeated measures one-way ANOVA and the one-way ANOVA that is not repeated.

When we select the independent variables for ANOVA, we will be asked to select fixed variables and/or random variables. For wearable testing, we typically have fixed independent variables in the ANOVA analyses, because we are usually interested in a fixed number of discrete treatments. For example, if we have three specific types of materials we want to test, and these are not three randomly sampled out of a large population of material types that we want to know about, then the materials variable is called a fixed variable. There are three and only three types we care about. If, on the other hand, we have a continuous range of material properties, perhaps a durometer measurement, for example, and we want to know if there is a relationship between durometer values and comfort and we randomly select durometer values to be tested, then we have a random variable. However, if we are using a durometer and we have a measurement to characterize it that is continuous or scalar, it would probably be better to do a linear regression analysis rather than ANOVA.

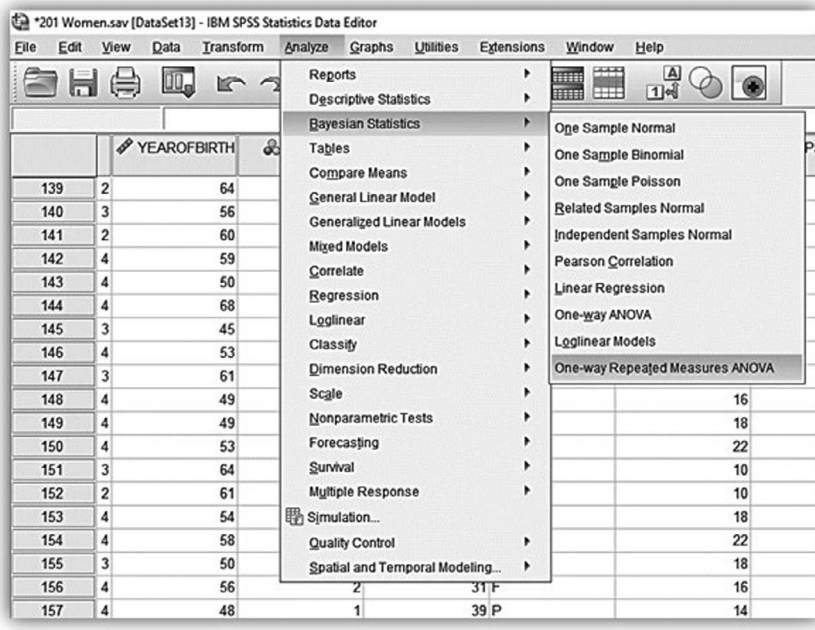


FIGURE 4.4 Running ANOVA in SPSS® (version 26).

Linear Regression/stepwise Linear Regression

Usually, when we refer to linear regression analysis, we are referring to the method of least squares. Least squares regression finds the line where the deviation of points above or below the line is minimized. When there is more than one independent variable, it is called a multiple regression.

Regression equations can be calculated by entering the independent variables all at the one time or by entering them one at a time in a stepwise manner. The latter method is called Stepwise Regression. With Stepwise Regression each independent variable is entered separately to see which has the greatest ability to predict the dependent variable (the first step), then adding each additional variable one at a time to see which, if any, improves the ability to predict (the second step), and so on.

Stepwise regression is used to determine which variables are good predictors of some continuous (scalar or ordinal) dependent variable. It is useful for understanding multiple variable relationships such as the ability of self-reported stature and weight to predict waist circumference. It can also be useful for establishing the multiple variable relationships of anthropometric variables with size-of-best-fit, if the size is scalar or ordinal such that it is in the form of a continuous variable rather than a discrete variable.

To run regressions in SPSS, the GLM procedure can be used, but the simplest approach is to use the stand-alone regression procedure as shown in the pull-down menus in Figures 4.5 and 4.6.

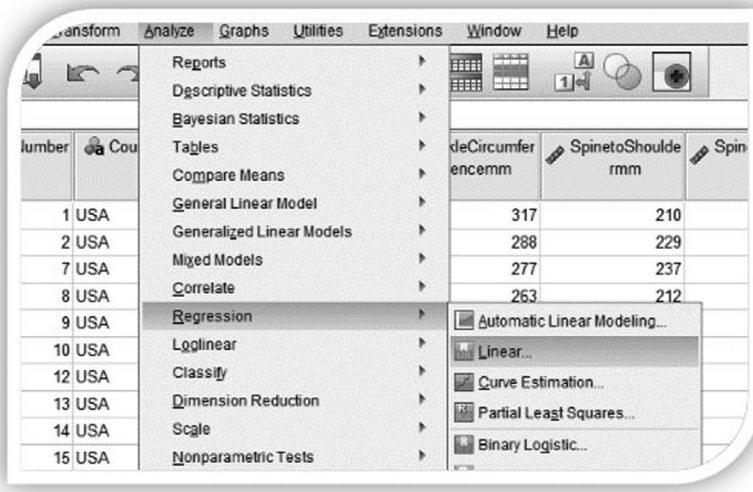


FIGURE 4.5 Running regression in SPSS® (version 26), part a.

Under the Analyze menu we find both GLM and Regression listed. If we select "Regression", we have several treatments. For linear regression, we select Linear. The Binary Logistic option is Logistic Regression which we will discuss below.

When we select Linear, it brings up the menu shown in Figure 4.6. This is where we indicate what our dependent (Y) variables and independent (X variables) are.

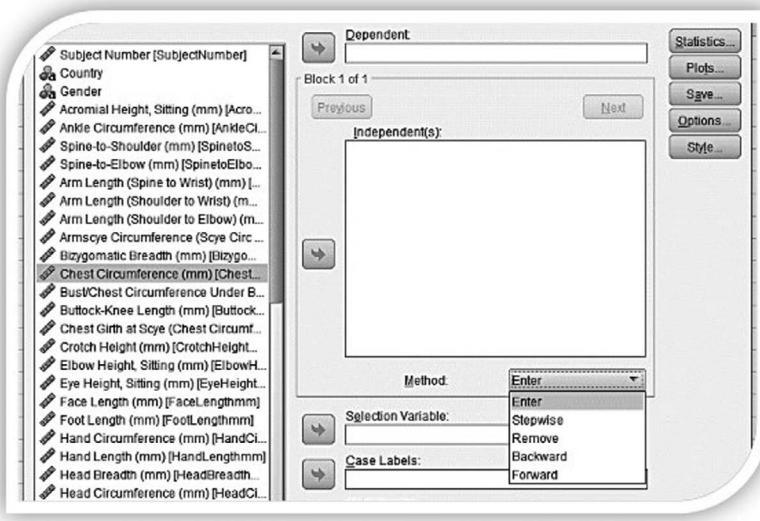


FIGURE 4.6 Running regression in SPSS® (version 26), part b.

We have options about how we have the model enter the X variables. If we select Enter, it just uses them the way we enter them in the list. We use this option if we know exactly what variables we want to use to predict our outcome (Y).

If we select Stepwise, it will enter them in a stepwise fashion. We use this option when we want to know which of the variables are the best predictors of Y. The other options called remove, backward, and forward are alternative stepwise methods that specify the order in which they are entered or removed. We might choose one of these if there are some variables, we specifically want to have entered first.

While it is called linear regression, the regression line is not limited to being a straight line. A dependent variable might be related to X in a curvilinear manner such as the quadratic equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

This equation describes a relationship between X_1 and Y that has the shape of a parabola. For example, if we are predicting the quality of fit in a single mid-range size helmet from an anthropometric variable such as head circumference, the fit might be poor for the smaller heads, then increasingly good toward the center and then poor again toward the larger head circumference. This is a parabolic shape that might be better characterized with a quadratic term in the model.

Discriminant Analysis

Discriminant analysis is much like linear regression except that the dependent variable is a categorical variable such as size-of-best-fit or fit pass/fail. It helps us understand which anthropometric measurements are good for sorting people into sizes or are related to fit issues. It can be used to predict discrete (nominal) variables that have more than two categories but that are not in a linear order.

Discriminant analysis is considered a classification method, so while it can be done under a GLM function, it is often found listed as a classification method. This is the case in SPSS as shown in Figure 4.7. When "Discriminant" is selected from the pull-down menu, we have another menu with several options. We select a grouping variable, which is our Y or dependent variable. In this instance it is ethnicity. We want to see if there are proportional differences between two ethnic groups that we must consider in our design. Note that we have an ethnicity variable that lists each as a number. We must indicate the range of values for the Y variable. We selected 1 as the minimum and 2 as the maximum because these are our two groups. Next, we input the independent or X variables that we want to consider. The classification method, display, and plots we used are the defaults.

There are many output tables including tables that show summary statistics for the groups, the steps that were taken, and variables entered and removed. The two most meaningful tables are shown in Figures 4.8 and 4.9. These are the test of the final discriminant function selected and the coefficients of the function. The first shows that the function is significant at 0.0001 and the second is the function itself. If we multiply each variable times the coefficient associated with it and add them together, we will get the predicted ethnicity.

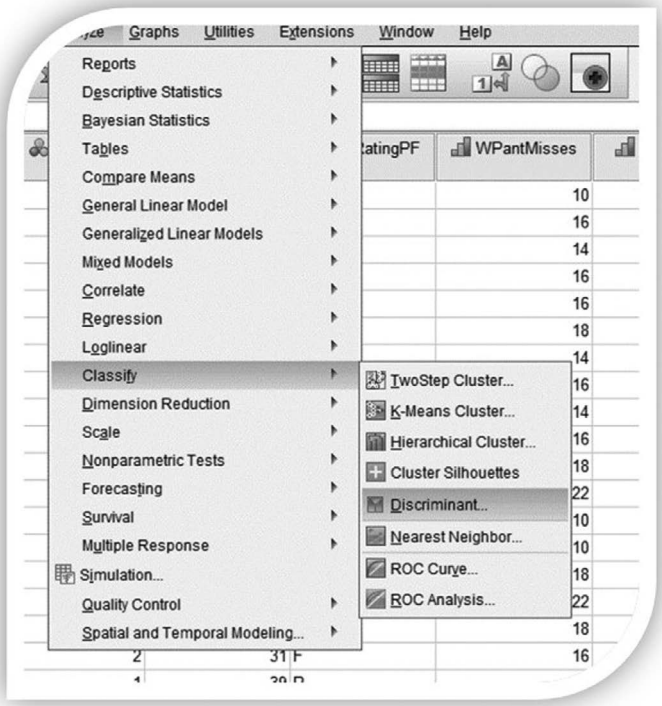


FIGURE 4.7 Discriminant analysis in SPSS® (version 26).

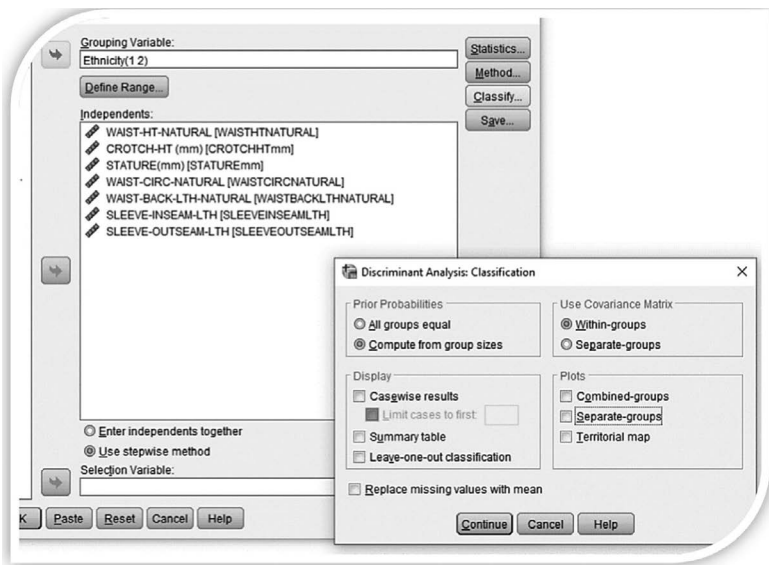


FIGURE 4.8 Discriminant analysis options.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.671	76.356	5	.000

Standardized Canonical Discriminant Function Coefficients

	Function 1
WAIST-HT-NATURAL	-1.452
CROTCH-HT (mm)	1.842
STATURE(mm)	-.741
WAIST-CIRC-NATURAL	.411
SLEEVE-INSEAM-LTH	.750

FIGURE 4.9 Discriminant analysis output.

The coefficients can also be interpreted. In this example, Waist-Ht-Natural is the first variable selected and it has a large negative coefficient. The next variable selected is Crotch-Ht and it has a positive coefficient. This indicates that there seems to be a proportional difference between the two ethnicities. One has long legs, with a short Waist-Ht, and the other has short legs with a tall Waist-Ht. To illustrate, we graphed these two variables and plotted the ethnicity in Figure 4.10. The proportional difference is clear. It means that the Black women in this sample have a shorter lower torso (Waist to Crotch Distance) but longer legs. This will need to be accounted for in lower body garments.

Logistic Regression

Logistic regression is used to predict dependent variables that have just two options (dichotomous or binary), such as pass or fail in a size, or tight or not tight in a particular area such as the waist. To understand it we must first understand probability versus odds.

The chance of being in one category versus being in any category is called the probability (p) of being in the category. It is estimated by the ratio of observations that were in the category over the total number of observations. The odds of being in a category are the ratio of the probability of being in the category (p) to the probability of not being in the category ($1– p$). For example, if a tight fit occurred four out of ten times the probability estimate for a tight fit is 0.40 or 40% and the odds estimate for a tight fit is $0.40/1 – 0.40 = 0.667$ or 66.7%.

Logistic regression uses a logarithm of the odds to predict the likelihood of all kinds of yes/no, in/out, or pass/fail outcomes, such as fit success in a size or fit failure in a size. A logarithm is the inverse function to exponentiation (the function that raises a quantity to a power). The logarithm of a number x in base b is the exponent to which b must be raised to produce x. For example, in our usual base 10 numbering

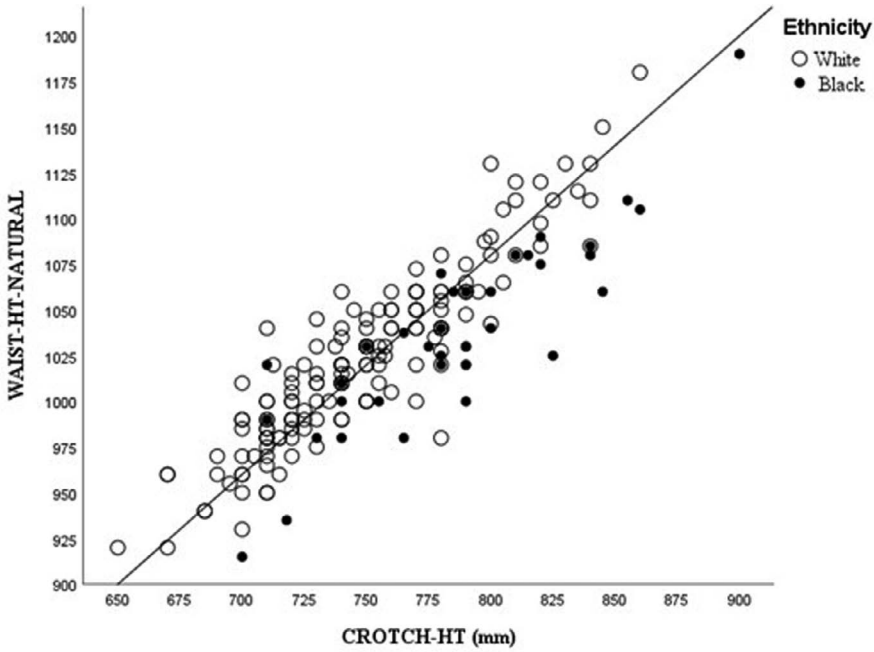


FIGURE 4.10 Waist-Ht by Crotch-Ht bivariate with ethnicity.

system, $8 = 2 \times 2 \times 2 = 2^3$. The exponent for 2 in base 10 to get 8 is 3 and this means the logarithm in base 10 of 8 is 3. This is written $\log_{10}(8) = 3$.

A *base*, in mathematics, is the number of different digits that a system of counting uses to represent numbers. For example, the most common base used today is the decimal system, which uses numbers 0 to 9 to count. Computers use a binary base, 0 and 1. Logistic regression uses a base called e, because of its special properties. The number e is called Euler's number after the Swiss mathematician Leonhard Euler and is approximately 2.718. When this base is used it is called the natural log, so it is sometimes denoted Ln rather than Log or Log_e.

Logit is the term for the natural log of the odds, $\text{Ln}(p/1 - p)$. It is used to transform the dependent variable (Y) for logistic regression. For simplicity of calculations the two options for Y are entered as 1 and 0, with 1 representing being in the category and 0 representing not being in the category. Then the probability (p) is the probability that $Y = 1$. The logistic regression equation becomes $\text{Ln}(p/1 - p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$.

Basic logistic regression is also called binary logistic regression or binomial logistic regression and we show where to find it in SPSS (version 26) under the Regression function in Figure 4.5. It is like linear regression in some ways. Some of the differences are listed in Table 4.7.

Multinomial logistic regression (MLR) generalizes binomial logistic regression to more than two possible discrete outcomes. For example, if we have one question asking if the pant was tight, good, or loose, there are three binary variables (0 or 1 responses) resulting from the question: (1) is the waist tight yes or no, (2) is the waist

TABLE 4.7
Differences Between Linear Regression and Binary Logistic Regression

	Linear Regression	Binary Logistic Regression
Predicted/Response/ Dependent Variable	Continuous, scalar, or ordinal typically with more than two values. For example, Stature, Likert scale comfort score –5 to +5.	Categorical or nominal with only two values (dichotomous). For example, pass or fail, win or lose, in or out.
Equation format	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$	$\text{Ln}(p/1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$
Equation fit method	Least squares	Maximum Likelihood Estimation
Output values	Measurement units such as cm, years, dollars	Probabilities less than 1 such as 0.41, 0.25, and 0.99

loose yes or no, and (3) is the waist good yes or no. When we answer we can only say yes, (1), in one of the three so it is like three binomial questions in one. This type of question would the dependent variable in an MLR equation.

Having a question with three or more discrete answer categories can provide some additional information beyond just pass or fail. For example, in addition to predicting the best fit size, it can also give a wearer an indication of the type of fit to expect in a size. An example of this is shown in Figure 4.11.

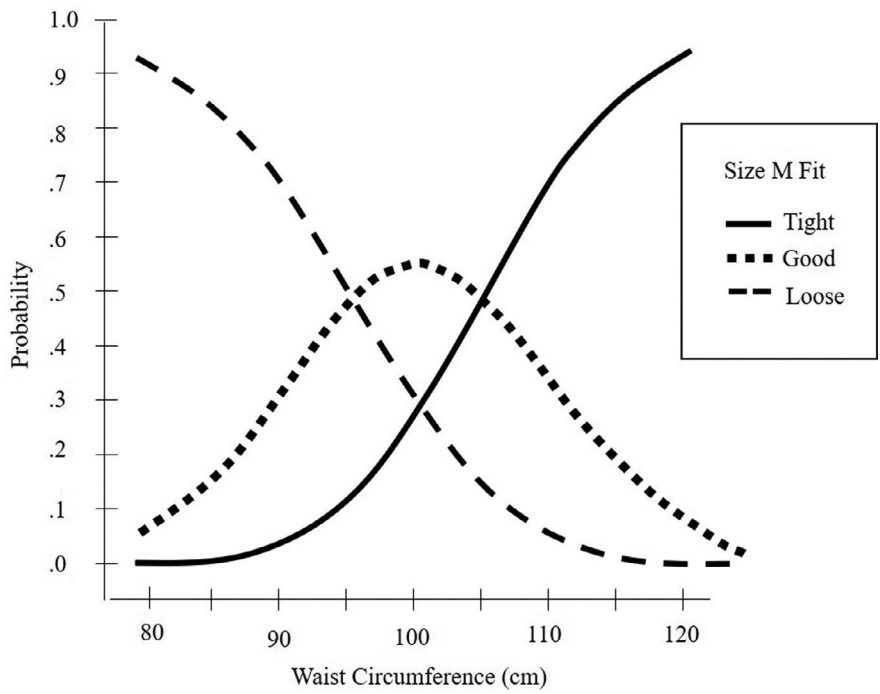


FIGURE 4.11 Example of fit estimation for hypothetical size M.

A size M pant was scored by an expert fitter for tight, good, or loose for each subject and we also measured their waist circumference. We predicted the pant tightness/looseness (dependent variable) from waist circumference (independent variable) in an MLR equation. This gives us average probability curves for each possible outcome. Here we see that if the subject's waist circumference is 100 cm, they have nearly a 0.6 or 60% probability of a good fit in the waist, with less than a 30% probability of it being either tight or loose, whereas, if they have a waist circumference of 120 cm, they have more than a 90% chance of it being tight. We can also see that at 95 cm, the chance of being good or loose is about even, and at 105 cm the chance of being good or tight is about even. This gives us some additional information that we can use to decide if we want the size or not. For example, at 95 cm it might be a little loose on us, but perhaps we like it that way.

The probability curves shown in this example are indications of the average probability curves when the independent variable (Waist Circumference) was correlated with the fit probability. If there is not a strong correlation these values might not be better than random chance. We look at the significance of the independent variables' coefficients to determine if they significantly influence the probability or predict fit.

Another way to examine how well a logistic regression performs, we use a function called the receiver operating characteristic (ROC) curve. This curve uses two metrics, sensitivity, and specificity. Sensitivity is the probability that the model predicts an outcome for an observation that is positive when the outcome is positive. It is also called the true positive rate (TPR). Specificity is the probability that the model predicts an outcome that is negative when the outcome is negative. The false positive rate (FPR) is calculated as 1 – specificity. The ROC plots the TPR against the FPR.

The ROC curve is a classification function, so it is often found in software tools under classification functions. The ROC menu can be seen in Figure 4.7 underneath the discriminant function.

DESIGN LOOP TESTS AND ANALYSIS

Design loop test outcomes are intended to either influence the product design or influence the test procedures. Pilot tests are pre-experiment guidance tests that evaluate our test procedures. They ensure that tools, processes, and the COF are reasonable and effective. The other two types of tests help us verify that our design concept is reasonable and help us choose between design treatments. Stand-alone trade studies compare a small number of treatments to help us decide between them, and prototype fit tests evaluate the entire human/product system.

We present the design loop test types in the order of simplest and easiest to most complex and challenging namely: pilot tests, stand-alone trade studies, and prototype fit tests. This is not necessarily the order in which they will be done. Tests are done when a decision is needed. For example, sometimes we find a problem during a prototype fit test and we may skip subsequent steps and go right to fixing the product. After we fix the problem, we may not need to repeat a pilot test and might skip right to another prototype fit test. Also, there are times when we discover a material or component, we planned to use in our product and that we tested in a prototype fit

test, is no longer available, so we need to check the efficacy of a new component we want to trade out. In this instance, perhaps only a stand-alone trade study may be needed. As with all testing, the purpose is to make better informed decisions to save time, money, and resources and have satisfied customers.

Pilot Tests

"Measure twice and cut once" is an old carpentry proverb indicating it is a good idea to double-check your measurements and your plan before you cut the wood, otherwise you may have to cut again and waste time and money. Just like with carpentry, the purpose of pilot testing is to save time and money by avoiding having to repeat tests or re-do the product. Pilot tests verify that our tools and COF are giving us the information we need within the time and cost constraints that we have. They guide (i.e., pilot) our subsequent fit tests. They may be done in conjunction with other tests.

Pilot tests can occur at two points in our design process. This is not to say that two pilot tests are required. For products that already exist on the market, or products that are like products that have been on the market, the process may skip most of the design loop and start just before the sizing loop. In this instance, the pilot test needed may be the one to verify the final COF and procedures.

The risks associated with not doing a sufficient pilot test include:

- None of our measurements help determine fit or why something is not fitting
- We waste time taking measurements that do not tell us anything useful
- We measure the wrong fit or prioritize the wrong aspects of fit
- We get confusing results due to differences in the people doing the measuring or assessments
- We have wasted time and must redo our fit tests
- We have a test procedure that takes too long so we cannot meet our deadline

Prior to pilot testing the team members should have been suitably trained on how to measure and record, as we discussed in Chapter 2, and the tools being used have been validated prior to selection. Therefore, before the start of the pilot test, the data collection team should be familiar with all procedures.

The pilot test consists of a complete run-through of the full-fit test multiple times. This is like a practice run and each run-through is timed to determine if there are any bottlenecks or places that can be streamlined. This is usually done with 3–5 subjects and each subject should be repeated to see if the results are the same each time. This can occur within one day but for subject repeats it is good to have the repeat on a subsequent day.

The number of subjects is a subjective judgment call based on any issues that appear and any changes we may make. The purpose is to reassure us that the test will run smoothly without the need for changes. If after running through the procedures a few times we are confident, then we do not need any more subjects. It is common to run through the procedures using ourselves as subjects first and we might modify the procedures and repeat on ourselves several times. Then, we might run through the procedure with convenient subjects of varying body proportions.

Running through the procedures with a few varied subjects enables us to evaluate the COF, the test protocol order, the way we ask questions to ensure consistency, the timing, the need for additional items and more. While the repeated data can be analyzed with statistics, this is typically not necessary, assuming the tools have already been validated. For these types of tests, we are not evaluating the TP, so we do not need a random sample of the TP and a convenience sample of 3–5 subjects is typically sufficient.

It is important to have some variety in the small number of subjects used. For example, if the product is a pair of pants, make sure to include both small and large waisted subjects in the pilot test. If the product is something to be worn on the face over the nose, have a variety of nose sizes and shapes represented in the subjects used. This will verify that the test procedures will work for most types. If the procedures work for some body types but not others, then we introduce bias into our data and that will make it difficult to sort out the sizes and fit. If we check all these things and verify that they work for all types ahead of time, we will save time and money during the analysis phase and reduce the chance that we will have to re-do tests.

Stand-alone Trade Studies

Trade studies are comparisons of a small number of treatments, such as different shapes, paddings, adjustment mechanisms, environmental conditions, or user tasks for the purpose of narrowing down design treatments or verifying design changes are effective and do not cause issues elsewhere. Trade studies can be done at many different stages of development. If, for example, it is found that the supplier for a particular material is not available and a substitute material must be used, it may be necessary to do a quick trade study to find a suitable replacement. These treatment options are categorical, such as treatment A versus treatment B. The outcome or effect we are measuring is usually scalar or ordinal, such as the comfort score from 1 to 10, or the distance from the optimal location.

Trade studies can be done as a stand-alone test or can be done as part of a larger prototype fit test. Doing a stand-alone trade study can enable us to drop some of the treatments before doing a larger more complex test. If a stand-alone trade-study result indicates that a change to the product is needed, the change may be made to the product before any further testing or analysis is done. This means it will have a reduced design loop.

This section on trade studies addresses stand-alone trade studies that have just one independent variable (called a one-way analysis) and one dependent variable (called a univariate analysis). More complex trade studies will need to be done as part of a larger prototype test. Therefore, they will be discussed in the prototype fit test below.

Trade studies often use a convenience sample, and this is a sample of company employees or close friends that have been sworn to secrecy (signed a non-disclosure agreement). The results are intellectual property that companies typically want to protect so recruiting subjects from the TP randomly is not advisable. A convenience sample has an added risk of bias, both in the assortment of subjects used as well as in their opinions about the fit of the product. Company employees may be more likely

to be either positive or more critical of the product. If we keep these risks in mind, the results can still give us valuable information to help us make better decisions.

With stand-alone trade-study tests, there are only one or two questions per test. Typically, we want a quick and inexpensive answer just to narrow down options and rule out options that are clearly poor choices. We are willing to accept a lot of risk and do not want to spend a lot of time or money. Often, we know we will have more tests in the future.

For quick stand-alone trade tests that have more than one fit variable, it is usually best to reduce our fit score to one variable, such as an overall pass versus fail score. Anything more complex will require more time and money. This is especially true if there will be prototype or full-fit tests to follow that will delve into the details of fit.

There are two basic trade-study questions for wearable products:

1. Which material or component is best?
2. Is one shape, size, or design concept better than another?

The first question may require a subjective opinion, but it may not need a representative sample if, for the item being tested, size, shape, and demographics will not affect the outcome. The second question assumes size, shape, and demographics will affect the outcome, so a more representative or varied sample is needed. The purpose of this test is to make a choice about which design concept to move forward with.

The simplest trade-study test is a comparison of two options (two treatments), but more than two can also be compared. If there are just two treatments, the analysis consists of a test to see if they are different. The candidate tests are:

- Student's *t*-Test
- Paired *t*-Test
- Proportion Test
- Wilcoxon Signed-Rank Test
- Wilcoxon Rank-Sum Test

If there are more than two treatments, the analysis begins with an ANOVA to test if any of the treatments are different from the others, and if so, then an additional multiple comparisons test (such as Tukey's or Bonferroni's Test), is done to see which ones are best.

The tests can be done using either an independent design or a repeated measures design. An independent design consists of using different participants for each treatment. This is often the case when we are comparing new treatments to treatments we tested previously. A repeated measures design consists of testing the same individuals on two or more treatments. This is common for small trade studies when all treatments are available and can be randomly assigned because it has more statistical power for finding differences with small samples of subjects. When there are just two treatments, the repeated measures design is called a paired design.

For material or component trade studies it is sometimes reasonable to assume that demographics and anthropometry will not affect the outcome, particularly if we are using a repeated measures or paired test sample. Subjects will either find one better than the other or not different regardless of their body size, shape, or demographic.

If this assumption is made, it would not be necessary to have a sample representing all the different kinds of size and shape variability, so the sample size can be small, and the test can be very quick and simple. Here are some examples.

Comparison of Two Treatments Using Paired Test Design

We found we did not have a supplier for the type of padding we were using for our wearable, so we want to test a new type to see if it is as comfortable or better than the old option (in other words, it is not less comfortable). Since we only want to know if it is as good or better this will be a one-sided test. We selected 10 subjects and we had each subject try both treatments and rate comfort on a scale of 1–10, with the order of the treatments randomized. For purposes of illustration, we analyzed the results twice, once using the Wilcoxon signed-rank nonparametric test and once using the paired *t*-test, a parametric test.

The Wilcoxon signed-rank test results are shown in Table 4.8. The comfort score for the first treatment (T1) was subtracted from the score for the second treatment (T2) to arrive at the difference (Diff.). Then the sign of the difference was recorded, as was the absolute value of the difference. The absolute values were ranked, and the sign was applied to the rankings to arrive at the signed rank. The sum of the positive ranks (W_+) was calculated, as was the sum of the negative ranks (W_-). The smaller of these two numbers is *W*. *W* was compared to the critical values for the signed rank to determine significance. W was smaller than the critical value; therefore, it is significant. Here, we see that the T2 scored lower for all but one person and the difference between the options was significant for the one-sided test with $p < .05$. This means that the probability that T2 is better than or equal to T1 is small, T2 appears to be less comfortable.

The critical values for *W* are determined by the probabilities for the number of paired observations and can be looked up in a critical values table or found using statistical software and calculators. We used a calculator at the Social Science Statistics website (https://www.socscistatistics.com/).

TABLE 4.8
Example of Wilcoxon Signed-Rank Test

Subject	T2 Score	T1 Score	Diff.	Sign	Abs	Rank	Signed Rank
1	2	7	–5	–1	5	7	–7
2	1	10	–9	–1	9	10	–10
3	3	9	–6	–1	6	8	–8
4	7	8	–1	–1	1	1.5	–1.5
5	2	10	–8	–1	8	9	–9
6	5	7	–2	–1	2	3	–3
7	7	6	1	1	1	1.5	1.5
8	3	7	–4	–1	4	5	–5
9	4	8	–4	–1	4	5	–5
10	3	7	–4	–1	4	5	–5

Source: $W_+ = 1.5$, $W_- = 53.5$, $W = 1.5$. Critical value for *W* at $n = 10$ ($p < .05$) for one-sided test is 10.

The Wilcoxon signed-rank test is the nonparametric method (does not assume a distribution) and it tests to see if the medians of the signed ranks are different. An alternative method for this example is the paired *t*-test which assumes a normal distribution of the means (because of the CLT) and tests to see if the means of the treatment scores are different.

The results for this method are shown in Table 4.9. The difference between the scores for the two treatments is calculated (Diff.) and the mean of the differences is calculated. Then, the deviation (Dev.) is calculated by subtracting the mean from each difference value. This value is squared (Dev.^{2}), and the sum of these squared values, called the sum of squares or SS, is calculated. The SS is used to calculate the Standard Error (Std. Err.), the Standard Error of the Mean (Std. Err. Mean), and along with the Mean, the Paired *t* score. Again, we find that the difference between T2 and T1 is significant with T2 scoring lower.

Which test is better? Both are easy to compute, but the signed-rank test has fewer assumptions. In addition, if the samples or populations are skewed, it may be better to test the medians rather than the means which makes the signed-rank test better as well. However, if the distribution is reasonably centered and normal, the *t*-test can be more powerful, and most people will be more familiar with the *t*-test, so communicating the results to the clients may be easier.

We only had ten subjects in this example, yet we were able to find significance. In this instance, ten subjects seem to be sufficient. However, if the differences between

TABLE 4.9
Example of Paired *t*-Test

Subject	T2 Score	T1 Score	Diff.	Dev. (Diff. – Mean)	Dev.^{2}
1	2	7	–5	–0.8	0.64
2	1	10	–9	–4.8	23.04
3	3	9	–6	–1.8	3.24
4	7	8	–1	3.2	10.24
5	2	10	–8	–3.8	14.44
6	5	7	–2	2.2	4.84
7	7	6	1	5.2	27.04
8	3	7	–4	0.2	0.04
9	4	8	–4	0.2	0.04
10	3	7	–4	0.2	0.04
Mean			–4.2		
Sum Dev.^{2} (SS)					83.6
Std. Err. (SS/*df*)			9.29		
Std. Err. Mean (Std. Err./N)			0.93		
SQRT SE Mean			0.96		
Paired *t* Score = (Mean/SQRT SE Mean)			–4.36		

Source: The value of *t* is –4.36. The value of *p* is .00091. The result is significant at $p < .05$.

the two treatments were more subtle, ten subjects might not be sufficient. If we do not find a significant difference, there can be several reasons, including: (1) our subjects were biased, (2) our sample was too small to detect the difference, or (3) there is no substantial difference. There are two reasons we might want to increase the sample size: (1) to have confidence that adequate variability is represented in our sample and (2) to be able to detect a true small difference if it exists. More subjects provide more power to find differences. If it is important to have more confidence that T2 is better before we choose to use it (perhaps T2 is more expensive than T1), and even a small improvement would be worthwhile, then adding subjects is a good idea.

Comparison of Two Treatments Using Independent Samples

In the example above, we had the ability to test both treatments on each subject. If we had a situation where it is not practical to have each person wear each treatment, so each person only tests one treatment, the samples for each treatment are called independent samples. This requires a different analysis because we no longer can subtract scores for each subject. For example, if we need to have the item worn for many hours before assessment, we may not want to run a test on a second treatment in the same day and it may not be feasible to have the subjects return for a second day.

There are both nonparametric and parametric tests for independent samples as well. For this example, we analyzed independent sample data using the nonparametric Wilcoxon rank-sum test (also called Mann-Whitney U Test) and the parametric Student's *t*-test. The Wilcoxon rank-sum test tests the medians. The Student's *t*-test assumes normality as well as equal variances in the two samples and tests the means.

We again want to test a new type of padding to see if it is as comfortable as or better than the old treatment (in other words, it is not less comfortable). We had a previous test of comfort for T1 that used ten subjects. We collected a new sample of ten different subjects to try T2 and rate comfort on a scale of 1–10. Example results for both are shown in Table 4.10.

Here we see that both tests found the treatments to be significantly different at $\alpha = 0.05$ with T2 having the higher comfort scores. Both tests use *p*-values, which indicate the probability of not being different. Smaller *p*-values indicate a higher probability that they are different.

In this example, the Student's *t*-test has a smaller *p*-value than the Wilcoxon rank-sum test and indicates a significance at $\alpha = 0.01$. If the sample is close to the normal distribution, as in this example, then the Student's *t*-test can be more powerful and more likely to find a true difference than the rank-sum test. However, if the data are skewed or we don't want to assume a normal distribution, the nonparametric rank-sum test can be more powerful.

Comparison of Three or More Treatments Using Repeated Measures Design

We have three types of candidate padding materials and would like to know which one is most comfortable. We selected ten subjects and we had each subject try every treatment (in random order) and rate comfort on a scale of 1–10. This

TABLE 4.10
Two Hypothesis Tests for Two Sample Comparison

Treatment 1	Statistic	Subject	Score	Treatment 2	Statistic	Subject	Score
		1	4			11	7
		2	3			12	10
		3	6			13	9
		4	9			14	8
		5	4			15	10
		6	7			16	7
		7	9			17	6
		8	5			18	7
		9	6			19	8
		10	2			20	7
	Mean		5.5		**Mean**		7.9
	STD		2.37		**STD**		1.37

Wilcoxon Rank-Sum Test (One-Tailed)	
***U*-Value**	19
Critical *U* at α = 0.05	27*
***z*-Value**	–2.30558
***p*-Value**	.01044*
Student's *t*-Test	
***t*-value**	2.77334
***p*-value**	.006266*

* Significant at α = 0.05.

question is a two-part question: (1) are the treatments different, and (2) if different, which treatments are better? The first part is answered with an ANOVA, and the second part, with a multiple comparisons test. If the ANOVA does not find any types that are significantly different from each other, then there is no need to do any comparison tests.

When there is only one independent variable, in this instance, the Treatment Number, it is called a one-way ANOVA. Since each subject has repeated the assessment on each option, it is called a one-way ANOVA with repeated measures.

In this example, there is only one dependent variable, the comfort score. When there is only one dependent variable, it is called a univariate analysis. Some software packages will drop the term univariate when they list the type of analysis.

One-way, univariate ANOVAs are the simplest and there are free calculators available on the internet. This can be run in a statistical package, but for this example, we again used the one at the Social Science Statistics website (https://www.socscistatistics.com/) for the data shown in Table 4.11. The data for a repeated measures analysis are usually arranged with a different column containing the score for each option as shown.

TABLE 4.11
Example One-Way ANOVA With Repeated Measures

Subject	T1 Score	T2 Score	T3 Score
1	1	8	9
2	10	1	8
3	8	3	7
4	7	2	9
5	1	2	6
6	5	5	10
7	7	10	10
8	10	8	9
9	9	1	9
10	10	7	10

Source: The *F*-ratio value is 5.546. The *p*-value is .01329. The result is significant at $\alpha < 0.05$.

ANOVA uses the *F*-ratio statistic to evaluate if there are differences. In this example, we see that the *F*-ratio value is 5.546 which is significant at $\alpha < 0.05$. This indicates that at least one of the three options appears to be different from the others. Since we found that there was a significant difference with the ANOVA, we did an additional analysis using SPSS (version 26) software to examine the differences between the mean comfort scores and plotted the means with the associated error bars (see Figure 4.12). Here we see that Treatment 3 appears to be more comfortable than Treatment 2 because the error bars do not overlap. Treatment 2 is not significantly better than Treatment 1.

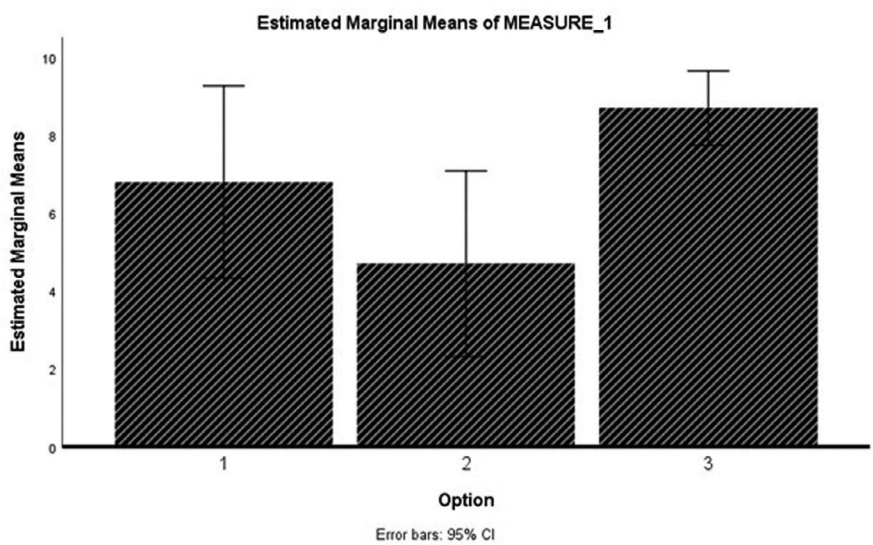


FIGURE 4.12 Repeated measures plot of option means with error bars.

Comparison of Three or More Treatments Using Independent Measures Design

This example is the same as the previous example, except that each subject is only tested in one of the treatments. In other words, the sample for each treatment is different or independent of the others. The one-way ANOVA when the subjects only test one treatment each is called a one-way ANOVA with independent measures, which is the same as saying ANOVA with one independent variable with independent measures. This confuses some people because the word independent is used in two places. We mention it so that the correct analysis can be found regardless of the software used.

We have three candidate padding materials and would like to know which one is most comfortable. We selected ten subjects and each subject evaluated one treatment (randomly assigned) and rated comfort on a scale of 1–10.

With an ANOVA with independent measures, the data will need to be arranged as shown in Table 4.12, part a. We sorted the data by treatment for ease of entry into some of the calculators. Each subject has only one comfort score and that is for only one treatment. We did this analysis in SPSS, and ANOVA is found under General Linear Model. The steps used to calculate this ANOVA are shown in Figure 4.13. Our treatments are fixed variables or factors. This means we only want to know about this set of treatments.

When ANOVA indicates there is a significant effect or difference in treatments, we next want to know which ones were different. That requires an additional test called a multiple comparisons test. This test can be selected when we run the ANOVA. We chose Tukey's HSD test. The ANOVA results are shown in Table 4.12, part b. Since the F-ratio was 4.82976 and was significant at $\alpha = 0.05$, it is appropriate to examine the multiple comparison results shown in part c. Here, the three treatments are labeled T1 through T3. Tukey's pairwise comparisons indicate that T2 and T3 are significantly different, but the other pairs are not.

Prototype Fit Tests

Prototype fit tests evaluate the fit and performance of the entire product-human system using mock-ups or prototypes of the product. That means body size and shape are necessary factors, and both the experimental design and the analysis are more complex than stand-alone trade studies.

A mock-up is a prototype, but it is one that is not fully functional and/or not made from the intended materials. We make this distinction to emphasize that a lot can be done before a fully functioning prototype is available, and it is usually less costly to make design changes. When we progress to the fully functional products made from the intended materials, we simply call them prototypes. The test procedures are the same regardless of whether the test uses a mock-up or a fully functional prototype.

A prototype fit test is usually a combination of many tests or experiments corresponding to the different aspects of fit and the many questions we need to answer. One prototype system test might answer a series of questions such as:

- How well does the system perform?
- What percentage of the TP fits in a size?

- Is there a large group of people for whom no size is available?
- Is there a fit issue that affects all sizes?
- Is there a population subgroup that has a specific fit issue?
- How much size overlap or redundancy is there?
- What are the body areas with fit issues?
- What body measurements are good fit predictors for one size?
- What body measurements are good predictors of the size of best fit?

TABLE 4.12
Example One-Way ANOVA with Independent Measures

a. Raw Data

Subject	Score	Treatment
2	10	1
6	5	1
8	10	1
10	10	1
12	1	1
13	7	1
17	1	1
20	7	1
26	9	1
30	8	1
4	1	2
5	3	2
9	7	2
11	8	2
14	2	2
16	5	2
18	8	2
19	1	2
23	2	2
29	10	2
1	10	3
3	9	3
7	10	3
15	7	3
21	9	3
22	8	3
24	9	3
25	6	3
27	10	3
28	9	3

b. ANOVA Result

The F-ratio value is 4.82976.
The p-value is .016102.
The result is significant at $p < .05$.

c. Tukey HSD Pairwise Comparisons

Treatment (T)		HSD.05 = 3.192	Q.05 = 3.5064
Means (M)		HSD.01 = 4.092	Q.01 = 4.4948
T1:T2	M1 = 6.80	2.1	Q = 2.31
	M2 = 4.70		(p = .25024)
T1:T3	M1 = 6.80	1.9	Q = 2.09
	M3 = 8.70		(p = .31813)
T2:T3	M2 = 4.70	4.0*	Q = 4.39
	M3 = 8.70		(p = .01189

* Significant at α = 0.05.

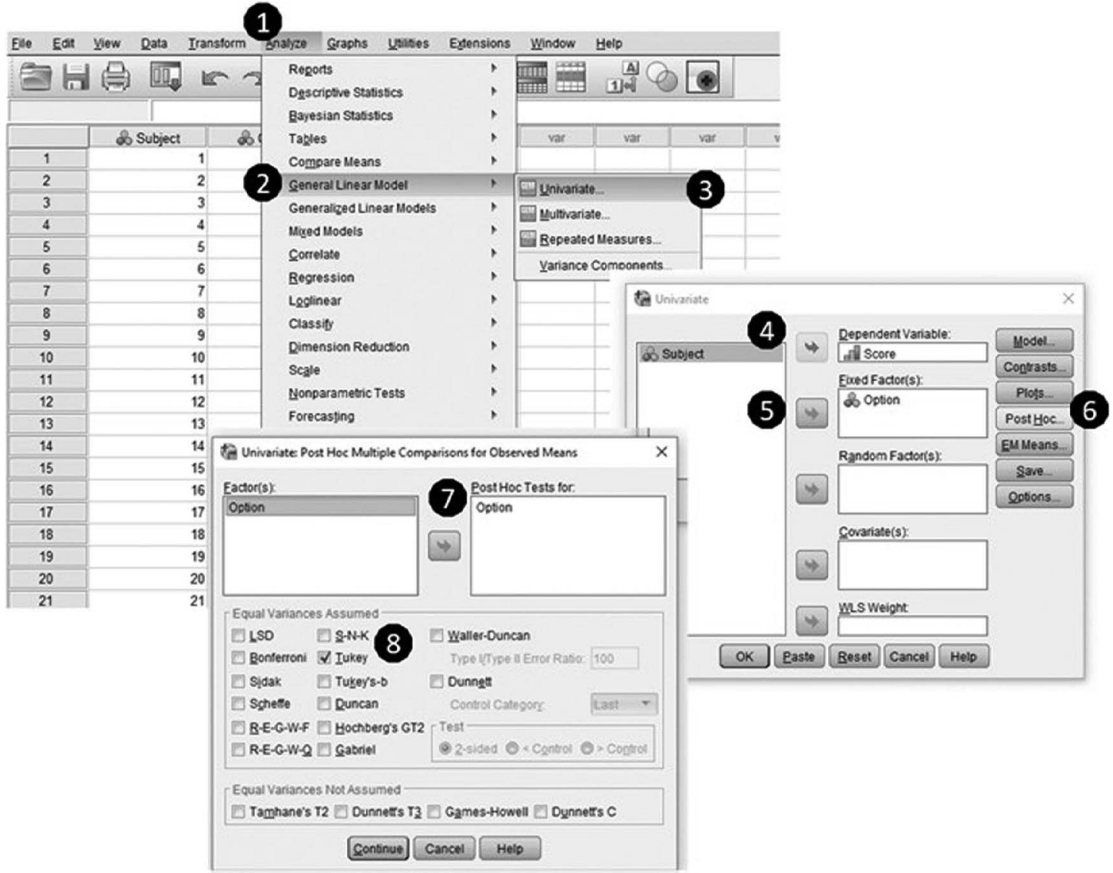


FIGURE 4.13 ANOVA steps in SPSS® (version 26).

The design loop is iterative, and the prototype fit test purpose and experimental design varies depending on where we are in the process, or where we are beginning the testing process. For example, if we are developing a new product that has no predecessors, we might start with a one size prototype to learn if we need more than one size. This test will have just one treatment, the size, so treatments will not need to be randomized. If we are developing a product that is a modification or a new design for a product that is similar to previous products, such as a new type of pant, we might want to use the prior product to check the sizing or we might have a fit standard and want to verify that the sizing meets the standard or is consistent with the prior products. Therefore, this prototype may come in several sizes, and we will need to determine how to randomize or assign treatments.

There are several good experimental designs for prototype testing. The important factor in making a choice is having enough information or statistical power to make informed decisions within an acceptable risk range. For example, we might start with just the base size and test 60 randomly selected subjects. If, during analysis, we find subjects from a particular demographic are having issues, we might add more in a later study from that group to understand the issue. This is a good strategy if the product is still developing (not mature) and changes made to it might resolve the problem in the meantime.

In the design loop, we do not always need a full TP sample. Sometimes a smaller fit test sample will suffice to determine if we need to make changes to the base size or size range. We will need enough subjects to do a range-of-fit analysis for at least one size. Range-of-fit analysis and the subsequent fit mapping require that we either have: (1) subjects who fit and subjects who didn't fit in at least one size or (2) we have subjects from the full range of the TP and everyone achieved a good fit in one size.

Fit failures can occur for different reasons for different groups of people, and this needs to be considered when designing the experiment. For example, females have wider hips than males, but narrower shoulders. If we are designing a protective vest for both genders, we may want to evaluate the fit in the hip area and shoulder area separately and we will want to include a large enough sample of each gender in case we need to evaluate the range of fit separately as well.

We want at least 30 subjects who achieve a good fit, and enough subjects who do not fit, to understand how much size variation can be accommodated in one size. Therefore, we expect a minimum sample size for a given prototype fit test to be about 60 subjects.

Alternatively, we might start with 30 subjects and add subjects as we learn what types of subjects we need. For example, we may find that we have a clear definition of the upper end of the size, but not of the lower, so we would need to add small subjects to our sample and stop when we have enough to define the range of fit within a size.

We might start with a plan for 200–300 subjects, test them in all available sizes in random order, and stop early if we find that we have enough statistical power for all population segments. This requires doing the analysis as we go, rather than waiting until the end, but that is a good practice anyway. It is more common that we start with a plan for 200–300 random subjects and determine we need more subjects of a particular size. Sometimes, if there are enough subjects, this design loop test can serve as the first sizing loop test as well.

The number of subjects also depends on how many sizes are expected or tested, and how much is already known about sizes. If we were testing a product with precedents and using a full range of sizes from an earlier product, we would need more subjects than if we were testing a new product with just one size developed.

To begin a prototype system test we start with the purpose of the test, and the list of questions we would like to answer for that purpose. Then we design the experiment to answer the questions and test the experimental design and the COF with a pilot test. There may be an infinite number of scenarios for prototype systems testing. To explain the process, we walk through the experimental design and analysis for two scenarios:

1. Scenario 1: First test iteration with just one size
2. Scenario 2: Existing product with fit and sizing issues in multiple sizes

Scenario 1: First Prototype Test Iteration With Just One Size

Our prototype was a helmet with one size to test, the base size. There were two copies made of the base size to allow for two subjects to be tested at the same time. The helmets were randomly assigned to the subjects. Both were measured to ensure they met the specification and were indeed the same.

The TP included North American men and women ages 18–35. This was the first fit test of the product intended to determine if it needed modification before finalizing the sizing, therefore the sample need only include people who fit and people who do not fit. We did not need a full TP sample. Because the TP includes both men and women, we included at least 30 of each gender and selected 75 subjects total, 38 women and 37 men.

The fit test team consisted of three people: (1) a greeter/scheduler, (2) a helmet fitter and assessor, and (3) a measurer. The subject's start times were staggered by having one subject start by being measured, and the other start by being fitted in the helmet. The two subjects were assessed in two different rooms, so they did not see each other in the helmet or have a chance to discuss the helmet until after the test was completed.

The specific questions we wanted to answer included:

1. What body measurements best predict fit (key variables)?
2. What is the range of fit in one size?
3. Who are we fitting and who are we not fitting?
4. Is the size well-placed or does it need to be moved?
5. If we need more sizes, what should they be?
6. Where are there any general fit issues that need to be fixed?

The COF required the product to be simultaneously comfortable and stable (not slip on the head when moving) for at least an hour. After a pilot test, it was decided that the helmet would be comfortable if was loose enough, and it would be stable if it was tight enough. Therefore, the happy medium was a good fit for both.

Pilot testing also indicated that having the subject wear the helmet for 30 minutes, in an indoor environment at room temperature would suffice. During that time, they

were required to look around the room and go through a series of movements. It also revealed that two subjects could be processed every hour, so 16 subjects could be processed in an eight-hour workday, and potentially the data collection could be completed in one week.

Comfort was evaluated using a five-point scale of 1–5 with 1 being extremely uncomfortable and 5 being extremely comfortable. It was agreed that a score of 3 or higher would pass, and 2 or lower would fail. Stability (or slippage) was a yes or no question for the investigator. After having the subject move around doing different activities, the investigator evaluated whether it slipped (0) or not (1). The helmet was deemed to fit if it passed for both comfort and stability.

Six head measurements were taken:

1. IPD (Interpupillary Distance)
2. Head Breadth
3. Head Length
4. Head Circumference
5. Bitragion Breadth
6. Face Breadth

Stepwise discriminate analysis was done to determine which two variables were the best predictors of fit. Overall pass/fail was the dependent variable, and the head measurements were the independent variables. Head Circumference and Head Length were determined to be the best combination. We plotted the overall pass score on the Head Circumference by Head Length bivariate plot to estimate the range of fit seen in Figure 4.14. This range seems to accommodate nearly 45% of the sample in one size.

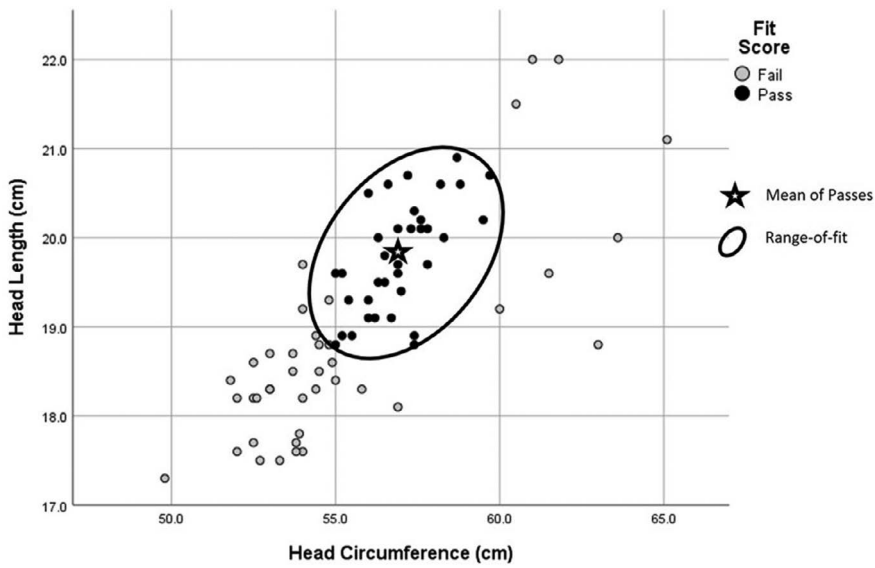


FIGURE 4.14 Helmet range-of-fit.

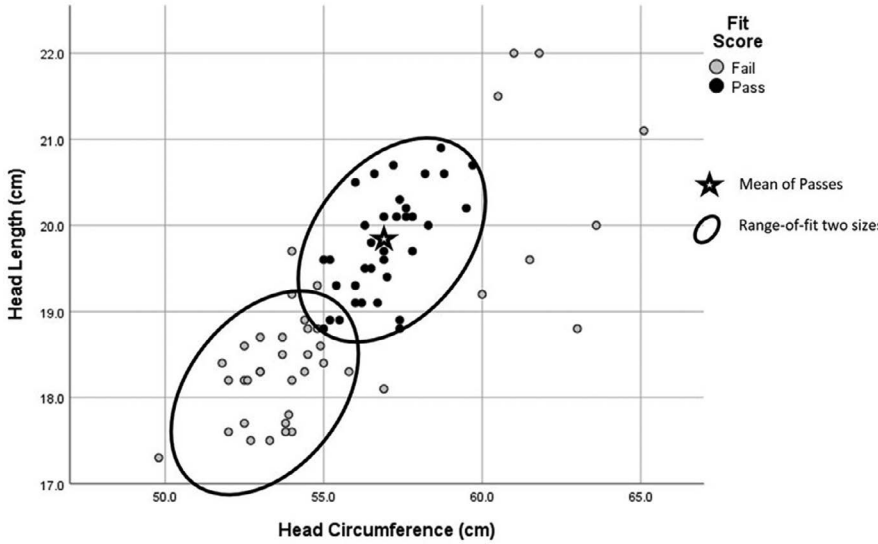


FIGURE 4.15 Range-of-fit in two sizes.

The size is not in the center of the sample, however, with this range of fit most of the sample might be accommodated in two sizes, with this size placement being the large size, accommodating most of the males. Therefore, it could be well placed. Adding a smaller size to accommodate smaller males and most of the females (see Figure 4.15), should accommodate about 85% of the sample.

There are a few subjects who seem to be extreme outliers with Head Circumferences larger than 60 cm, but Head Lengths less than or equal to 20 cm. There is also an outlier with a Head Circumference larger than 65 cm. We looked at the demographics of these subjects and found that they were all Asian. We determined we needed to do another fit test with at least 30 Asians to determine how well the helmet works for that demographic. There were only 14 Asians total in the sample, 7 male and 7 female and that is too few to characterize them. This needs to be done before we can decide on the final sizes.

It was decided to make a smaller size 1.5 cm shorter in length and 3 cm smaller in Circumference and do another Prototype Fit Test with the two sizes. The next test would examine: (1) the size overlap, (2) discomfort in specific spots on the helmet, (3) gaps that might be resulting in slippage between the helmet and the head using 3D scanning, and (4) evaluate the fit issues for the Asian demographic group.

Scenario 2: Existing Product With Fit and Sizing Issues in Multiple Sizes

This scenario was an evaluation of a pant for military women that came in eight sizes, even sizes 6 through 20, but only fit approximately 25–30% of the TP without the need for major alterations. The pants came unhemmed and were hemmed to the correct length when the pant was issued, so the length of the pant was not the problem. The alterations needed were in the hip and waist areas. The military must fit 100% of their population. Therefore, if a woman cannot find a pant to fit, she must

have it altered or custom-made. This is expensive and time-consuming. The goal for this organization was to fit at least 95% of the women with stocked sizes and reduce those needing alterations from 70–75% to 5% or less.

The best thing for the organization to do would have been to use the existing pant in a prototype fit test and determine how to improve the sizing. However, they believed they would solve the problem by adding odd sizes 7 through 19 and a larger size (22). At the time even-numbered sizes were referred to as Misses or Missy sizes, and odd-numbered sizes were referred to as Junior sizes and apparel companies believed these sizes were different. They used the same fit model (case) as the existing pant, and merely added the new sizes to the grade. With these changes they were confident they could skip the prototype test and go right to a verification of the sizes in a sizing loop test. They also skipped the pilot test because the COF for this pant was well established and documented in the military uniform code, and they thought they knew what measurements they needed.

The test results indicated their theory about the sizes was wrong. Doubling the number of sizes offered no significant improvement over the previous pant. The percentage who achieved an acceptable fit was only 30.3% and only 6% with an excellent fit. Instead of verifying a 95% fit, they had merely increased the testing complexity, costing more, and taking longer to complete. Even though the test was done poorly it did provide some useful information and is useful as a teaching example highlighting the need for prototype fit testing. Therefore, we summarize the study here.

A sample of 250 women, ages 18–55 was selected. This is a large sample size for a prototype test because this sample was intended to serve the dual purpose of representing the full TP sample. To meet their goal of 95%, only 13 women (or fewer) of the 250 should fail to have a size that fit. An example of the data collection sheet is shown in Table 4.13.

There were only two fit scores, one by the investigator and one by the subject. The fit scores were a scale of 1–4, with 1 being poor, 2 being fair, 3 being good, and 4 being excellent. Scores of 1 or 2 were fails, as they would not meet the uniform code without alternation. They assumed they would not have fit problems, so they didn't ask questions about areas that had fit issues. For example, it would have been nice to have questions about fit such as those shown in Table 4.14. These questions would have helped us identify the problem areas and determine how to fix them.

Each subject tried on all sizes that might be best and the one deemed to be the best fit by the investigators was chosen for fit testing. This size was called the Best Fit Size.

While one size might be acceptable for a sizing loop test after prototype fit tests have established the range of fit in a size, it was not ideal for this test, because the range of fit in a size was not known. It would have been better if the women would have been tested in more than one size so we could determine how much overlap there was in the sizes. It is possible (and turned out to be the case), to get a good fit in more than one size. If there is too much overlap the number of sizes can be reduced without loss of fit quality which saves time and money.

It is also possible to have a poor fit in all sizes, but for different reasons, so the choice of best fitting size can be arbitrary without a pilot test and an agreed upon decision guide added to the COF. For example, sometimes it can be because the hip is too tight, and others because the waist is too large.

TABLE 4.13
Example of Data Collection Sheet

Subject No. Date:	Date of Birth:
Place of Birth:	Rank:
Ethnic Group: W B A O	Hispanic: Y N
Weight	
Stature	
Neck Circumference	
Shoulder Circumference	
Bust Circumference	
Waist Circumference (preferred)	
Hip Circumference	
Waist Back Length	
Sleeve Inseam	
Sleeve Outseam	
Sleeve Length	
Waist Height (Outseam)	
Crotch Height (Inseam)	
Best Fit Pant Size	
Commercial Pant Size	
Subject Rating*	
Investigator Rating*	
Comments:	
* 1 = Poor; 2 = Fair; 3 = Good; 4 = Excellent.	

The ethnic groups recorded were White, Black, Asian, and Other. The purpose of recording this is to ensure that people with different body sizes and shapes are included. For this purpose, these groupings were sufficient. Hispanic people can be any ethnicity, so it was recorded separately. This is consistent with the way it is recorded for the US Census.

TABLE 4.14
Area Scoring Example

Area	Tight	Good	Loose
Hip (circle one)	–1	0	1
Waist (circle one)	–1	0	1
	Short	**Good**	**Long**
Crotch Length (circle one)	–1	0	1
Rise (circle one)	–1	0	1
	Too Far Back	**Good**	**Too Far Front**
Side Seam	–1	0	1

Thirteen anthropometric measurements were taken. Some of these were not relevant to pants but these were the measurements they were familiar with, and they added to the database of body measurements for other applications. Unfortunately, there were other important pant-related measurements that were needed but were not included, such as Crotch Length and Hip Breadth.

The data were evaluated to understand if any of the measurements that were taken were good predictors of the Best Fit Size. If so, they could be used as key variables to understand the range of fit within a size.

First, a stepwise regression analysis was done to determine the best anthropometric measurements for predicting the Best Fit Size. The size numbers were in a linear order and were treated as a scalar variable to be the dependent or Y variable. All anthropometric variables were input into the stepwise regression procedure as independent (X) variables. This was done using SPSS (version 26).

There were three models that were statistically significant at $\alpha = 0.001$. In other words, they predicted the size with a probability greater than chance. The first model was:

$$Y = -32.66 + 0.047X_1$$

where Y was Best Fit Size, X_1 was Hip Circumference. This means that Hip Circumference was the single variable that controlled the most variability in the size selection. The correlation (R), of the independent variable part of the model with the Best Fit Size was 0.885 and the R^2 was 0.782. This indicates that approximately 78% of the variability in the size selected is related to Hip Circumference.

The second model added Waist Circumference at the natural waist (Waist-Circ-Natural) to the first model and was:

$$Y = -33.437 + 0.036X_1 + 0.017X_2$$

where X_1 was Hip Circumference X_2 was Waist Circumference. This means that these two variables together controlled the most variability in the size selection, and the improvement in size selection by adding the second variable was significant. The R for this model was 0.909 and the R^2 was 0.826. This is an improvement of about 4.4% over the first model.

The third model added Shoulder Circumference (Shoulder-Circ) and was:

$$Y = -36.148 + 0.034X_1 + 0.013X_2 + 0.007X_3$$

where the first two variables were the same as the other models and X_3 was Shoulder Circumference. This seems odd, given we are evaluating a pant. However, it could indicate there is some other aspect of body shape that affects the size selection, but we didn't have a better variable in our dataset. Perhaps Hip Breadth might have given us a better result, but it was not measured. In any case, Shoulder Circumference's contribution was small. It improved the model correlation with Y by just 0.002, and with a significance value of just 0.032. Therefore, even though it was significant it was not meaningful and was ignored.

Next the Best Fit Size was plotted on a Waist Circumference by Hip Circumference bivariate chart. There was so much overlap we had difficulty seeing the distinction

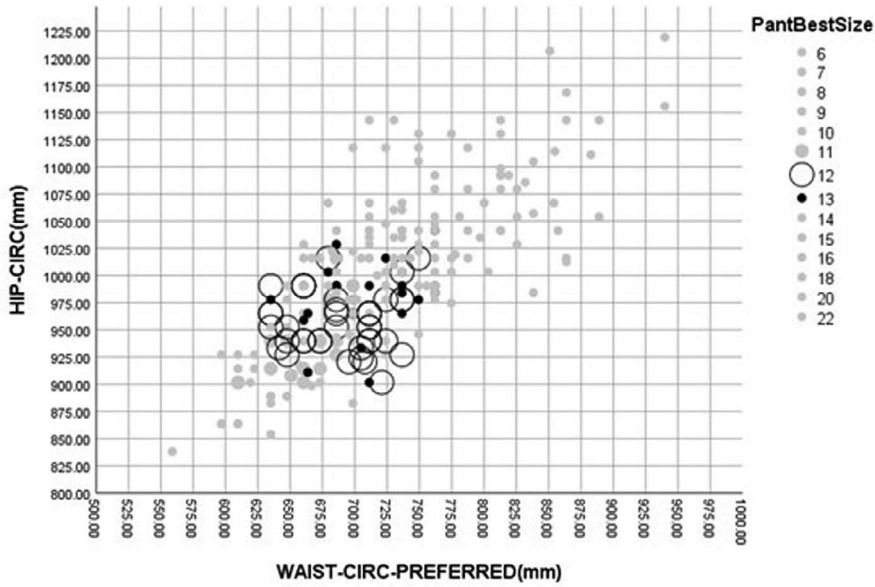


FIGURE 4.16 Sizes 12 and 13 showing the overlap.

between the sizes. Therefore, to clarify, we focused the graph on three sizes in the middle. We created a plot (Figure 4.16) with sizes 12 and 13 indicated with black circles and the others shaded a light gray. Size 11 was indicated with a larger circle than the other gray circles.

We see in this plot that sizes 12 and 13 are fitting the same size women. There seems to be a 100% overlap in sizes 12 and 13. In other words, these two seem to be duplicates of each other. Size 11 has substantial overlap but is fitting women who are a little smaller than the 12/13.

Since 12 and 13 seemed to be the same size, we looked at pairs of the other sizes and saw the same result. Sizes 10 and 11 seemed to be duplicates, as did sizes 8 and 9. The manufacturer was then asked about this result and they noted that pairs of even and odd sizes, such as 12 and 13, were within sewing tolerance of each other, so they used the same pattern for both. In other words, *they were the same size with different labels* so adding the odd sizes had no added value. If sizes are within sewing tolerance, they will not have a noticeable fit difference even if they were made with different patterns.

The plot in Figure 4.16 included all subjects, whether they got a good fit or not and we didn't know why they were given the size they were given. If we had had area ratings to tell us the source(s) of the fit issues, we could have looked at those to determine the fit issues. Likewise, if they had been tested in more than one size, we might have been able to find a relationship between size and body measurements or re-assigned the Best Fit Size in a more consistent manner. In the absence of these information, we compared those subjects for whom the fit was rated excellent (called it a pass) to the others (called it a fail). We chose just the excellent score because this

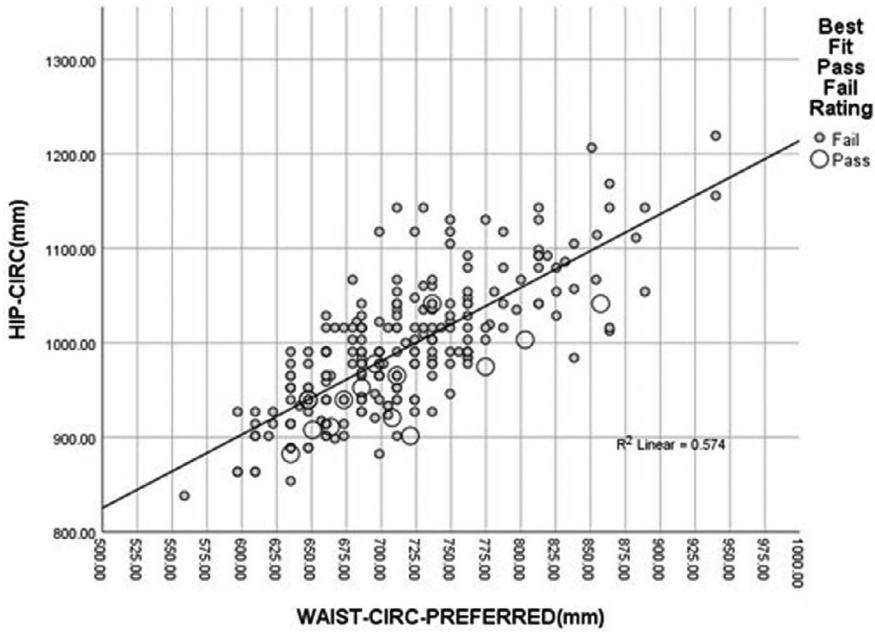


FIGURE 4.17 Pass versus fail in the pant.

score is certain. The next best score, "good", may have been a fit that was not really that good, but the rater didn't want to say it was only fair. With the excellent criteria we found a pattern shown in Figure 4.17.

The regression line for predicting Hip Circumference from Waist Circumference is shown in this figure, along with the R^2 value that indicates the degree of relationship between Hip and Waist Circumferences. This line indicated the estimated mean Hip Circumference value given the Waist Circumference value. Approximately half of the population will fall above this line and half will fall below. All but one of the people who received an excellent fit in the pant fell on or below this line. This indicates approximately half of the women were not getting a good fit because the pant hip area was too small.

There also seems to be a cluster of excellent fits for women between 625 mm and 750 mm in Waist Circumference. This is the size 12/13 area. We circled the excellent fit area that was within the area where size 12/13 was the Best Fit Size in Figure 4.18.

Eight out of 13 excellent fit scores (61.5%), were in the 12/13 Best Fit Size area. This was the size 12/13 women who had smaller hips. Some of the women in this area who did not rate the pant excellent were not given the size 12/13. Many were given the 11, which would have had a smaller hip. The number of women who did rate it excellent in this area suggests that the size 12/13 might be able to serve as a good base size with some pattern adjustments. We do not have any information about what the individual fit issues were. Perhaps they were a proportioning problem in the crotch or a problem with the rise, or perhaps the front to back proportioning was putting the side seam in the wrong place. If they had recorded fit in the different areas

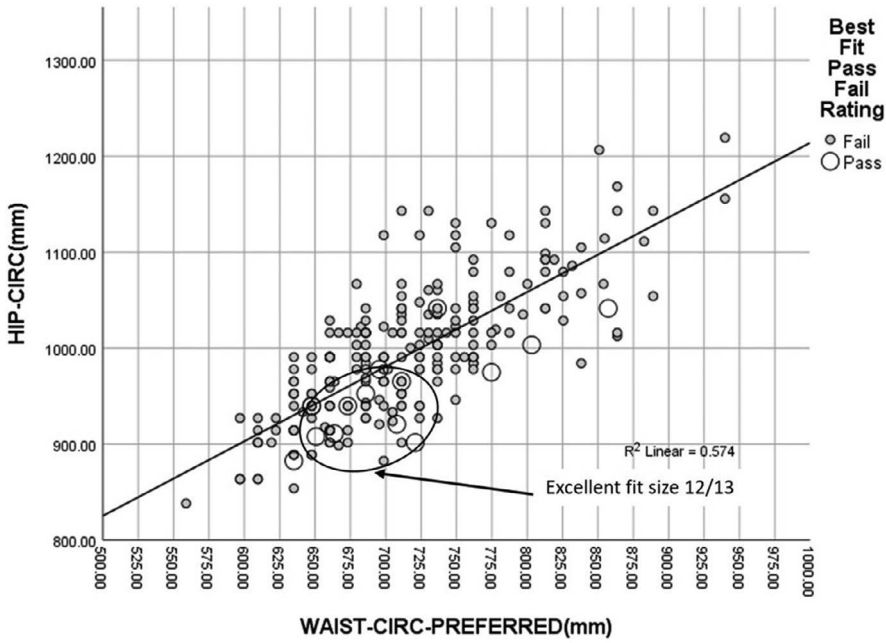


FIGURE 4.18 Range of fit in 12/13 versus full sample.

of the pant, we might have been able to identify the proportioning problem. As it was, we didn't have enough information, so determining the issues required another prototype fit test with improved fit scoring.

The other four excellent fit scores that are below the line seem to be in a line that is approximately 50 mm below the regression line. Only 4% of the women fall farther from the regression line than this. Since there were excellent fit scores on and close to the line in the size 12/13 range, but not in the larger and smaller size areas it suggests there may be an additional problem with the grade.

This manufacturer could have saved a lot of time and money and ended up with a better product if they had started with a prototype fit test using fewer sizes and fewer subjects to refine the base size pattern and adjust the grade before doing a full sizing loop fit test.

The range of excellent fits in the size 12/13 area suggests that if the pattern issues in the base size can be corrected it might be able to accommodate a 100 mm range in the waist and a 100 mm range in the hip. If this is the range of fit within one size, then a better sizing scheme for this pant might look like the one in Figure 4.19. This sizing scheme would accommodate approximately 95% of the TP with just eight sizes. It has two size ranges, one starts with the base size where the 12/13 was located, and the other with a base size with a larger hip at the waist size of the original 12/13.

This is not a new idea. In a Navy uniform sizing study (Robinette et al., 1991) we referred to the extra hip size range as "Women's". Some manufacturers are using a similar scheme and are calling the larger hip size ranges a "curvy" fit size range.

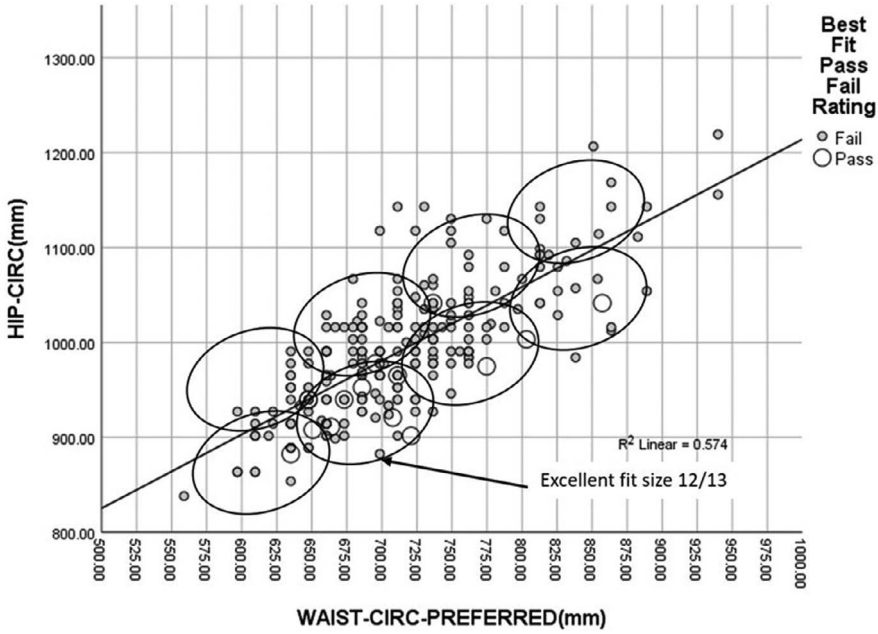


FIGURE 4.19 Eight sizes with 95% coverage of the TP.

In conclusion, the test in this scenario was not ideal and wasted time and money but it did give us some useful information including:

- Starting point for the base size
- Indication about the potential range of fit in a size
- Indication about the amount of overlap to expect
- Importance of Hip Circumference to pant sizing
- Database of 250 women
- Importance of pilot testing and prototype fit testing

Of course, we won't know what the size range should be until we do another round of fit testing, starting with a pilot test to refine our fit testing procedures, followed by an improved prototype fit test. The pilot test should be done with 4 or 5 women who are expected to be in the base size. This will not only help us refine our COF but will often indicate pattern issues that need to be fixed before testing. The prototype fit test does not need to include 250 subjects. The purpose is to create a good base size and determine the size spacing to minimize overlap. For this purpose, a sample size of 50–60 subjects is usually sufficient. However, it should include:

- Sizes selected from just the original eight sizes either:
  - All eight or
  - Just 10, 12, 14
- Area fit ratings such as those in Table 4.14.

- Additional anthropometric measurements related to the pants including:
  - Crotch Length
  - Hip Breadth and
  - Rise
- Testing every subject in at least three sizes to figure out the overlap

SIZING LOOP TESTS AND ANALYSIS

If we start with the design loop, then by the time we arrive at the sizing loop we should have a mature design that we are confident is good. This means we should have a complete and functional product with all components and functional features in at least one size. After this, we are ready to move toward the sizing loop.

Sizing loop testing: (1) verifies that the final product sizing is acceptable, (2) examines the cost versus benefit of each size before selecting the set of sizes to produce or purchase, (3) determines how many of each size to build or purchase (*the tariff*), and (4) provides charts or algorithms to help the wearer find the best size and adjustment.

The products of a sizing loop test are:

- Sustainable fit standards
- Cost-effective set of sizes for a TP
- Accurate size selection charts/algorithms
- Sustainable Tariff

Evaluating if the product meets our fit standard or works for a new TP is called a *fit audit*. Most sizing loop tests are fit audits. With a fit audit, we have a full set of sizes with a defined fit range for each and we want to check them. This can be an existing product that we want to test on a new TP, a new product using a prior product's sizing, or a new product that has gone through prototype fit testing with all sizes that we need to check on the full TP. For fit audits, we must either have or collect a full TP sample with raw anthropometric data that can adequately represent the TP for all important subgroups. This is important to understand who we are capturing in our sizes and who we are missing.

Evaluating an existing product for a new TP is important if the new TP is thought to be substantially different. For example, a product produced for sale in Japan may have sizing and proportioning issues when sold in Europe or the Americas. A more common example might be a product that was made for one age group being adjusted for a different one. In this instance, the number and variety of sizes might not change, but the amount of each size produced might change. An older population may require more of the larger sizes and fewer of the smaller sizes.

For existing products, products with a precedent, or products that must meet a fit standard, we might begin a fit audit with the COF and pilot test that falls between the design loop and the sizing loop. For example, organizations that are not manufacturers of the product but need to fit everyone in their organization (such as medical personnel, firefighters, or the military), might start with a COF that is validated and approved in a pilot test, then run a sizing loop test with

subjects from their organization (the TP), before they purchase in quantity. The testing, if done correctly, will reveal how many of each size they will need so they will not purchase too many of some sizes and not enough of others. It is usually the case that there are sizes that will not be needed by anyone in the organization and needed sizes that are not available. This is particularly true if the product was created for a different or an unknown TP since people from different occupations have different body sizes. It is also very common for there to be duplicate sizes in a size range, just as we saw in Scenario 2 in the prototype fit testing section. This happens when manufacturers do not include sizing loop testing or fit audits in their development process. The manufacturers never learn that they have duplicates. With sizing loop testing duplicate sizes are revealed and duplicates can be dropped from the purchase.

Some existing products are evaluated by manufacturers in a sizing loop before producing the new season's line. This adjusts the sizing for changes in the design or materials used, keeps up with market changes, and ensures the sizing is effective. If the manufacturer has a sustainable sizing standard, sizing loop testing can ensure that the new product sizes will fit the same people as other product sizes in a line of wearable items. This provides customer confidence in size purchasing and maintains an effective fit for minimizing waste. If sizing loop testing reveals that the changes in the design or materials for the new season cause more serious design issues, or do not meet the fit standard the results from the sizing loop test may suggest the need for product changes and another design loop or sizing loop test.

The sizing loop has six steps:

1. Conduct full-fit test
2. Establish range of fit in each size
3. Map against TP
4. Create size prediction equations and charts
5. Conduct size cost/benefit comparison of sizing alternatives
6. Produce sizing loop output

The first three steps are the same as the three steps in the design loop starting with prototype fit testing. There are two main differences in the sizing loop: (1) the prototypes are complete final products, and (2) a full TP sample is needed, as described in Chapter 2. The experimental design can be identical to the experimental design of a prototype fit test and it can be helpful to standardize the test procedures. Standardized procedures make analysis easier to follow and faster to complete for new versions of a product or new products of a similar type.

The full TP sample should have at least 200 or more subjects of each gender and if we are using a stratified random sample, we should have 30 subjects in each stratum, so there are enough for weighting the sample to match the strata proportions in the TP. If we have a large sample from the TP for whom we have the appropriate anthropometry and demographics when we begin the sizing loop, a smaller sample with the same measurements and demographics can be used for the fit testing. If we do not have a sufficiently large sample from the TP, then we will need to collect a large enough sample during fit testing to represent the TP.

The other three steps are additional analyses of the test results to produce and finalize the sizing loop products, specifically the:

1. Cost-effective set of sizes
2. A tariff of how many of each size to produce and/or purchase
3. Accurate size prediction charts and algorithms to help the wearer get the right size
4. A sustainable fit standard for future versions or products

Evaluating the Cost Versus Benefit of Sets of Sizes

The Cost versus Benefit analysis is simply a comparison of different sets of sizes and the estimated percentages accommodated. This is done by visually mapping the sizes on bivariates of the key variables from the full TP sample. The cost of producing each size, and the benefit of having the size for expected sales or customer satisfaction are then considered.

We illustrate this in the first case study in Chapter 5. In that study, we estimated the range of fit for five sizes of a woman's pant using a fit test. (Note this can be a sizing loop test or a design loop prototype fit test.) We mapped the estimated range of fit of the five sizes against a weighted full TP sample, and we compared them to another set of sizes with the same range of fit in each size, but a different base size that was more centered over the TP. The two sets of sizes are illustrated in Figure 4.20. The original five sizes are shown in part a. at the left and the new sizes in part b. at the right. By adjusting the sizes to make the Waist Circumference larger by 20 mm (a half size up for this measurement only), then dropping the smallest size and adding one larger, we increased our coverage of the TP by more than 30%. While the coverage would be increased further by dropping the smallest size remaining and adding another larger size, the company felt that would be too dramatic of a change for their current customers. Their customers are used to them having very small sizes so they felt they might not be able to attract the larger women to try their products. This is an example of using statistics to guide decisions but taking all factors into account before making them.

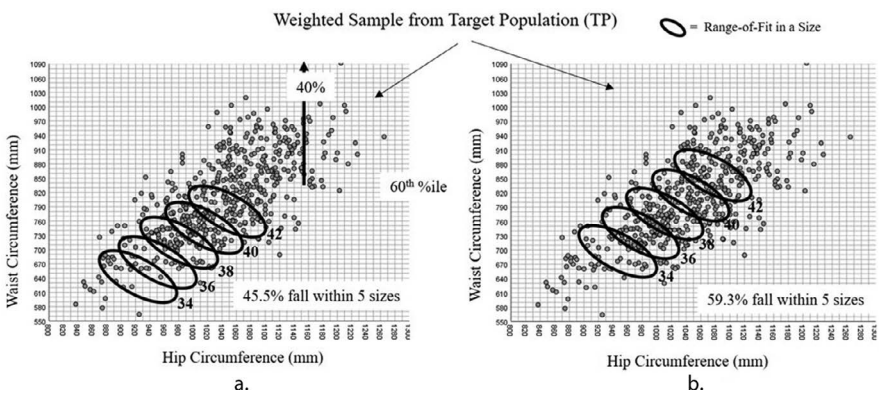


FIGURE 4.20 Cost/benefit size range comparison. (a) Original five sizes. (b) New five sizes.

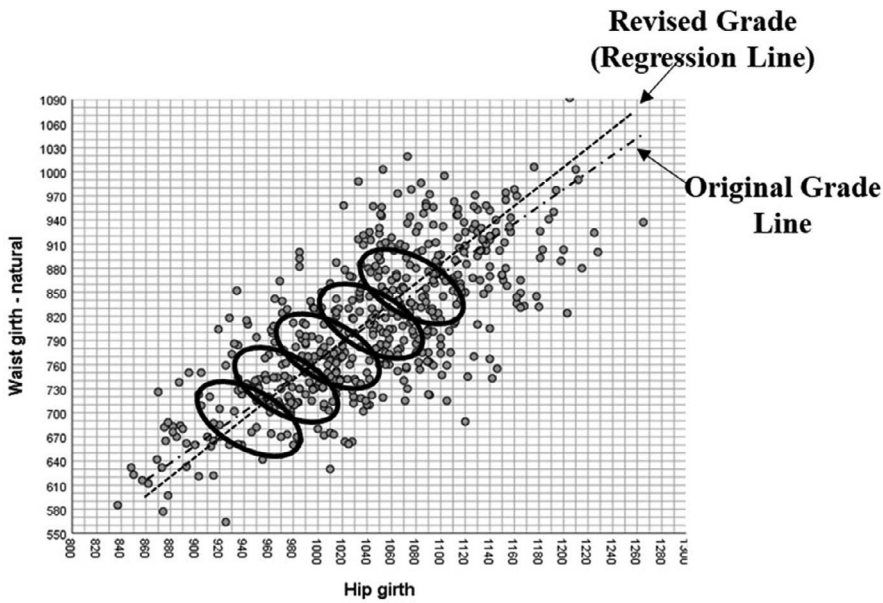


FIGURE 4.21 Regression line to revise grade.

The Hip Circumference grade change to 35 mm was based on the linear regression line for predicting Hip Circumference from Waist Circumference as shown in Figure 4.21. This line is the best estimate of the most likely Hip Circumference for the Waist Circumference location and around which the greatest concentration of subjects will occur.

Since the grade starts from the middle size, the impact of having the wrong grade is not noticeable until we get to sizes that are farther away from the middle. This is one reason companies have particular issues with the largest and smallest sizes. For example, in the article by Spedding (2019) one company had too many size 16 returns and the other had size 16 always sell out. Both came to the same conclusion that they needed another size. A third company did not say what the problem was but also thought the answer was more sizes. They all knew there was a problem, but because they did not know who they were fitting and who they were missing, they could only make arbitrary guesses at how to solve it. Their guess was to add more sizes with its added expense and no idea if it would solve the problem. Maybe they did need another size, but they added the wrong one. If they would do a fit audit, they could find out and may find they do not need more sizes at all. They need better placed sizes just as in our pant example.

Since we made a change to the grade, it is important to do an additional sizing loop test to verify that our changes were effective and did not cause any other issues. It is possible to fix one thing and break something else in the process. This verification sizing loop test need not be extensive. It can be more like a trade-study test where we use a convenience sample of people with selected key variable sizes to test each new size. Once we verify that the new size works, we can assume the range of fit is the same.

Determining the Tariff

After deciding and verifying the sizes, we must next determine how many of each size to purchase and stock. A table listing the percentage or proportion of each size to purchase is called a *tariff*. The number of people who need each size differs depending on the size, so it is not a good idea to purchase the same amount of every size.

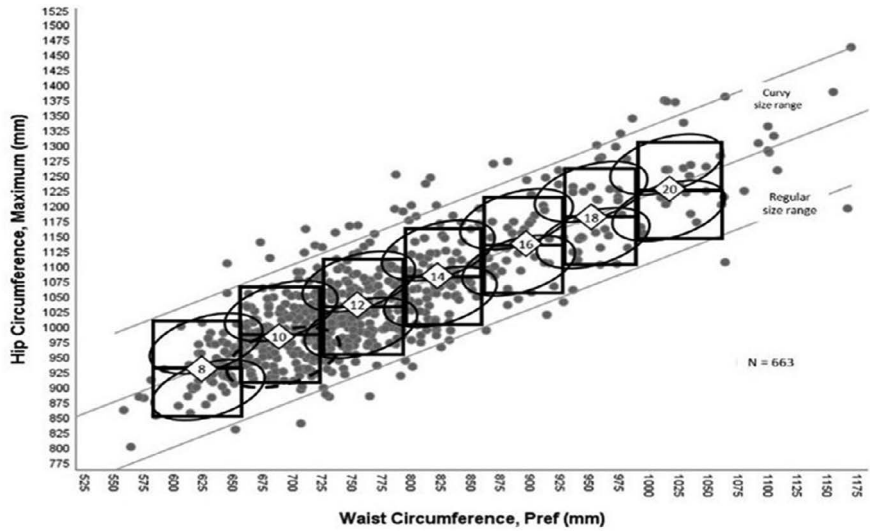
The percentage of potential wearers who will fit in each size is important, but it is not the only consideration when purchasing sizes. To create the tariff, we must also consider the people who will not get a good fit but who will buy or select it anyway, such as those who will alter it or tolerate a poor fit. When we count the number of people who will need each size, we need to count these people as well.

To count the number of people in each size, we must first assign a size to each of the subjects in our sample. If we have fit scores for each subject, we can use the size of best fit as their assigned size. If we do not have fit scores, then we need to predict an assigned size. The simplest way to do this is to create simplified size categories based on our range-of-fit categories and our assumptions about what sizes people will buy who do not get a good fit.

For example, we had a pant sizing system with 14 sizes in two size ranges, a regular range and a curvy range that had a larger hip. To create a tariff for this pant we drew rectangular categories using the range-of-fit ellipses as guides. The rectangular size boxes have no overlap for counting purposes. This is shown in part a. in Figure 4.22. Then we added rectangular areas to the sizes based upon our assumptions about what sizes people would use who did not get a good fit but would purchase or obtain the size anyway. This is shown in part b. For example, size 20 curvy (20C) includes subjects who have a size 18 waist but larger hips than an 18 curvy shape (18C). These are people who will also select the size 20C. Size 20 regular (20R) also includes people who have smaller hips than the 20 regular shape. These are people who will also select the size 20R. Size 18 includes subjects who have a size 16 waist but larger hips.

Rectangles are used because they create an easy way to count, and while not perfect, they will be close enough for a tariff. To do the counting, we use the simplified size category rectangles to assign a size to each person in our sample. For example, all people who are within the rectangle (box) with Waist Circumference greater than 990 mm and less than or equal to 1060 mm and a Hip Circumference greater than 1055 mm and less than or equal to 1225 will be assigned the size 20R. Size 20C is the sum of people in two rectangular areas (boxes), the box with the size 20 waist and curvy hip and the rectangle with the size 18 waist and the 20 curvy hip. The size assignment statements used to assign a size for everyone in the TP who falls within a size are shown in Table 4.15.

After all sizes are assigned, we count the number who were assigned each size and divide this number by the number of people who were assigned any size to get the proportion of people who will select each size. There will be some subjects who are not assigned any size. These are people we expect will not purchase the product, so we do not need to have a size available for them.



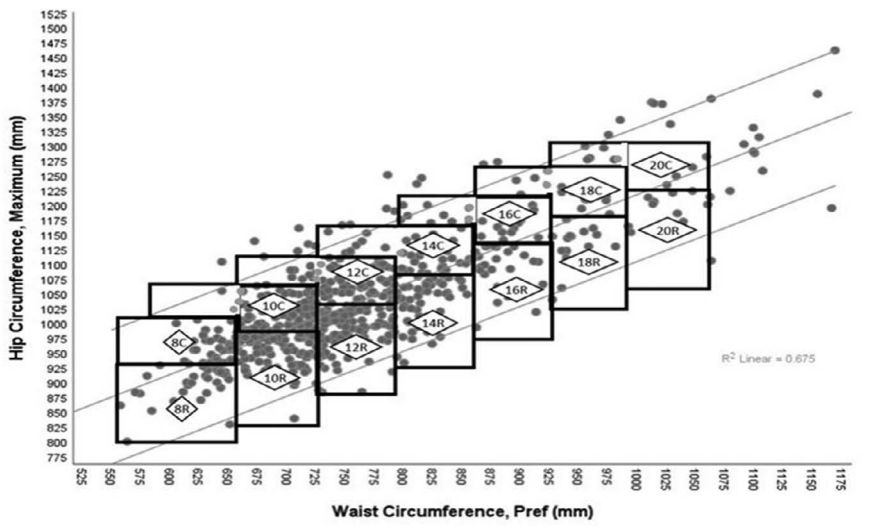


FIGURE 4.22 Size categories for creating tariff. (a) Simplify size ranges with rectangles. (b) Add poor fit wearers areas.

We counted the number in each size and calculated the number of each size to be purchased in an order of 10,000. This is the tariff and is shown in Table 4.16. The table shows the count for each size, the proportion in each size with respect to the total in a size, the percent in each size with respect to the total in a size, as well as the number in each size per 10,000.

TABLE 4.15
Size Assignment for the Tariff

Size	Waist Min	Waist Max	Hip Min	Hip Max	Statement
8R	550	655	800	930	If Waist ≥550 AND Waist <655 AND Hip ≥800 AND Hip <930 then Size = "8R";
8C	550	655	930	1025	If Waist ≥550 AND Waist <655 AND Hip ≥930 AND Hip <1025 then Size = "8C";
10R	655	725	825	990	If Waist ≥655 AND Waist <725 AND Hip ≥825 AND Hip <990 then Size = "10R";
10C Box 1	655	725	990	1070	If Waist ≥655 AND Waist <725 AND Hip ≥990 AND Hip <1070 then Size = "10C";
10C Box 2	585	655	1025	1070	If Waist ≥585 AND Waist <725 AND Hip ≥1025 AND Hip <1070 then Size = "10C";
12R	725	795	875	1045	If Waist ≥725 AND Waist <795 AND Hip ≥875 AND Hip <1045 then Size = "12R";
12C Box 1	725	795	1045	1125	If Waist ≥725 AND Waist <795 AND Hip ≥1045 AND Hip <1125 then Size = "12C";
12C Box 2	655	725	1070	1125	If Waist ≥655 AND Waist <725 AND Hip ≥1070 AND Hip <1125 then Size = "12C";
14R	795	860	925	1080	If Waist ≥795 AND Waist <860 AND Hip ≥925 AND Hip <1080 then Size = "14R";
14C Box 1	795	860	1080	1175	If Waist ≥795 AND Waist <860 AND Hip ≥1080 AND Hip <1175 then Size = "14C";
14C Box 2	725	795	1125	1175	If Waist ≥725 AND Waist <795 AND Hip ≥1125 AND Hip <1175 then Size = "14C";
16R	860	925	975	1130	If Waist ≥860 AND Waist <925 AND Hip ≥975 AND Hip <1130 then Size = "16R";
16C Box 1	860	925	1130	1225	If Waist ≥860 AND Waist <925 AND Hip ≥1130 AND Hip <1225 then Size = "16C";
16C Box 2	795	860	1175	1225	If Waist ≥795 AND Waist <860 AND Hip ≥1175 AND Hip <1225 then Size = "16C";
18R	925	990	1025	1180	If Waist ≥925 AND Waist <990 AND Hip ≥1025 AND Hip <1180 then Size = "18R";
18C Box 1	925	990	1180	1275	If Waist ≥925 AND Waist <990 AND Hip ≥1180 AND Hip <1275 then Size = "18C";
18C Box 2	860	925	1225	1275	If Waist ≥860 AND Waist <925 AND Hip ≥1225 AND Hip <1275 then Size = "18C";
20R	990	1060	1055	1225	If Waist ≥990 AND Waist <1060 AND Hip ≥1055 AND Hip <1225 then Size = "20R";
20C Box 1	990	1060	1225	1305	If Waist ≥990 AND Waist <1060 AND Hip ≥1225 AND Hip <1305 then Size = "20C";
20C Box 2	925	990	1275	1305	If Waist ≥925 AND Waist <990 AND Hip ≥1275 AND Hip <1305 then Size = "20C";

TABLE 4.16
Tariff for Example Women's Uniform Pant

Size	Count	Proportion	Percent	Per 10,000
8C	33	0.052	5.2	524
8R	32	0.051	5.1	508
10C	79	0.125	12.5	1254
10R	94	0.149	14.9	1492
12C	75	0.119	11.9	1190
12R	120	0.190	19.0	1905
14C	43	0.068	6.8	683
14R	49	0.078	7.8	778
16C	25	0.040	4.0	397
16R	22	0.035	3.5	349
18C	18	0.029	2.9	286
18R	18	0.029	2.9	286
20C	13	0.021	2.1	206
20R	9	0.014	1.4	143
No Size	33			
Total	663			
Total in a size	630	1.0	100.0	10,000

After examining the tariff some sizes might be dropped. For example, it might not be worth the manufacturing cost to make just 143 of the size 20R. If a size is dropped the proportion out of 10,000 must be re-calculated either by adding the dropped size's numbers to another size (such as adding 143 to the 206 in the 20C), or by dropping the numbers altogether and removing them from the total in a size (such as dropping 20R and reducing the total in a size to 630 – 9 = 621) and recalculating the proportion, percent, and per 10,000 numbers.

A more complex method for assigning size can be done using logistic regression, provided the fit test sample is collected appropriately. First, the fit test must be done such that every person is tried in every size they can get on. Second, the overall fit in every size must be scored either as a two category (binomial), pass/fail that can be scored 0 or 1, or it must be a three or more discrete category score (multinomial) that is two directional, such as small, good, large that can be scored –1, 0, 1, respectively. With this information a logistic regression equation can be calculated from the fit test sample that can be used to indicate the most likely size from a person's body measurements.

For example, a woman's pant that came in four sizes was fit tested on sample of 47 women. The four sizes were Extra-Small (XS), Small (S), Medium (M), and Large (L). Each woman wore every size they could don, and every size was scored either small (–1), good (0), or large (1). Forty-five anthropometric measurements were also taken. Then MLR equations were calculated for each size for predicting fit quality probability from body measurements. The probability of being too small, good, and too large can then be calculated for new or additional subject from the TP. The

TABLE 4.17
Multinomial Logistic Regression Prediction of Fit Quality for Each Size

Subject No.	Fit Quality Probability												Likely Best Size
	Too Small				Good				Too Large				
	XS	S	M	L	XS	S	M	L	XS	S	M	L	
1	1.00	0.00	0.00	0.00	0.00	**0.98**	0.00	0.02	0.00	0.02	1.00	0.98	S
2	1.00	1.00	0.97	0.00	0.00	0.00	0.03	**0.90**	0.00	0.00	0.00	0.10	L
3	1.00	0.00	0.00	0.00	0.00	**0.91**	0.00	0.02	0.00	0.09	1.00	0.98	S
4	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.02	**0.97**	0.99	1.00	0.98	XS-1
5	1.00	0.00	0.00	0.00	0.00	**0.98**	0.00	0.02	0.00	0.02	1.00	0.98	S
6	0.00	0.00	0.00	0.00	**0.94**	0.01	0.00	0.02	0.06	0.99	1.00	0.98	XS
7	1.00	0.99	0.18	0.00	0.00	0.01	**0.81**	0.07	0.00	0.00	0.00	0.93	M
8	1.00	0.98	0.00	0.00	0.00	0.02	**0.93**	0.05	0.00	0.00	0.07	0.95	M
9	1.00	0.01	0.00	0.00	0.00	**0.99**	0.01	0.02	0.00	0.00	0.99	0.98	M
10	0.00	0.00	0.00	0.00	0.04	0.01	0.00	0.02	**0.96**	0.99	1.00	0.98	XS-1
11	1.00	1.00	1.00	**0.88**	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	L+1
12	0.01	0.00	0.00	0.00	**0.99**	0.51	0.00	0.02	0.00	0.49	1.00	0.98	XS
13	0.02	0.00	0.00	0.00	**0.97**	0.19	0.00	0.02	0.00	0.81	1.00	0.98	XS
14	0.00	0.00	0.00	0.00	0.02	0.09	0.00	0.02	**0.98**	0.91	1.00	0.98	XS-1
15	1.00	1.00	1.00	0.00	0.00	0.00	0.00	**0.97**	0.00	0.00	0.00	0.03	L
16	1.00	1.00	0.96	0.00	0.00	0.00	0.04	**0.80**	0.00	0.00	0.00	0.20	L
17	1.00	1.00	0.10	0.00	0.00	0.00	**0.89**	0.86	0.00	0.00	0.00	0.14	M
18	1.00	1.00	0.01	0.00	0.00	0.00	**0.99**	0.87	0.00	0.00	0.00	0.13	M
19	1.00	1.00	0.00	0.00	0.00	0.00	**0.99**	0.14	0.00	0.00	0.01	0.86	M
20	0.00	0.00	0.00	0.00	**1.00**	0.40	0.00	0.02	0.00	0.60	1.00	0.98	XS

probability predictions for 20 new subjects are shown in Table 4.17. Each subject has a probability estimate for being too small in each size, for being good in each size and for being too large in each size. The probability values go from 0 to 1, with 0 being very unlikely and 1 being very likely. For example, a score of 1 in the too small area means it is very likely to be too small.

The likely best fit size from the logistic regression predictions is shaded in gray and listed in the right column. When one or more sizes have a high probability (0.8 or higher) of being a good fit, then the one with the highest probability is the most likely best fit size and is shaded gray. When no sizes have a high probability of a good fit, such as for subjects 4, 10, 11, and 14, then the subject needed a size that was not available. In this situation, the "too small" and "too large" probabilities were examined to indicate what size they needed. For example, we see that subject 4 did not have a probability of a good fit in any size. All her good fit probability scores were less than 0.05. She also had a near 0 probability of being too small in any size, and the smallest size, had a 0.97 probability (or 97%) of being too large. This indicated that she needed an even smaller size. Therefore, the most likely best fit size was listed as XS-1.

To get the tariff we calculate the most likely best fit size for every subject in our full TP sample, count the number of people in each size, and divide by the number of people who will wear any size. We can also decide to put the people who didn't fit in any size into the next closest size if we think that is reasonable.

Size Prediction

Size prediction is simply a set of tools to help the wearer get the best fitting size. Most size guide tables for fashion apparel, such as the women's pant size selection guide in Table 4.18, do not illustrate the range of fit for combined measurements. They usually only list a single value for each measurement. It is difficult to use these guides to find the best fitting size.

For example, if a woman has a Waist Circumference of 690 mm (69 cm) and a Hip Circumference of 950 mm (95 cm) she falls more toward the medium for Waist Circumference and slightly above the small for Hip Circumference. Which size should she choose? The chart suggests the small will be too tight with a waist that is 3 cm smaller and a hip 1 cm smaller than her measurements. However, the small would not only fit her better. But she would be happier because she fits in a smaller size!

Bivariate charts with a combined variable range of fit help people select the actual best size, and the one we created for the tariff is a good one to use. It is simple and easy to read, and it can be published in a catalog or posted online. The wearer can see the range of fit in each size so even if they don't know their measurements, they can often make a reasonable judgment about where they fall. To make the chart easy to read it is important to remove the data so only the size categories appear. It would look like the chart in Figure 4.23.

The logistic regression method we described for the tariff is another method that might be considered. For this method, the subjects cannot be expected to create or use the equations, but the fit prediction algorithms can be put into an app that does all the calculating given a subject's body measurements.

A comparison of the accuracy of the logistic regression method versus using the pant company's size selection chart for size prediction was done. For this test, the subject first selected the size she thought would be best based on the company's chart. Then all the sizes were fit tested and the true best size was found. The MLR equations from the earlier fit test were used for predicting the best size for each subject. The results indicated that the MLR predictions were correct 95% of the time, and the size selected using the company's chart was only accurate 45% of the time.

TABLE 4.18
Example of a Typical Size Guide Table (cm)

SIZE	XS	S	M	L	XL
WAIST	62	66	70	74	78
HIP	90	94	98	102	106

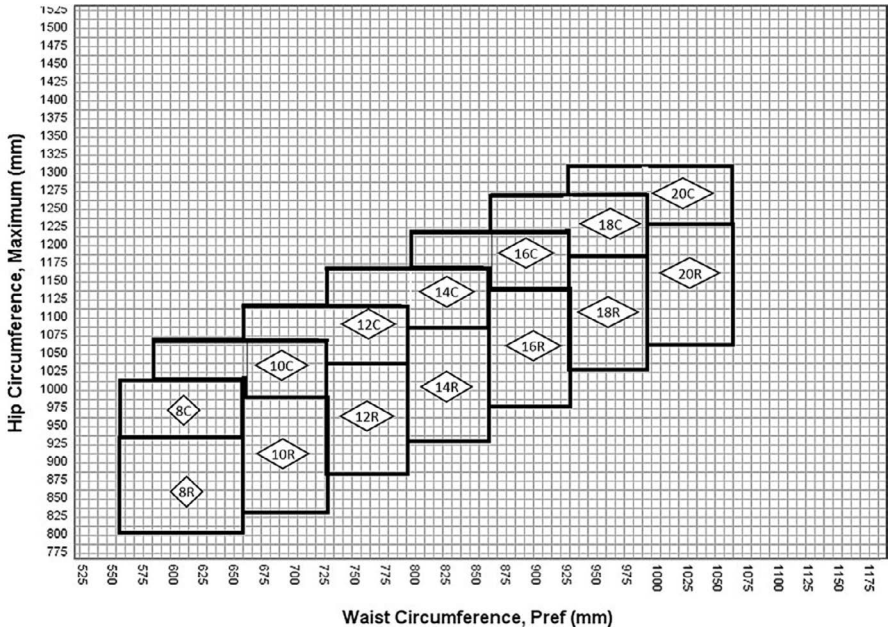


FIGURE 4.23 Example size selection chart using tariff categories.

However, it is not always possible or feasible to use predictive equations, because many people do not have the tools or the ability to take the body measurements in the same way they were taken to create the equations. This will throw all the estimates off. There have been some very promising studies on using 2D cameras, such as a cell phone camera, to adapt 3D models to represent an individual (Ballester et al., 2016, 2018, 2015). These methods show promise for apps to help people predict their size in the future.

REFERENCES

Ballester, A., Parrilla, E., Piérola, A., Uriel, J., Pérez, C., Piqueras, P., Nácher, B., Vivas, J. A., & Alemany, S. (2016). Data-Driven Three-Dimensional Reconstruction of Human Bodies Using a Mobile Phone App. International Journal of the Digital Human, 1(4), Article 4. https://doi.org/10.1504/IJDH.2016.084581

Ballester, A., Pierola, A., Parrilla, E., Uriel, J., Ruescas, A. V., Perez, C., Dura, J. V., & Alemany, S. (2018). 3D Human Models from 1D, 2D and 3D Inputs: Reliability and Compatibility of Body Measurements. Proceedings of 3DBODY.TECH 2018 - 9th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, Lugano, Switzerland, 16–17 October 2018. 132–141. https://doi.org/10.15221/18.132

Ballester, A., Valero, M., Nácher, B., Piérola, A., Piqueras, P., Sancho, M., Gargallo, G., González, J. C., & Alemany, S. (2015). 3D Body Databases of the Spanish Population and its Application to the Apparel Industry, Proceedings of the 6th International Conference on 3D Body Scanning Technologies, Lugano, Switzerland.

Budurka, W. (1984). Developing Strong Systems Engineering Skills. IBM Technical Directions, *10*(4), 40–48.

Cochran, W. G. (1977). Sampling Techniques. John Wiley & Sons, Inc.

Harrison, C. R., & Robinette, K. (2002). CAESAR: Summary Statistics for the Adult Population (Ages 18-65) of the United States of America (Technical Report AFRL-HE-WP-TR-2002-0170), United States Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division. https://apps.dtic.mil/dtic/tr/fulltext/u2/a406674.pdf

Lapin, L. L. (1978). Statistics for Modern Business Decisions (Second). Harcourt Brace Jovanovich, Inc.

Montgomery, D. C. (1976). Design and Analysis of Experiments. John Wiley & Sons, Inc.

Myers, R. H., & Montgomery, D. C. (1995). Response Surface Methodology: Process and Product Optimization Using Designed Experiments. John Wiley & Sons, Inc.

Ott, L. (1977). An Introduction to Statistical Methods and Data Analysis. Duxbury Press, A Division of Wadsworth Publishing Company, Inc.

Robinette, K. M., Mellian, S. A., & Ervin, C. A. (1991). Development of Sizing Systems for Navy Women's Uniforms (Technical Report AL-TR-1991-0117; Issue AL-TR-1991-0117), Armstrong Laboratory, Air Force Systems Command. https://apps.dtic.mil/docs/citations/ADA250071

Spedding, E. (2019, May). Editors, Designers and Model Agents on the Fashion Industry's Sizing Problem, Who What Wear, United Kingdom. https://www.whowhatwear.com/fashion-industry-sizing-problem

Stangroom, J. (2021). Social Science Statistics [Computer software]. https://www.socscistatistics.com/

5 Mass-Produced Apparel

Daisy Veitch

ABSTRACT

This chapter discusses how to apply the Sustainable Product Evaluation, Engineering, and Design process (SPEED) process to mass-produced apparel products. Examples of the concept-of-fit (COF), pilot testing, trade studies, prototype tests, and applications in the sizing loop are presented in the form of real-world case studies. The case studies include examples from manufacturers, retailers, and procurement groups for organizations such as the military, firefighters, hospitals, and so on. Retailers and procurement groups may or may not be able to influence the design and sizing of existing products. However, they want to understand the sizing for several reasons such as: (1) to ensure it will accommodate their target population (TP), (2) to decide which sizes to purchase, (3) to decide how many of each size to purchase, and (4) to specify design criteria and sizing for future products. This knowledge enables better decisions that lead to improved efficiencies, profitability, and sustainability of their business. The case studies include examples of full fit tests, creating and checking a size selection chart, creating a sustainable fit standard, grading, and calculating tariffs.

The commercial apparel industry has been successfully making and selling apparel for a profit for more than 100 years. Obviously, there are some things they are doing well and some of the procedures in our process come from experience with this industry. However, there are also many fit issues, dissatisfied customers, lost sales, and wasted sizes that end up on sales racks and garbage dumps so there is a lot of room for improvement. In this chapter, we review industry practices, discuss what works and what doesn't, and demonstrate how to improve upon them for more sustainable products and sizing.

BACKGROUND

There are three types of commercial apparel production: (1) individualized apparel, (2) mass production, and (3) mass customization. *Individualized apparel or made-to-measure* products are products styled and fitted to one individual. The process of fitting to the individual is called *tailoring*. Until the First World War, tailoring was the method by which most apparel was produced. Individualized fit does not have sizes and only the individual being fitted will wear the product. In tailoring, fit standards that apply to more than one person are not needed.

Mass production, also known as ready-to-wear or pret-a-porter, is the manufacturing of large quantities of standardized products. It includes not only fashion apparel, but also technical products worn on the body, such as body armor and personal protective equipment (PPE). Technical garments usually have specialized requirements

DOI: 10.1201/9781003397533-5

that must be included in the COF and testing procedures, but the overall development and assessment process is the same. This chapter focuses on mass production because this is where the opportunities for improving sizing, maintaining fit, and reducing waste are most evident.

Mass customization is a hybrid between mass production and individualized apparel. It most commonly refers to customizing a garment by making design selections including fabric, color, pocket positions, custom embellishments like embroidery, or minor shortening or lengthening of the sleeve or leg lengths. These adaptations use a mass production pattern and sizing system, so the procedures of mass production apply.

In the apparel industry today, manufacturing practice follows the process of design specification (drawing and measurements), product development (PD) including pattern and prototype, fit assessment on a single fit model representing the base size, alterations and iterative fittings on the fit model prior to final approval, then pre-production steps (grading and production specs) and finally production. The fit model is a case that is represented both as a human subject who provides feedback about the fit, and with a physical manikin. Sometimes the case can be represented as body measurements in a chart and a digital model as well.

The problems with this process arise with the fit model and grade selections and their applications via the concept-of-fit (COF). "Fit model selection and grades are standards often unique to each company, separate from product development (PD). They are part of quality control standards" (Janice Larsen, formerly Lead of Fit and Product Development, Lululemon, personal communication, 2023).

Unfortunately, they are typically guesses based upon what others are doing or what has been done in the past. Sometimes they are agreed upon guesses, but they are guesses nonetheless, and they are not optimized for the target customers or population. When the companies never learn who they are fitting in their sizes they cannot determine how to best fit their market.

There are international sizing standards for commercial apparel but they are only loosely followed, if at all, for many reasons: (1) it is impossible to find a fit model who exactly meets the standard, (2) the standards are expressed as body measurements, not garment measurements so without the three-dimensional (3D) body and a COF it is impossible to test whether the garment meets the standard, (3) the standards are for generic populations and company products are for more specific subpopulations such as specific age groups, or regions of the world, or fitness level, (4) the standards are for generic purposes and company products are often for more specific purposes such as uniforms or work clothes for specific occupations, garments for specific sports such as cycling or rock climbing, and (5) the sizing standards have not been fit test verified so no one knows who will fit in the sizes or if the sizes are appropriate or not. Individuals and companies can do better.

Some companies have multiple departments for a garment type, and each department might have its own fit model and grade standard. For example, they might have multiple departments for women, such as Juniors, Misses, and Women's, that all produce the same items. For example, a pant might be produced in both Junior and Misses sizes, with the Junior sizes labeled with odd numbers, 3 through 15, and Misses labeled with even numbers, 2 through 14. The companies often believe these

two size ranges are fitting different body sizes and ranges but if they use similar base size fit models and grades there can be little difference between them. As a result, they produce equivalently sized garments in the two departments, doubling the cost, without adding any improvement in customers accommodated.

Some companies are beginning to differentiate size ranges so, for example, a Women's 14 might use a fit model that has larger hips for the size 14 waist size than the Misses size 14. Or a Misses Petite 14 fit model might be shorter but the same Waist size as the Misses 14 Regular. However, currently, they are mostly chosen with educated guesses, with little or no feedback about fit and size as it relates to the target population (TP). Without a fit audit, it is not clear who is fit in their size ranges, and there is a lot of money to be saved, and customers to be added by finding out.

In contrast, the Sustainable Product Evaluation, Engineering, and Design (SPEED) process provides validated fit ranges for each size and for each garment. The apparel producers who follow the process can provide retailers and customers with accurate and validated size prediction charts and algorithms. This ensures retailers purchase the best size assortment for their customers and customers get the correct size. This minimizes both the risk of lost sales and wasted sizes.

The process is improved by: (1) using methods for selecting the fit model as described in Chapter 3 to hit the TP sweet spot, (2) fit testing on additional people to optimize the design (trade studies in the design loop), (3) fit testing on enough people to establish the range of fit in the base size (prototype testing in the design loop), and (4) sizing loop full fit testing combined with fit mapping against the TP to determine the most cost-effective grade, cost-effective size assortment, and tariff. This entire process does not have to be repeated for every new product if it is used once to create an effective and sustainable fit standard. This is called a fit audit. After that, only limited trade studies and/or prototype fit tests are needed to verify that the new product meets the fit standard.

FIT AUDIT AND THE SUSTAINABLE FIT STANDARD

With a fit audit, we learn the range of fit in each size and the percentage of the TP who will fit in each size. We can use this information to make better decisions about the quantities to produce, buy, and stock. This means less waste. Furthermore, it enables us to develop size prediction algorithms based upon who will get an acceptable fit in each size which helps the consumer find the best size quickly and accurately. This has the potential to improve and increase online sales and minimize returns.

A fit standard holds fit and sizing constant for all styles. With a fit standard, we only change non-fit design elements, that is, those that do not affect fit negatively, when we produce a new version or a new style. Its function is to ensure consistency in fit and sizing between styles within a brand or company. A *sustainable fit standard* is a fit standard that can be followed and maintained for all products of a type and provide a consistently good fit for the TP with no unnecessary sizes and no size duplication. It is determined and defined after a fit audit. It is seasonless and may vary with each brand's identity. The goal is to consistently fit the same TP with different designs. This is illustrated in Figure 5.1.

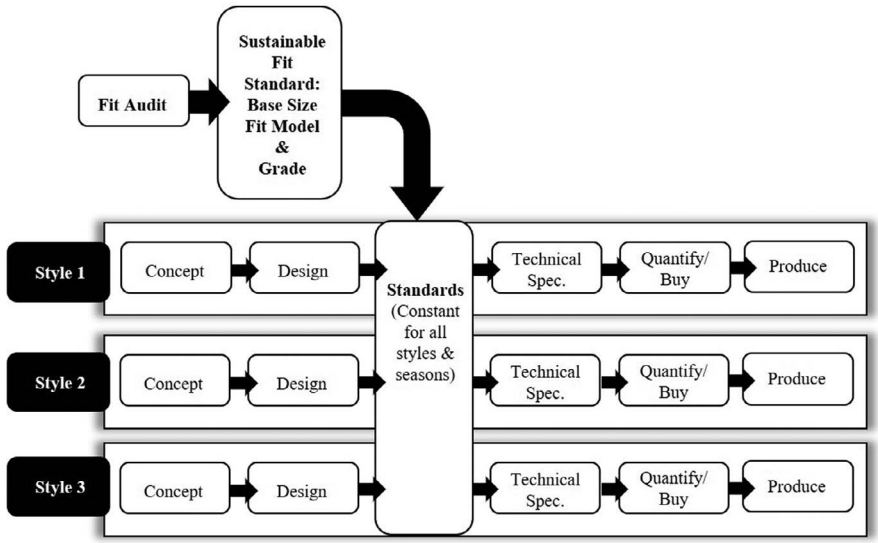


FIGURE 5.1 Fit audit and sustainable fit standard input.

The fit standard represents the people who buy or use that brand. The physical characteristics, body size, and shape, of these people don't change and this should be reflected in keeping the fit model and grading consistent, therefore maintaining brand consistency and identity.

"It is OK that each company's standard for quality and fit is different from every other company as this is part of the brand identity or brand DNA. However, it is essential that these selections are based on each company's knowledge of their TP (target population) and how to fit it. This includes how to create and validate the company's own standard base size blocks, fit model, and grade. In addition, each company needs a way to communicate its sizing to potential buyers and wearers. So many companies with which I have worked have allowed the design and/or production departments to play with grades with disastrous results. They should keep to execution and not reinvention. Fit manikins help keep the QC standards" (Janice Larsen, formerly Lead of Fit and Product Development, Lululemon, in personal communication 2023).

To maintain the sustainable fit standard for different styles and across seasons or product versions, it is necessary to validate and confirm that the standard is followed. For example, it can sometimes be difficult to know if a change to the design will affect the fit without a fit test. When in doubt it is useful to do a small trade study or prototype test comparing the before and after versions of a product. These can be studies with small samples of participants just to reassure ourselves that we have not strayed from our standard. The size of the test depends upon our confidence that we have met the standard and the risk we are willing to take that we might not have met it.

Brands may adjust their fit standards if they develop new visions of customer profiles or are expanding into new markets. If their standard is based on a fit audit, they can more easily adjust their standards because they know the range of fit in each size

and have good size prediction algorithms and charts. Not only will this accommodate the new profiles or markets, but also it enables them to communicate the sizes and changes to merchants to estimate their potential sales. Manufacturers may not want to share all the results, but they could share validated and effective size-of-best-fit prediction algorithms or charts that even consumers could use. Unlike size charts used by most retailers today that are unvalidated guesses that list a single value for each measurement for a size, the size prediction tools resulting from the sizing loop testing would produce a data-validated range of measurement values and a combination of measurement values that would be accommodated within a size.

Currently, most merchants rely only on past years' sales to inform their buying metrics, but it is not known why one style sells and another doesn't, and they can't have any metrics on the sale of sizes they didn't stock. If the previous year's sales included markdowns and returns from internet sales due to poor fit, then this year is likely to be the same. The retailer might still buy the brand if they made a profit despite the problems. Without fit data, neither the manufacturer nor retailer can improve their fit or know how to communicate sizing in person or online sales thus reducing the returns due to poor fit. This is where a fit audit of different brands is useful. In addition, if the retailer develops their own brand, then they need to understand their own products and sizing as well as how to communicate this to their customers. A sustainable fit standard enables all these goals.

Buyers or purchasing organizations who are not manufacturers also need to know who is fit and who is not. For this reason, some large organizations, such as the military or organizations that buy uniforms or protective equipment in large numbers, do fit audits to evaluate product sizing. Their populations can differ substantially from the population for which the manufacturer designs the products. Also, they sometimes have unique requirements for protection, tasks to be performed while wearing the product, and for integration with other products. These organizations benefit by doing a sizing loop full fit test with their specific COF, and with mapping against their TP. This enables them to buy only the sizes that are needed and avoid the cost of duplicate sizes, thereby reducing cost and waste. It also ensures effective integration of each item with people of all sizes and shapes and the other apparel and equipment needed.

Retailers and procurement agencies can do fit audits to gain knowledge of who is fit, who is not, and how to get the right size to the user. This helps with determining how many of each size to buy, whether the product will fit a given population or not, and helps the user get the right size the first time without the need for alterations or returns. In addition, knowledge gained from fit audits allows retailers and procurement agencies to compare competitor brands and/or can be used to develop tests or standards that will encourage improved products for their organizations.

Without a fit audit and a sustainable fit standard, we are blind. We are missing the following:

1. Body size or demographic data on the customer TP (who we are trying to fit)
2. Fit data describing who and how the company's key styles fit the TP
3. Fit model relationship to the TP
4. Fit model sizes that can effectively represent the base size

5. Manikin relationship to the TP
6. Manikin relationship to the fit model
7. How the product COF relates to the manikin, the fit model, and the TP
8. Knowledge of who besides the fit model will fit in the base size
9. How the grade relates to the TP
10. How the grade will affect fit in the large and small sizes
11. Knowledge of duplicate or overlapping sizes
12. Knowledge of large groups of potential customers for whom there is no good size
13. Knowledge of sizes we have that no one needs
14. Customer fit preferences

GRADING AND THE SWEETSPOT

Since actual people are used for fit models, the product will fit at least one person, the fit model, and will probably fit the people who are shaped like and near the fit model in size. This is the good part. However, it is just luck if it fits the area of the population where most customers are clustered and is in the best location for starting the grade. Unless they compare the model to the raw data distribution from the TP and determine with fit testing, the range of fit for the base size, the fit model location may be out of place making the whole size range offset from the maximum accommodation area. This is exactly what happens with most apparel companies today. In other words, they will have fewer sales and more wasted sizes.

When standards that control the quality of fit are missing or not used consistently each product can vary in who it fits. The variations can occur within the same company or even the same PD team. Companies often produce multiple labels and sometimes each of these labels may have different fit models. Sometimes no fit models are used. Sometimes if the regular fit model is on leave or sick then a substitute fit model is used. Without a fit audit and sustainable fit standard, there is no way to know if the different models are enough alike to represent the base size or if they will result in different sizes. Using one fit model this week and another next week may result in changes to fit one and then changing again to fit the other. This leads to wasted time and money.

Figure 5.2 shows the cost of having the base size (where the fit model falls) and grade at locations that are not quite right. It is a bivariate scatterplot of Waist Circumference and Hip Circumference values for 1262 women in North America that represent a TP. Sizes 2 through 20 from a sizing standard (ASTM International, 2015) are overlaid on the scatterplot along with a band called the Sweetspot. The Sweetspot is the area that would be most effective and accommodate the most people. This ASTM sizing standard has been updated since this version, but the even grade it uses (where Waist Circumference and Hip Circumference are changed by the same amount), is still in common use. The percentage of people contained within the ASTM grade (called the existing grade in the figure), for the sizes ±40 mm is compared to the percentage of people contained within the Sweetspot line ±40 mm. We see that the existing grade has 14.7% fewer people.

As the grade moves away from the base size it can quickly go into an area where there are few customers (see sizes 16 through 20) and miss the areas where there

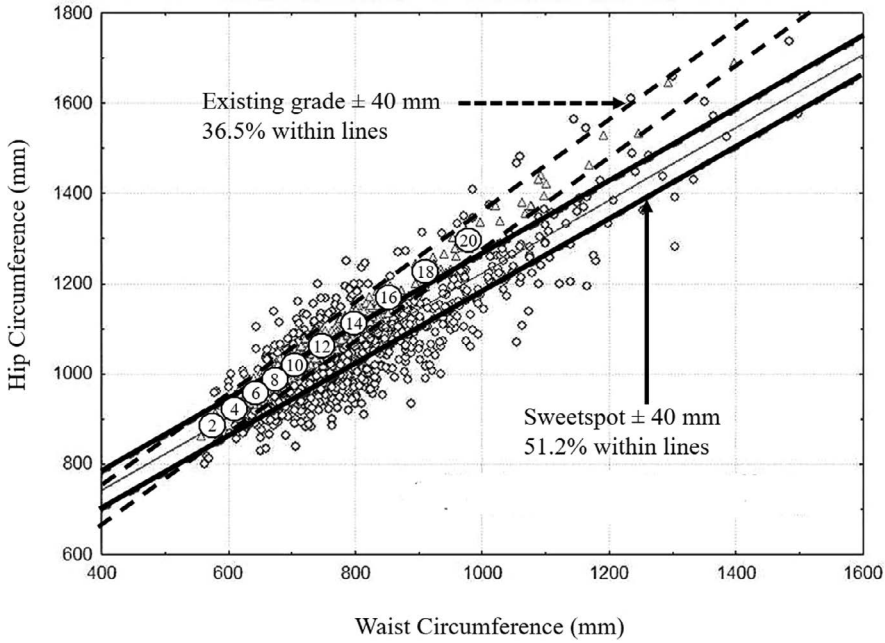


FIGURE 5.2 Waist Circumference by Hip Circumference scatterplot showing size range versus Sweetspot.

are many potential customers. When the grade goes into an area where there are few customers this can cause the smallest and largest sizes to end up on the sales rack more often than the middle sizes. When this happens companies will often delete sizes rather than explore if they have a grading issue. This reduces sales even further.

Sometimes the relative growth of one part of the body is different from other body parts. The study of the relative growth of parts of the body in relation to the growth of the whole is known as allometry. When we grade by matching the body growth in different body areas, we call it allometric grading. The Sweetspot shown in Figure 5.2 is the allometric grade for Waist Circumference and Hip Circumference. The allometric grade for other body dimensions is determined using regression prediction from the key dimensions.

In Figure 5.3, we show just the base size 10 versus the Sweetspot. Here we see that it is off-center or larger in the hip than the Sweetspot center-line. The other sizes are determined by adding and subtracting from the base size, so it is important to have a good base size location. In this size range when the base size location was combined with the even grade, the size range ended up having nearly 15% fewer people in the same number of sizes than with a better base size location and allometric grade. With a fit audit, we would have known this and been able to correct it in the fit standard.

In Figures 5.2 and 5.3, we used a ±40 mm range to simulate the range of fit, but the true range of fit is determined by the product and varies from product to product. The true range of fit is estimated by fit testing. We illustrate this in Figure 5.4 where

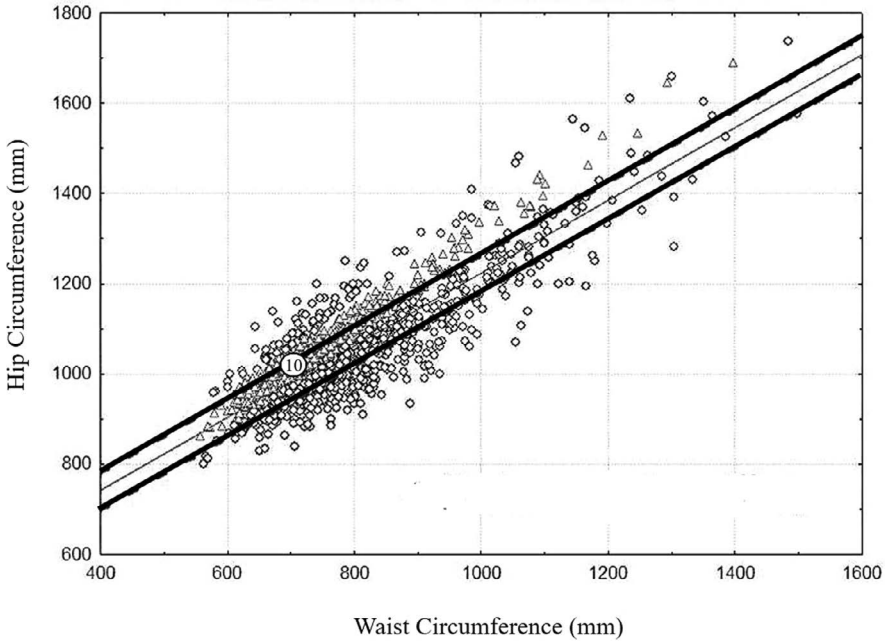


FIGURE 5.3 Waist Circumference by Hip Circumference scatterplot showing base size 10 versus Sweetspot.

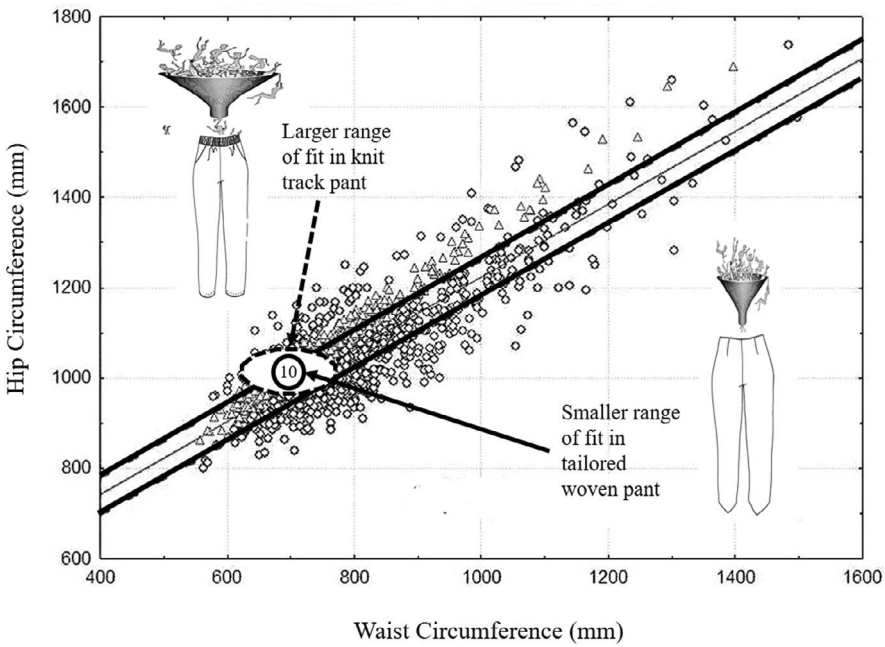


FIGURE 5.4 Scatterplot showing fit range differences with different garment types.

we show the base size 10 from the previous two figures with the range of fit ellipses varying between a tailored woven versus a knit track pant. The product acts like a funnel that captures some people and loses others. The size and shape of the track pant range of fit ellipse, or funnel, is larger and shaped differently to the range of fit ellipse for the woven pant. The track pant has stretchy knit fabric and an elastic waistband so it will fit a much larger range of people than the same size pant with a more inflexible woven fabric without elastic in the waist. The elastic waist also means the range of fit in the waist is larger than the range of fit in the hip. For the woven pant, the waist fit range and the hip fit range are very similar, hence the fit ellipse is more circular.

Sometimes different product types will require different size ranges. A stretch pant might be labeled small, medium etc., and come in fewer sizes. Sometimes the company might leave the sizes as 8, 10, 12, etc., and allow much more overlap in the sizing. Fit testing is needed to determine what the range of fit is so that the best combination of sizes can be determined. We don't have to go through the whole SPEED process for every new product if we establish a fit standard for the different product types using a fit audit. After that, we may only need a small trade study or prototype test when the first product sample is produced to verify the product meets the standard.

BENEFITS OF THE SPEED PROCESS

The use of the SPEED process varies depending on the history of the product, the existing knowledge base, the confidence we have in the sizing and fit standards we are using and other factors. Some examples of specific issues for mass-produced products and suggestions to resolve them are listed in Table 5.1.

CASE STUDIES

A case study is a detailed example in a real-world context. In the theoretical world things always go as planned. In the real world, not so much. We have prepared a few real-world examples to better explain our process, and how it looks when used in the real world. They illustrate how, with careful planning, even when things go wrong, we can often still glean the information we need to make good decisions. They also illustrate the need for pilot testing, measuring everything to ensure things are within spec and carefully planning before we start. The case studies include:

- Manufacturer/retailer design, sizing, and fit standard development
- Assessment for purchasing existing products aided by 3D scanning
- Assessment for purchasing tariff
- Prototype test to determine the correct alteration

TABLE 5.1
Opportunities for Improved Fit

Problem	Potential Solutions
Sizing and fit inconsistencies across styles	• Trade study of base sizes, check against fit standard • Prototype test of base sizes and grades (large and small sizes)
Potential inadvertent size duplications, such as petites and regular ranges fit the same people	• Prototype test of base sizes and grades
Fit issues in base size	• Trade studies test of base size
Grade is off in some areas causing poor fit in a specific body area (such as sleeve)	• Prototype test of large and or small sizes • Regression analysis (see Chapter 3 and Case Study 4)
Slow product development due to excessive fit alteration cycles	• Pilot test COF • Full fit audit and sustainable fit standard development or revision
Poor sales due to customer fit dissatisfaction	• Sizing loop full fit test with sample from TP
Size confusion due to different sizing for different brands	• Trade study of base sizes, check against fit standard • Prototype test of large and small sizes to check for grade differences • Check grading specifications against TP
Excess product in some sizes (markdowns), and insufficient product in others (missed sales)	• Fit audit using sizing loop full fit test and map against TP • Check size prediction algorithms and charts • Check tariffs against the TP
Insufficient customer confidence in size charts (missed sales/excessive returns due to poor fit)	• Sizing loop fit test with verified size prediction algorithms and improved size communication
Don't know which vendor's product to purchase	• Sizing loop full fit tests using full TP sample to compare competitor products • If manufacturers did a fit audit, compare competitors fit standards, specification sheets, size prediction algorithms and charts
Don't know how many of each size to purchase	• If manufacturer did a fit audit, use their size prediction algorithms/charts and map against TP • Sizing loop full fit test to produce size prediction algorithms and size purchasing tariffs
Functionally duplicate sizes	• Sizing loop full fit test
Wrong sizes for TP	• Sizing loop fit test with sample from TP
Lack of integration of multiple items worn together	• Develop and pilot test integrated product COF • Sizing loop full fit test using integrated system sizing COF

Case Study 1: Manufacturer/Retailer Design, Sizing, Tariff, and Fit Standard Development

Company A was both a producer and a retailer of women's apparel. They had already been producing pants in a range of sizes and had them in their retail stores. They had received feedback that customers were complaining about the fit, even in their

best-selling items. We were asked to help them understand what the issues were and to improve their sizing. One of the first items we assessed was their best-selling pant that was already on the market and in stores.

The company had a fit model and a manikin representing their base size case, and they had an established grade for five sizes. This means they had completed the first three steps in the design loop, case selection, design, and prototyping. However, they did not know how representative their base size case was, nor how well placed it might be for their targeted market. They also did not know the range of fit within a size, nor if their grade was well placed for the market. They did not know who they were fitting and who they were missing with their size assortment. This is typical in today's apparel industry. This study answered all these questions and was completed in six months. While it did not follow the process in order, it essentially involved all the parts of the SPEED process.

The study outcomes improved the following: (1) the base pattern, (2) the size assortment, (3) the grade, (4) the cut ratio (tariff), and (5) the COF. The result was an improved fit in all sizes and an increase in the targeted market (TP) coverage from 45.5% of the TP to 63% (a 38.5% increase in potential sales) with a shift of the base size (added 31.8%) and adding one size (6.7%). This study also gave the company a new capability to make informed decisions about sizing for different and new markets.

Inputs

The study began by establishing the inputs: (1) product requirements and design concept, (2) resources, and (3) TP demographics and anthropometry. The product requirements were expressed in the COF document.

All company personnel who were involved with some aspect of fit were tasked with participating in its development. It is important to have consensus and this is only achieved by having all the stakeholders together so they can debate any areas of disagreement. A meeting was held with all of them to draft the first COF and to witness the pilot test of the COF to finalize it. The pilot test was used because different people who are technically the same size can have different fit issues. We needed to identify them and classify them as pass or fail before we began testing. Reviewing how the product fits on a variety of people helps us refine our definition of fit so it will apply to everyone. Some aspects of fit can be seen in a photo, but comfort-related issues cannot be seen, so we need subjects to tell us.

Those present were senior management (who have the final say on fit), design, PD, buyers/merchants, and the regular fit model who also tried on the garments. Each idea was tested. Then everyone thrashed out what they thought was a good fit in each area of the garment. We did not move on until we reached a consensus. Specifically, each fit statement had metrics that everyone signed off on. So, for example, we decided how long was too long in the crotch and how to measure it. We also decided what was a pass or fail. The questions to be used in the fit questionnaires were based directly on the COF and accounted for the subject's opinion overall as well as in a variety of pant areas so we could tease out problem spots.

To assist us, the regular fit model tried on sizes that were both too small and too big. We also had a second model try on the sizes to test ideas. This took about one day. Each decision was documented. This was written up and presented for initial signoff to the project manager.

TABLE 5.2
COF Definition of Fit Pass or Fail

Criteria	Pass/Fail
The customer says they wouldn't purchase the garment because it is deemed too uncomfortable to do normal activities like standing, sitting at a desk, or driving a car.	Fail
When the garment is not deemed too tight or loose or too short or long in any area	Pass
The hip curve is deemed extremely too high or extremely too low.	Fail
The waistline and/or the hip is deemed extremely too tight or extremely too loose.	Fail
The *crotch* is extremely too tight (as assessed by the subject).	Fail
The *crotch* is extremely too long (greater than 30 mm pinch).	Fail
There are three or more whiskers (horizontal folds at the front crotch).	Fail
There are six or more gluteal folds (horizontal folds at the back crotch)	Fail

The COF included two overall fit assessments: one from the customer (subject) and one from the investigator. The former indicated subjective preferences and the latter the more standardized fit assessment based upon the documented COF. The investigator's overall fit assessment was used to assign the size of best fit and considered what the subject said about comfort and mobility. The subject was asked what size they thought was best too. Disagreements were allowed and very valuable because when the customer disagreed with the investigator, areas were highlighted in the COF that the manufacturer didn't know were important. Why and where they disagreed was recorded.

All the definitions of too tight, loose, short, long, etc., were listed in the COF with pictures, scans, and measurement scales. We cannot show the specifics due to the need to protect the company's proprietary information. In general, the COF definition of fit pass or fail is as seen in Table 5.2.

The customer answers were recorded using a questionnaire for each size they tried on. The questions are presented in Table 5.3.

Scores of 1–4 were considered a fail and scores of 5–10 a pass. The subject was also asked:

1. Would you purchase the garment given the fit for normal activities? Yes or no
2. If not, is it because of the fit or style? Fit Style

TABLE 5.3
Customer Questionnaire

On a scale of 1–10, with 1 being extremely bad and 10 being extremely good how would you rate the following	Score (Circle One)									
a. Waist comfort	1	2	3	4	5	6	7	8	9	10
b. Hip comfort	1	2	3	4	5	6	7	8	9	10
c. Crotch comfort	1	2	3	4	5	6	7	8	9	10
d. Waist location	1	2	3	4	5	6	7	8	9	10
e. Hip location	1	2	3	4	5	6	7	8	9	10
f. Overall fit	1	2	3	4	5	6	7	8	9	10

TABLE 5.4

Questions for investigator	Score (circle one)				
1. How is the waist fit?	Extremely tight		Good		Extremely loose
	–2	–1	0	1	2
2. How is the hip fit?	Extremely tight		Good		Extremely loose
	–2	–1	0	1	2
3. How is the hip location?	Extremely low		Good		Extremely high
	–2	–1	0	1	2
4. How is the waist location?	Extremely low		Good		Extremely high
	–2	–1	0	1	2
5. How is the crotch fit?	Extremely tight		Good		Extremely loose
	–2	– 1	0	1	2

We asked about the style in order to give them the response options, so we could separate fit issues versus style preferences. An example of a style preference might be they would prefer a different color. Sometimes they want to talk about the style and that can bias their answers so if we let them talk about it, we get a better fit score.

Some of the investigator questions were recorded on a five point scale from –2 to +2, with the minus values meaning tight, or short, and the plus values meaning loose or long, and 0 being good or just right. The questions are presented in Table 5.4.

The data collection team was comprised of six people:

- Two trained measurers
- Two fit assessors
- One "meet-and-greet" person who
  - greeted the arriving subjects
  - coordinated the appointments
  - and gave the participant a gift at the end
- One supervisor who
  - was a criterion anthropometrist
  - resolved difficult to measure anthropometry or close judgment calls about the fit
  - collected body scan data and photos
  - conducted the data analysis

The supervisor also covered for team members during a break and filled in when one was sick or running late which made the team run smoothly and on time.

Test subjects were recruited using advertising and by providing incentives. The goal was to collect anthropometric measurements on 150–200 people in total, and we measured 170. This provided the company with a database for future use.

The company provided three facilities for data collection, each having parking, climate control, good lighting, and privacy for getting changed. These were booked for the duration of data collection. In addition, the company had its manikin and fit model measured and scanned at the Head Office.

Equipment included two anthropometers, tape measures, tablets for data entry, and sundry other equipment like photography and scanning equipment. We used a handheld scanner as well as a Cyberware WBX™ for the fit model and manikin. Software used was Google Forms™, Microsoft Office™ suite, SPSS®, and Blender™.

The company also had garment-specific anthropometric measurements they wanted to be collected and we added a few to these for cross-comparison with existing databases. A total of 56 measurements were included and these were documented in an anthropometry manual. We took anthropometric measurements of each subject, in addition to the fit scores in all the fit tests and trials.

The company had only collected basic demographics previously which included name, email, and geographical location of purchases. However, they had an "ideal" profile of their customer that included females, aged 18–60 and professionals, with a certain income band. They also estimated they would be relatively fit, so we limited them to women with hip measurements below 1300 mm and/or waist measurements less than 1200 mm. To characterize the TP it was decided to include some additional more specific information including age, race (as this can affect fit), occupation, exercise frequency, self-reported height, weight, bra size, usual clothes size, and size in the brand. This questionnaire was deployed to the retail outlets. Separately we asked if we could have contact details as potentially this database could be used to recruit participants, but this was explained as part of the privacy policy and permission status was recorded on the consent form. This demographic questionnaire was a self-reported questionnaire. We also had a version for each subject recruited which we the investigator team filled out. These could be merged later to expand the database.

Design Loop

For the first pilot test, we were evaluating both the COF and the test procedures. We tested the pant on five employees as subjects. We checked the overall anthropometry and fit questionnaires. We noted efficiency, errors, inconsistencies, and clarifications. In addition, we checked how our COF worked in field conditions and timed the tests to see if we could improve the workflow. We checked the training and performance of the team.

We made the necessary minor adjustments that included changing the order of some questions to make them flow better. We timed the workflow and made sure all the measuring and fitting could be completed in one hour per subject.

We tested and finalized examples of how we scored fit in the COF. For example, we had a subject that was wearing a pant that we would have described as slightly too big. This made the pant waist slip down the torso, to rest on a circumference below the waist. If the pant waist was measured in that location, then the waist circumference would have been rated as perfect and the crotch length as slightly too long. However, if the subject adjusted the waist location to her actual waist, then the waist would have been rated slightly too loose and the crotch length rated perfect. This needed clarification in the COF. We decided that for circumstances like that, we should adjust the pant to the correct waist location and code its fit problem as slightly too loose in the waist and perfect for the crotch length. By doing this, the fit coding

pointed to the correct fit problem as too big (the subject needed a smaller size) and made the crotch length a separate issue. This was one of several clarifications to the COF. We documented final COF decisions in a fit manual.

The products of this first pilot test were: a prototype that was within spec for testing, a tested COF, and validated test procedures and manuals. Once verified the final COF was officially signed off.

We next began a prototype test. The sample test garments were sizes 34, 36, 38, 40, and 42. Each subject was tested in every size they could don. The process was:

1. Greet the participant, explain the project, provide an information sheet, and obtain consent in writing
2. Complete the demographics questionnaire
3. Measure the subject in standard bike shorts and tops over their own underwear
4. Complete the fit assessments in multiple sizes

For the fit testing, each participant started with their estimated size of best fit. This was determined by reviewing the size they normally bought in pants from the demographics questionnaire and was reviewed visually with a quick try-on. This starting size was recorded, and the fit assessment questionnaire was completed. Then each subject was tried in a size smaller and the fit assessment was repeated. If there was an even smaller size this was assessed next. When the progressively smaller garments no longer were don-able then we went to size one larger than the starting size and assessed it. If the participant did not receive a fit in any garment they were recorded as none-fit. We also recorded what size we thought they would fit if they needed a size that was not available. If they achieved a passed fit in multiple garment sizes, then this was recorded along with selecting a single size of best fit.

In terms of experimental design, it might have been better to randomize the order of the presentation of the sizes to the subject. However, this is much more difficult to execute and keep accurate track of which size is being assessed. We decided that could lead to recording errors and would be much worse than any bias that might exist because of the lack of randomization in the order.

The size presentation order was controlled using Google Forms™. The starting size was recorded and labeled as size X in the form. X − 1 was listed as the size smaller, and X + 1 as the size larger. The first question with each size was "is the size donnable?" If the subject could not don the pant it was recorded as a fail and we moved to the next size or the next subject.

The original plan had been to do one full fit test (sizing loop), which we refer to as a fit audit. However, this assumed the garment's proportioning was good for the sizes already in the shops. In fact, after the first 30 subjects were tested it became clear that there were major proportioning issues. The subjects were saying they were afraid to sit down in case they broke the zip, and the pant was so uncomfortable that it was dubbed a "standing only" pant. These were the critiques of the best fit size!

We missed that there was a general fit problem when we did the pilot test because we did not have enough subjects. Consistently across all the sizes, the pant was rated good when the subject was standing but a significant group of these subjects reported

substantial comfort issues when they tried to sit or bend. We determined that the overall crotch length was good and that the problem could be fixed by shortening the front crotch length at both the front waist and front crotch fork, and simultaneously lengthening the back crotch length at the back waist and also the inside leg fork. This left the overall length unchanged but improved the sitting and bending comfort. The amounts were small, so the pant looked the same.

This general fit problem required alteration and the creation of new pant prototypes. Testing was stopped and resumed when the altered pant prototypes were ready. Since this was considered a new pant, the results of the two fit tests couldn't be combined. Therefore, the test with the first 30 subjects was used as a prototype fit test (a design loop test), to guide the design. It resulted in two adjustment recommendations for the base size of the pant, the one we described above as well as to slightly elasticize the waistband.

Next, prototypes were made incorporating the recommendations and tested with a small trade study test with eight subjects. This compared the original base size (size 36), to the new base size to test the design modifications. This is a paired comparisons test, and the Wilcoxon Signed-Rank test was used and indicated a significant improvement in fit in the new base size. Seven out of eight (87.5%) of the subjects had equal or improved fit with the new changes. Five of these had improved enough that they indicated they would now be comfortable enough to buy the pant, whereas previously they would not. The one person who preferred the original pant felt the new base size was too loose. Examining her waist and hip proportion it was determined that with the changes she would now probably be a better fit in the next smaller size, the size 34. That size was not available for this test so we could not confirm, but if true this would make a happier customer as well. Most women like to know they fit in a smaller size! This indicates that the change in the crotch was a good change and should be kept. The new pant was better than the original. This gave us enough confidence to move to the sizing loop.

Sizing Loop

Once all new pant prototypes had been produced, we did a full fit test. We followed the same procedures used in the prototype test. We measured and assessed an additional 40 people.

Prior experience had shown that Hip Girth and Waist Girth were good predictors of the size of best fit for a pant, so we began by plotting the best fitting size on a Hip Girth by Waist Girth bivariate chart of our fit test sample. We immediately realized there was something wrong with the size 38. As can be seen in Figure 5.5, while there is a clear range of fit for the base size, size 36, and for the larger size, size 40, there are many no fits in the area where the size 38 should be. We measured the size 38 prototype and found that it had a sewing error that had stretched the back crotch and made this outside of tolerance. This was something we should have verified before testing, and it shows the importance of measuring the prototypes and verifying that they are within specification before beginning testing.

We measured the other sizes and found they were all correctly cut and sewn, and since we had enough subjects to determine the range of fit for two sizes, sizes 36 and 40, we felt confident in the assumption that the range of fit for the correctly sewn size 38 would be similar. Once we had a good estimate of the range of fit in our sizes, we

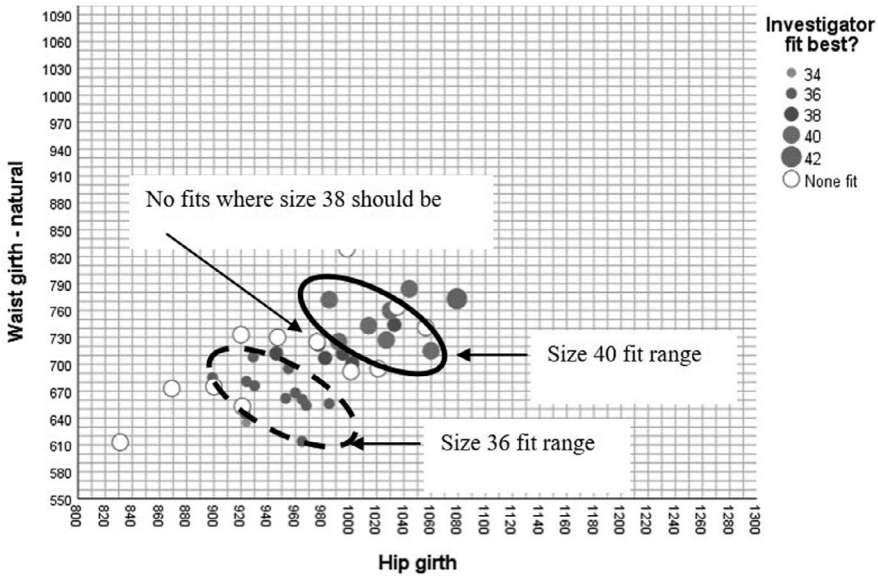


FIGURE 5.5 Hip Girth (mm) by Waist Girth with best fit sizes indicated.

began a cost/benefit analysis (CBA). The CBA begins with who we will fit in any size assortment, and who will be missing, and the percentage of people who will get a good fit in each size. With this information we can compare the percentage of potential new customers or users we might gain, current customers we might retain or lose, against the costs of producing or stocking the size. For this, we need to understand how our sizes accommodate our target market. There aren't enough test subjects in the fit test sample to see how these sizes will accommodate the TP. To increase our TP sample size, we added anthropometry data drawn from another (proprietary) data set to create the full TP sample. The raw data was then tailored to represent the TP. The gender was female. Anyone with a hip measurement above 1300 mm and/or a waist measurement of greater than 1200 was excluded. This reduced our full TP sample size to N = 521.

Next, we plotted the size 36 and 40 ranges on the full TP sample. This is shown in Figure 5.6. This indicates that the size range has a Waist Girth that is a little below the ideal coverage area.

Moving the sizes up (making the base size Waist Girth larger by 20 mm), will increase the TP coverage, as shown in Figure 5.7. Here we show all five sizes before and after the shift. This simple shift to a larger waist gains 6.5% more people (representing a 14.3% increase in potential sales). This change is in the base size to locate it over the Sweetspot. We haven't yet examined the grade.

Next, we examined the size assortment. We also see in this figure that there are many more larger people than smaller ones. So, we examined the addition of a larger size, size 44, and dropping the smallest size, size 34. This is shown in Figure 5.8. This set of sizes increases the TP coverage to 59.6%. This combination is a 14.1% increase in TP coverage and represents a 31.8% increase in potential sales with the same number of sizes. The company decided to add the size 44 and keep the size

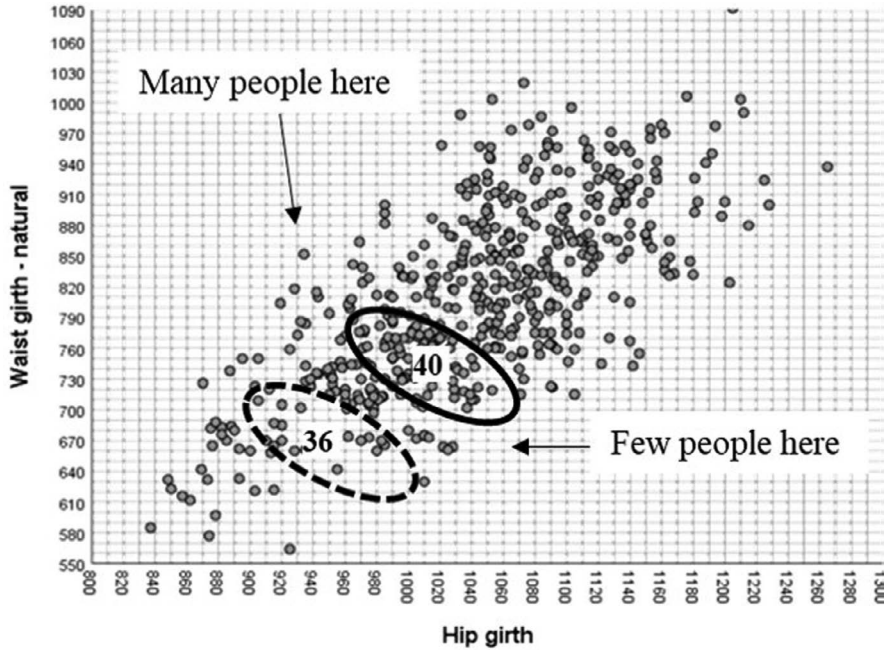


FIGURE 5.6 Two sizes plotted on full TP sample. Units are mm.

34, for a total of six sizes. This represented a 63% coverage of the TP and a 38.5% increase in potential sales. They felt that their customers were used to having smaller sizes. They will need to advertise that they have added a size to attract the new business. The final set of sizes is shown in Figure 5.9.

The company would have liked to increase the size of the grade (change it from a 4 cm to 5 cm grade) but the testing showed that if they had gone down that route there would have been people in-between sizes and they would have lost sales, so they decided against this course of action.

The company used the results to adjust their company fit standard for pants. The new fit standard would use the same fit model and manikin but make the Waist Circumference 20 mm larger and the corresponding COF 20 mm larger in the waist. The block pattern for the pant, would also have 20 mm added to the waist. The fit standard would have the same grade and add one size which would extend this grade out to the size 44.

To create the tariff indicating how many of each size to manufacture and stock we first created a size selection chart. This is a chart to help the customer select the best fitting size. This would be the garment they select, so counting how many are in each of the categories in the chart would be a good estimate of how many of each size garment to produce and purchase, that is tariff.

For practical purposes, counting for the tariff and communicating with customers, the size selection chart cannot have overlap in the size categories, and it needs to have squared corners that follow along the grid lines of the bivariate chart so that the user can follow the grid lines to find their size. The process to create the size selection chart is shown in Figure 5.10. We started by creating rectangles with the size ellipses (part a.) that will make it easy to count the number of people in the size. Then we outlined all

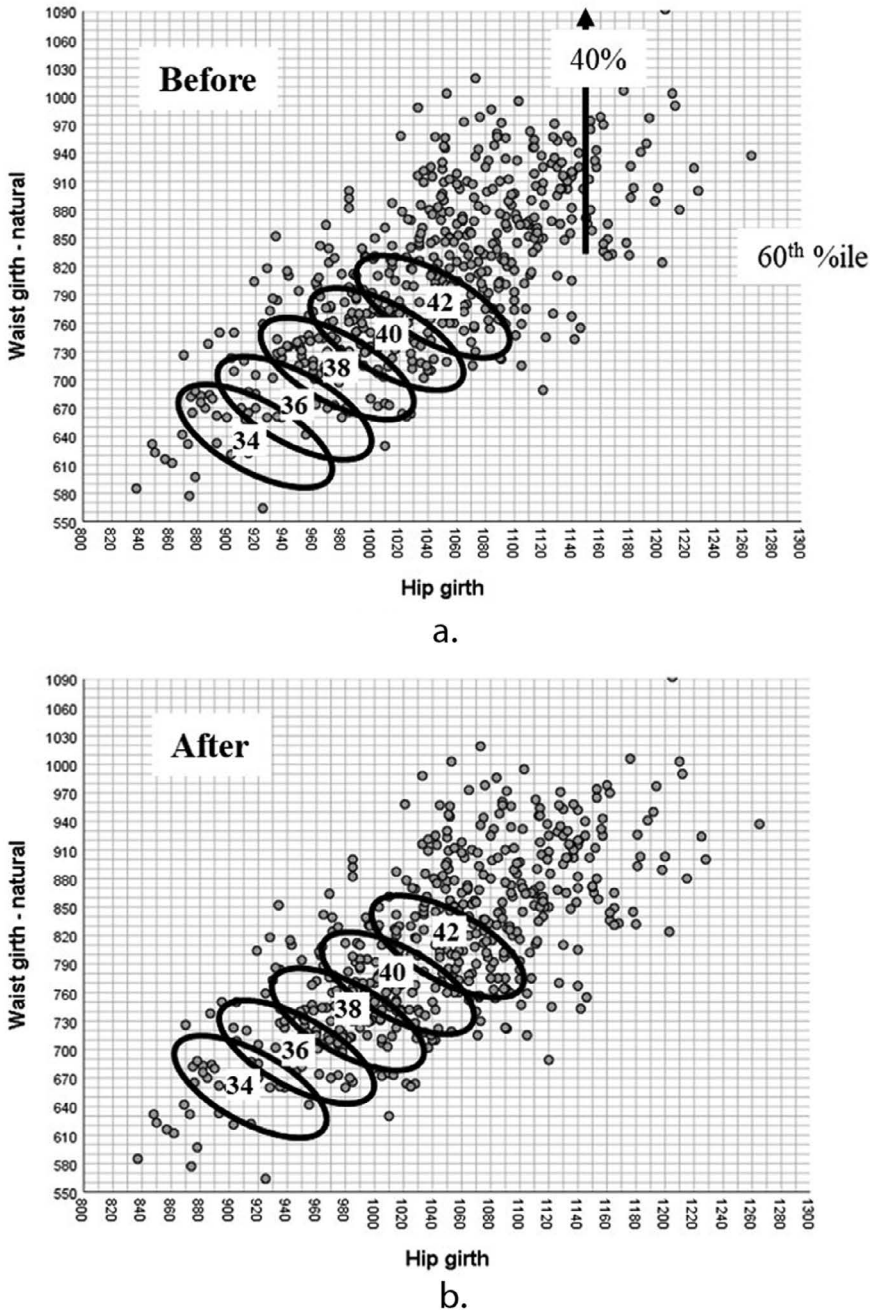


FIGURE 5.7 Effect of increasing Waist Girth. (a) Before: 45.5% fall within the five sizes. (b) After: 52% fall within the five sizes. Units are mm.

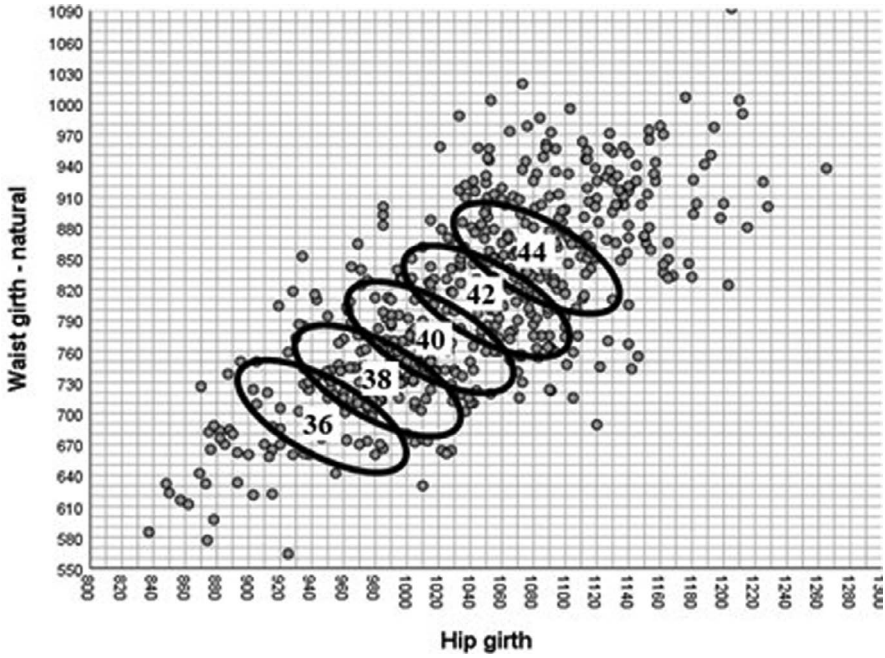


FIGURE 5.8 Adjusting sizes to better fit the population.

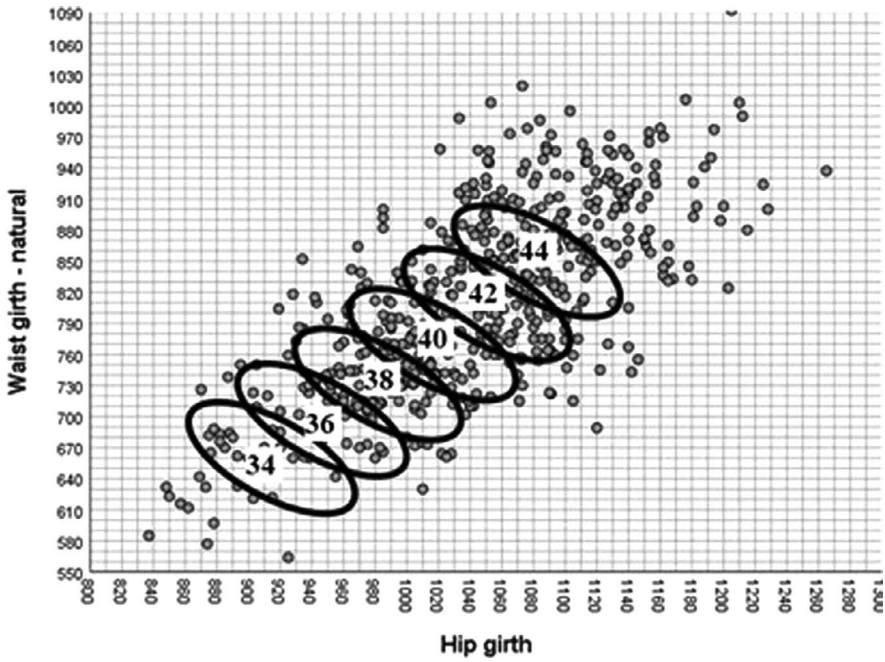


FIGURE 5.9 Final six sizes.

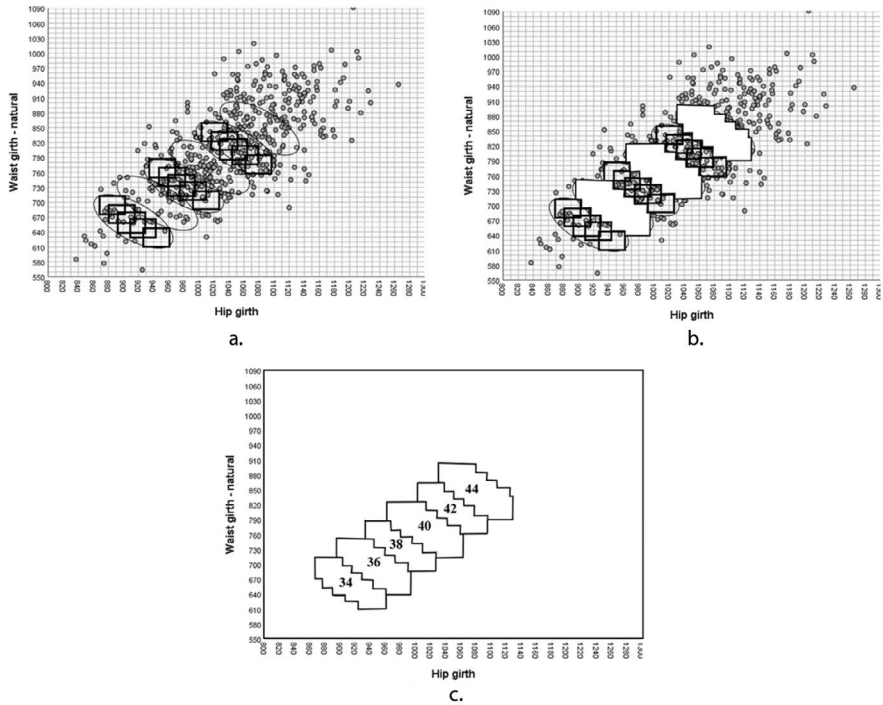


FIGURE 5.10 Creating size selection chart. (a) Fill in ellipses with rectangles for counting numbers in each size. (b) Outline rectangles in each size. (c) Remove grid and add size numbers.

the rectangles in each size as shown in part b. Then we removed the grid with the subjects, as shown in part c, and add the size labels. A plain grid can be added if desired.

Once we had the size selection chart, we used it to count the number of people who would select each size, calculate the cut ratio, percent of the total TP, percent of the people in the size range, and the tariff. This is shown in Table 5.5. The tariff was calculated as the number of units to produce or stock per 10,000 products.

TABLE 5.5
Determined the Percentage Accommodated in Each Size

Size	Number	Cut Ratio	% of Total	% Within Size Range	Tariff
34	19	0.036	3.6%	5.79%	579
36	44	0.084	8.4%	13.41%	1341
38	50	0.096	9.6%	15.24%	1524
40	117	0.225	22.5%	35.67%	3567
42	63	0.121	12.1%	19.21%	1921
44	35	0.067	6.7%	10.67%	1067
None	193	0.37	37.0%		
Total in Size Range	328	0.63	63.0%	100.0%	10,000
Total	521	1	100.0%		

Case Study 2: Assessment for Purchasing an Existing Product Aided by 3D Scanning

An Ambulance Service (AmSe) estimated future recruiting and found that although in the past the recruitment was 80% males and 20% females there was a large increase in female paramedic graduates. The AmSe recruits independently of gender so they anticipated that recruitment would soon reflect the gender proportion in the university paramedic cohort (50:50). In keeping with this anticipated changing workforce demographic, the AmSe reviewed uniform fit for women. It was found that there were substantial fit problems with the pants and even some female worker injury claims involving pant fit. These women were having difficulty with essential activities reportedly due to a tight fit in the crotch region. These activities included bending forward over patients, squatting when the patient was on the ground, and kneeling with one knee down on the ground and the other up (the Knight's Pose).

To better understand the fit problem a prototype fit test was done. The prototype selected for testing was the most used cargo-style pant. This pant was intended to reach the waist or sit slightly above the waist worn with a duty belt. It came in two size ranges, male and female. Males had a 27 size range assortment and women had a 10 size assortment in this pant. The manufacturer noted the female pants were not being purchased so they offered more limited sizes than the male size range. They did not investigate the reason for the poor sales of the female pant but assumed it was because there were fewer females in the AmSe. However, an estimated 80% of the women in the AmSe reported wearing the male pants. This seemed to indicate there were proportioning and sizing issues with the female pant.

The TP for the study was female emergency service paramedics and the assessments included 3D scanning with and without the pant, and software overlays of the scans. The 3D scanning and scan overlays allowed us to clearly visualize the issues and recommend solutions. This is an example showing how small prototype fit tests can inform the design.

Inputs

A scan of the pant and a photo of the pocket detail is shown in Figure 5.11. It was a polyester-cotton twill with minimal stretch in the ambulance corporate colors (gray with hi-visibility stripes and branding). It had a front opening fly and button at the waist, with a one-quarter elastic waist (two waist panels by one-eighth) to aid in comfort and increase the range of fit in a size. It had six pockets, two rear patch style, two front in-seam cut-away pockets (like jeans), and lastly two off-set cargo patch pockets: one on each thigh. Pockets are needed for numerous items including security ID badges, vaccination cards (for entry to aged care facilities), pager, radio, pens, torch/flashlight, spare mask, and mobile phone, at minimum. Car keys hang from the duty belt. Extras include safety glasses, PPE such as gloves and spare masks, and other items like lip balm. So, the ability to include these things in the pockets was part of the COF, and the subjects were tested with items in all the pockets.

Since a key issue with these pants is movement and performance, the COF included having the subjects kneel, squat, and bend before assessing tightness,

FIGURE 5.11 AmSe pant scan and pocket detail.

looseness, or overall fit. One of these poses, called the Knight's Pose, was scanned for fit visualization purposes, and was used to score tightness and bunching in the crotch area, which had been identified as a particular problem for females. The COF included two sets of fit scores as shown in Table 5.6: one from the customer/subject and one from the investigator. The former will indicate customer preference or COF and the latter the more standardized fit assessment based upon this documented COF. The investigator asked the subject comfort questions before rating the fit. Length adjustments were not scored for fit because the pants could be made long and hemmed.

Relevant anthropometric measurements for the pant when the paramedic was standing included: waist circumference, hip circumference, waist height, hip height, crotch height, crotch length, AND squatting measurements: thigh circumference, knee circumference, back crotch length. The subjects were scanned with the pant and without (in scanning garments) and the scans were overlaid using Blender™ software.

Demographic data included race, age, years in the job, rank (Intern, Clinical Instructor, Intensive Care Paramedic, Extended Care Paramedic), number of pregnancies, fitness level (number of times workout per week and type of workout), the current size of clothes worn, and self-reported height and weight.

AmSe defines their working population (the TP for this test), as females aged from 20 (all were university graduates) to 60 years old. Paramedic interns must pass fitness tests including a 40-kg squat lift to qualify for the internship. There was no

TABLE 5.6
Fit Scoring for AmSe Pant

Questions for Subject	Score (Circle One)				
1. Would you purchase the garment given the fit for normal activities?	0 = No 1 = Yes				
2. How would you rate the overall fit?	Extremely bad				Excellent
	–4	–3	–2	–1	0
Questions for Investigator	**Score (Circle One)**				
3. How is the waist fit?	Extremely tight		Good		Extremely loose
	–2	–1	0	1	2
4. How is the hip fit?	Extremely tight		Good		Extremely loose
	–2	–1	0	1	2
5. How is the hip vertical location?	Extremely low		Good		Extremely high
	–2	–1	0	1	2
6. Does the phone or gear in the front pockets interfere with bending?	Extremely bad				Not at All
	–4	–3	–2	–1	0
7. How is the waist location?	Extremely low		Good		Extremely high
	–2	–1	0	1	2
8. How is the buttock compression in the Knight's Pose?	Extremely tight		Good		Extremely loose
	–2	–1	0	1	2
9. How is the front crotch bunching?	Extreme				None
	–4	–3	–2	–1	0
10. How is the overall crotch length?	Extremely short		Good		Extremely long
	–2	–1	0	1	2
11. How is movement?	Extremely bad				Excellent
	–4	–3	–2	–1	0
12. How would you rate the overall fit?	Extremely bad				Excellent
	–4	–3	–2	–1	0

difference in the male and female fitness tests. Height and weight data collected from new recruits over the previous 12 months revealed that the height range for female recruits was: minimum 156 cm and maximum 186 cm. Minimum weight was 50 kg and the maximum weight was 90 kg. Recruits do not reflect the whole AmSe population and provision needed to be made for older paramedics who aged "in the job" up to age 60 and some of these had increased weight although they still had to pass annual fitness tests.

Design Loop

A prototype fit test was done with ten female paramedics who were in the mid-range for female sizes and normally wore either the size 77 or the next size up or down. They were fit assessed in every pant they could don. All of them were fit in the size that was deemed the best possible fit from the sizes within the two size ranges. None of the pants fit well, but it was determined that no other sizes would fit better. The testing was completed in one day.

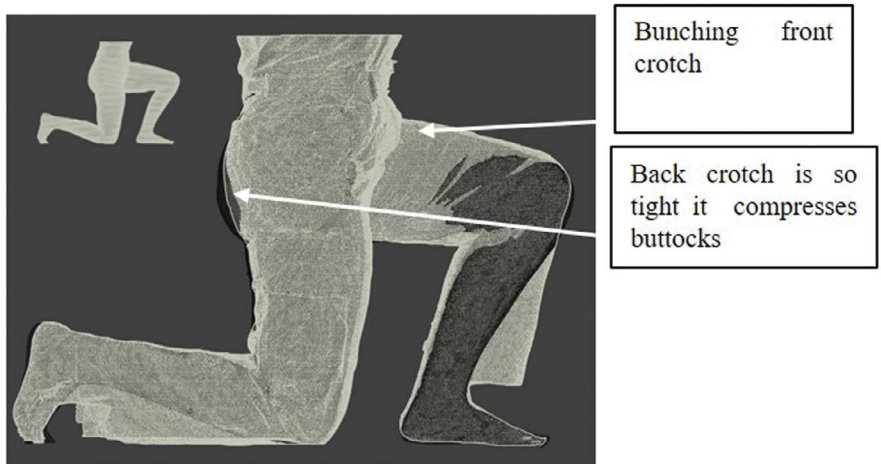


FIGURE 5.12 Female paramedic in men's size 77R overlaid on near nude scan, Knight's pose.

For all the pants the front crotch was too long such that it caused major bunching even when standing, and the back crotch was too short such that it caused compression of the buttocks and thighs and restricted movement. The men's pants bunching issue in the front crotch was surprisingly less than the women's pants, which may have been a factor causing female paramedics to purchase men's pants. However, even though it was better, it was still a problem. This issue was clearly identified by examining the overlaid scans.

An example showing a female subject in the men's size 77R is shown in Figure 5.12. The near-nude scan by itself is shown at the upper left to help visualize the overlay. The overlaid scans are at the bottom right. Note the compression on the buttocks from the extremely tight back crotch and there is also bunching at the front from the front crotch being too long.

The same subject is shown in the men's size 82R in Figure 5.13. Size 82 is more comfortable in the back but less comfortable with more bunching in the front crotch. Neither size fits well.

It is important to try the subject in multiple sizes. Often the incorrect assumption is that since the 77 was too tight the larger 82 will fit. This was not true in this case, as neither size fit well, but if forced to choose the size of best fit would be a 77. Further investigation by trying multiple sizes challenges the false underlying assumptions allowing the investigator to uncover the real problems.

It was determined that the women could achieve a good fit in the men's proportioned pants with the following modifications: (1) shorten the front crotch 14 mm at waist and 16 mm at crotch to remove bunching and allow forward bending, (2) increase the back crotch 30 mm at the fork allowing a back length for squatting or add a stretch reinforced gusset, (3) keep the overall crotch length the same effectively moving the pant leg forward to ease pressure on the thigh when bending, and (4) slightly elasticize the front waist by using cotton/elastomeric fiber mix for the waistband to ease forward movement.

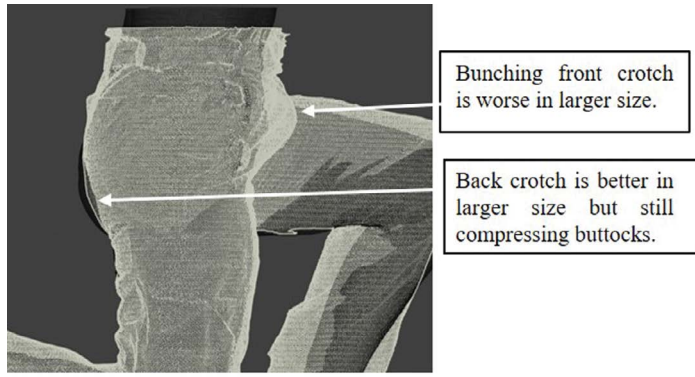


FIGURE 5.13 Same female paramedic wearing the next size up, men's size 82R.

These changes might not work for men. There should be a test of this new pant proportioning on the male population as well as the female and see if they both get an acceptable fit, in which case there could be unisex sizes. If not, then this becomes the new base for the female pant. If there is a female size range, the best choice for the base for women should be used to implement the changes. A full fit test will be needed to determine the best assortment of sizes.

Case Study 3: Assessment for Purchasing Tariff

This study demonstrates the need for fit testing before purchasing protective equipment. A Law Enforcement Officer (LEO) procurement department wanted to know what sizes of body armor vests to purchase. They had an existing database that included some anthropometric measurements and which size vest each officer was allocated when they were recruited and fitted out by the manufacturer. They wanted to use their existing data to predict which sizes to buy. This assumes that the size allocated and fitted out by the manufacturer fits them. An analysis of the data indicated that many of them likely did not fit according to the COF. Therefore, either the COF is wrong or many of the officers had large unprotected areas with the body armor they were wearing. The COF was established by the National Institute of Justice. Therefore, not complying with this guide raises the question about the safety of the officers wearing the current body armor.

The ideal process should have been:

1. Develop a COF
2. Pilot test the COF
3. Exploring which LEO body dimensions are important for product assessment
4. Assemble or obtain a TP database anthropometry
5. Conduct a fit test to test sizing and prediction charts

However, this is not what happened. We were brought in after the purchase decision for which company's body armor to purchase, and after they had already done a survey that did not include a fit assessment. They only collected the designation of the size allocated, presumably the size of best fit, with no quality of fit score. So, if it didn't fit or if it fit poorly that was not recorded. They included only three body measurements and that did not include height and weight which made the data hard to compare to other surveys that have 3D scan data for visualization, such as CAESAR (Blackwell et al., 2002; Robinette et al., 2002). Finally, there was no time or money allotted for fit testing, so this survey was all we had, and no live subjects were used.

Inputs

The COF was obtained from the National Institute of Justice's body armor guide (NIJ, 2014). It indicated that the armor front length should extend from just below the jugular notch (1.75 cm below) to no more than 2–3 finger widths (about 5 cm) above the officer's omphalion (belly button) when standing. The unstated and untrue assumption is that the officer's duty belt and omphalion are located in the same place. The intention is that this clearance allows for the shortening of the front length of the body when a LEO sits. In the field of anthropometry, the jugular notch is the top of the sternum and is also called Suprasternale. The location of the duty belt is the preferred waist as described by Blackwell et al. (2002). The measurement on the body from Suprasternale to the preferred waist in the front is called Waist Front Length.

The NIJ guide also indicated the vest should not interfere with the LEO's ability to shoot in their preferred stance and at the same time provide good coverage in the armhole area. The standing and shooting stance meant that the Interscye measurement was likely significant. References to the number of acceptable overlaps at the side (although this is adjustable with Velcro) were associated with circumferences of the waist and chest. The neck opening was associated with both neck circumference and the location of the neck in relation to the armhole.

For the TP, the LEO survey had data on 4000 new recruits (both male and female although this analysis is for males only). The data were collected by the manufacturer of the existing vest. The data recorded were:

- Chest Circumference
- Waist Circumference at Omphalion (belly button)
- Waist Front Length, Standing (to Omphalion)
- Size of Existing Vest worn

There was no data on interscye, bust point to bust point or neck measurements so these aspects of the NIJ COF could not be examined from the LEO data alone. Also, since the waist location used for the waist measurements was the Omphalion landmark, we did not have the duty belt location either. The duty belt waist location is the preferred waist for their belt. Normally height and weight allow the comparison of data from different sources as these two measurements are usually standardized, easy to take and very reliable. In the absence of these data, we used

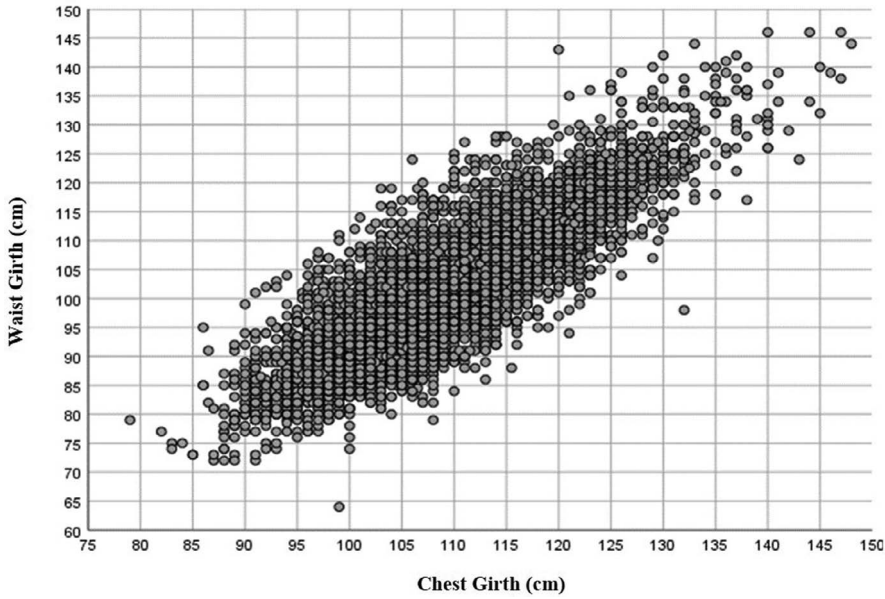


FIGURE 5.14 Bivariate chart of Chest Circumference by Waist Circumference for males.

Chest Circumference to match the samples. The CAESAR database measured Chest Circumference in the same manner as this study and included the duty belt waist measurements, as well as neck and interscye measurements. This provided us with some estimates of this missing data.

Scatterplots were prepared in SPSS® for all these measurements for males to see which data might be relevant and what range should appear in specification sheets for the vest comparisons.

Analysis of the LEO data showed the average Chest Circumference to be 109 cm with a Standard Deviation (SD) of 9.3 cm. The average Waist Circumference for the matched CAESAR subjects was 101.4 cm, and the average Waist Front Length was 42.1 cm with an SD of 3.2 cm. A bivariate chart of Chest and Waist Circumferences is shown in Figure 5.14.

A bivariate chart of male Chest Circumferences and Waist Front Length, Standing is shown in Figure 5.15. Individuals with a Chest Circumference of 109 cm can have a Waist Front Length as short as 35 cm and up to 52 cm. This indicates that no matter what the Chest Circumference nearly the full range of lengths are needed to cover that person's vital organs.

The Waist Front Length distribution is shown in Figure 5.16. The range to accommodate 95% of the Waist Front Lengths is approximately 34–50 cm, which is a range of 16 cm. This suggests the vest will need to come in at least two and possibly three different lengths to fully protect the front torso.

Specifications sheets were obtained from the existing supplier that included vest grading. These are shown in Table 5.7. This vest came in nine width sizes (36–68),

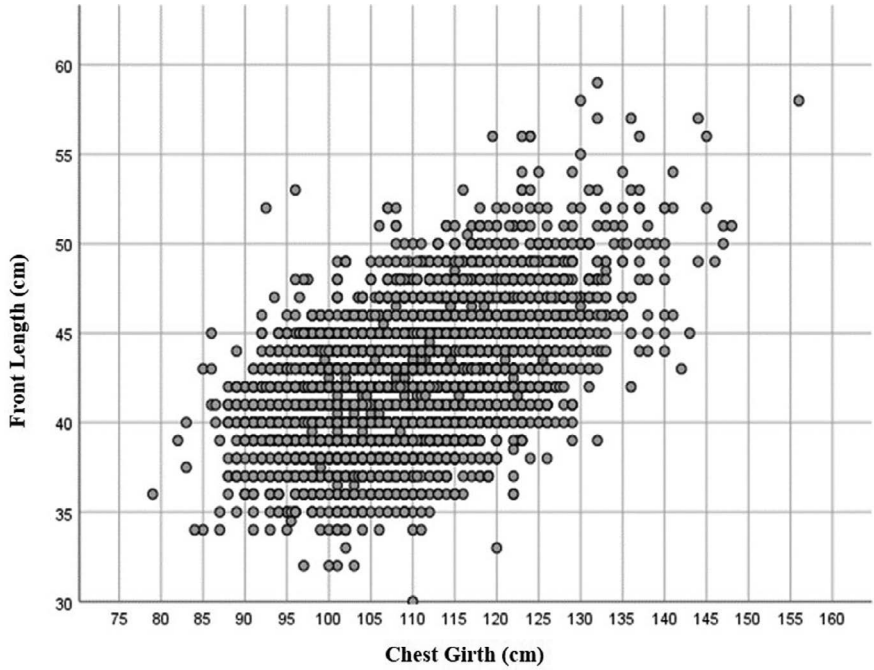


FIGURE 5.15 Chest Circumference by Waist Front Length, standing.

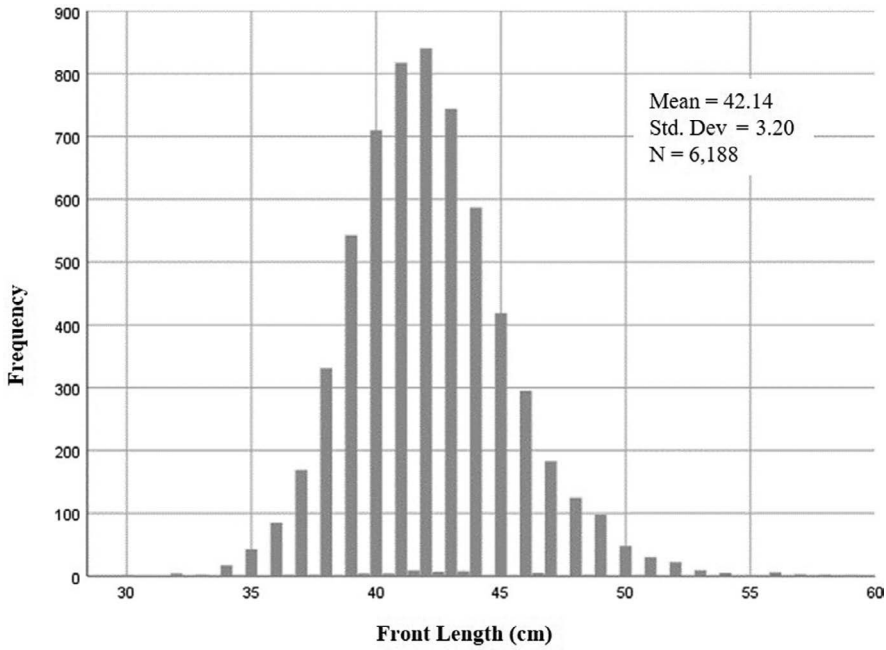


FIGURE 5.16 Waist Front Length distribution.

TABLE 5.7
Body Armor Vest Size Specifications

	Size								
Length (cm)	**36**	**40**	**44**	**48**	**52**	**56**	**60**	**64**	**68**
Short	30.3	30.9	31.5	32.1	32.7	33.3	33.9	34.5	35.1
Regular	33.0	33.6	34.2	35.1	36.1	36.0	36.4	37.1	37.6
Long	35.2	35.8	35.8	37.1	37.7	38.3	38.9	39.5	40.1
Range	4.9	4.9	4.3	5.0	5.0	5.0	5.0	5.0	5.0

each with three lengths, short, regular, and long. There were a few men who needed the larger sizes 72 and 76, but there were no specifications for those sizes.

We added 6.75 cm to the length values to match the garment lengths to the body length (Waist Front Length) that would be deemed a fit. That is adding 1.75 cm to move it up to the Suprasternale, and 5 cm to move it down to the duty belt. These values are shown in Table 5.8.

Surprisingly, even though the vest came in three lengths, the range of vest lengths was only 5 cm. That is less than one-third of the body Waist Front Length 95% range. To visualize the impact, we examined the vest lengths from shortest to longest for each size which were then overlaid on a bivariate chart of the size worn by the LEOs by Waist Front Length in Figure 5.17.

None of the men wore the smallest two sizes. For all sizes, there were some men for whom the longest vest was too short. This gets worse as the size gets larger. Of those men who wore the sizes 64 and 68, more than half had a vest that was too short, meaning they did not have the torso coverage indicated as needed in the COF. In size 64, the gap in coverage was as much as 12.9 cm and for size 69 it was up to more than 10 cm. There were also problems with the vest being too long. When it is too long, it presents problems when sitting or crouching, pushing the vest up into the neck. Sizes 52 and 56 were the worst for this problem with the vest being as much as 7.5 cm too long in the 52 and as much as 10 cm too long in the 56. This indicated that the range of lengths was too narrow. With the range of Waist Front Lengths in each size being approximately 20 cm, the difference between the lengths should be double to triple the spread now, or 5–7.5 cm.

Also, shown in Figure 5.17 is the average body grade for Waist Front Length. The middle range of the best length should ideally fall on this grade line to ensure that the point with the greatest number of people near it will fit in the middle length size.

TABLE 5.8
Lengths Adjusted to Match Corresponding Body Waist Front Length (cm)

Adjusted	**36**	**40**	**44**	**48**	**52**	**56**	**60**	**64**	**68**
Short	37.1	37.7	38.3	38.9	39.5	40.1	40.7	41.3	41.9
Regular	39.8	40.4	41.0	41.9	42.9	42.8	43.2	43.9	44.4
Long	42.0	42.6	42.6	43.9	44.5	45.1	45.7	46.3	46.9

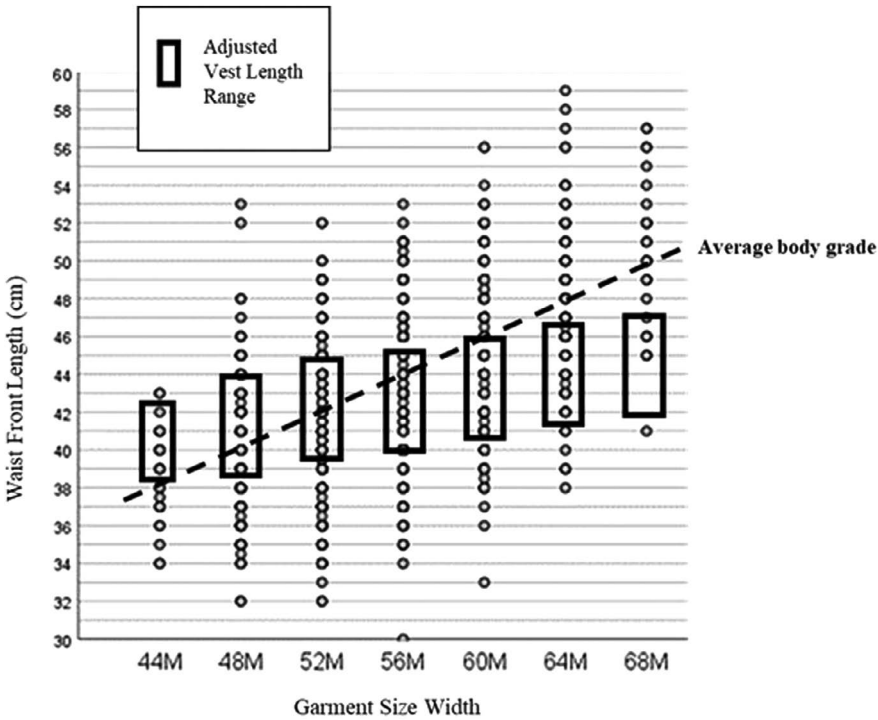


FIGURE 5.17 Short to long vest sizes by width size.

In this case, the largest lengths of vests fall below the grade line in the larger sizes, and the shortest length of vest falls above the grade line in the smallest size. This was a clear indication that the grade was off. This vest has missed the fit sweet spot.

The combination of the range being too narrow and the grade being off-center meant that the percentage of men who met the COF was only approximately 30–40% of the LEO male population. If the COF was accurate this meant this was the approximate percentage who would get a safe and effective fit in the current set of sizes.

Fixing the grade alone would increase the percentage fit to as much as 75%, provided it is the rigid ballistic plate that is graded and not just the fabric. This is probably a large source of the problem. Changing the ballistic plate is expensive and there is a tendency to want to use the same plate size for every garment size. The manufacturers who use one plate for every size can probably provide a cheaper proposal or offer, but poor protection and performance. The degradation in performance and safety might not be noticed unless the product is evaluated for protection coverage.

Creating three to five plate lengths that are each 5 cm apart should provide more plate options to help fit the center area represented by the average grade line for each size. This will also increase the percentage of men who get a fit according to the NIJ COF, and who will have a more comfortable fit as well. However, since the NIJ COF has never been validated or evaluated on live subjects we do not really know that it is a good definition of safe and effective fit. To ensure a safe and effective fit, a fit audit is needed, with testing on live subjects representing the LEO population.

Case Study 4: Prototype Test to Determine the Correct Alteration

This is a practical example based on an actual problem from a womenswear apparel company in the 1990s in Melbourne Australia. It illustrates how customer complaints can be investigated to determine and fix an underlying problem in the fit standard.

The apparel company was a very well-known and respected family-run company that had been in business since the 1960s. It had experienced sustained growth, especially during the 1990s, producing millions of garments annually both under its own brand names and under the private labels for large Australian department stores like Myer, Target, and David Jones. The company was very successful and achieved accolades from the department stores of preferred manufacturer status and was rightly proud of its reputation for quality. When working for itself, the company was vertically integrated and had control over its brand identity. When working for the large department stores it acted as both PD and production, with the cost of PD being recouped during the production phase once it had secured orders. This distribution of the cost of PD was standard in the Australian industry at that time. It meant the company took a risk every time it developed a style – if it didn't sell, the company would lose money – so the pressure was on the technical team to do a good job, fast, first time, every time!

The job of the pattern maker directly involved not just making the patterns, but also overseeing and approving the sample-making and producing technical specifications, while at the same time thinking about how the company could save money during future production, solving any technical problems that arose in the most cost- and profit-oriented way possible. This was comparable to a product engineering position. The quality of these solutions directly affected the company's profitability. The company made a lot of money and was doing very well; however, it still had problems.

In the 1980s and 90s, almost all the garments made by this apparel company had shoulder pads. The trend going into the 2000s was that shoulder pads got smaller and then they were removed. When a garment has shoulder pads the armhole is bigger and correspondingly the sleeve is wider. Removing the shoulder pads allowed the sleeve widths to narrow. This then caused a new problem when some customers reported that several different styles had tight sleeves in a size 16 (large size ladies wear). The patternmakers were asked to investigate it and solve this problem. In this case study we compare what they did then and its result with how it could have been solved if they had known about and used the procedures in this book.

Standard practice in the company was to produce a base size (bust measurement 90 cm) and grade the pattern based on x and y coordinates to produce the other sizes automatically. It used the then state-of-the-art newly emerged computer technology for this. The grading was the 50 mm bust grade. The sleeve was graded 12 mm per size, as was consistent with what was taught in fashion design school in Australia in the 1980s.

The company always checked the fit of the base size with a garment sample on a manikin and a fit model, both which represented their unverified fit standard. The fit model was knowledgeable about the implicit fit expected and knew how to check the functionality of the garments. Once the sample was approved and sold to the customer there was no further checking of the sewn garments for fit.

All sizes were subject to quality assurance procedures, but these only involved making sure the sewn garment matched the specifications listed in the "spec" sheet. For example, they would check to ensure that the pocket was in the correct

location. Millions of garments were cut, sewn, and sold. The only feedback the company had was from the retail sales figures. This meant they were acutely aware of "best-sellers" and so they developed variations on these styles, slavishly hoping to have more best-sellers. This often worked and made the company more money.

It was in this environment that the company received feedback from some retailers suggesting the sleeve was too tight in the upper arm for some of the ladies who needed larger sizes, in particular those ladies wishing to buy a size 16. The retailers also suggested a solution. They asked the technical team in the company to increase the sleeve in the base size to fix this problem. The technical team was immediately concerned that the proposed solution would cause a cascade of other problems. If the technical team accepted the suggestion, they had to decide what to do with the extra length in the sleeve head. The sleeve needed to be sewn into the armhole, so should they make the armhole bigger or add more ease? Also, if it was a problem for that style what about all the other similar or future styles? Did they need to be "fixed" in the same way? If not, why not?

This was the mid-1990s and coming out of the 1980s all the styles had shoulder pads. This extra length in the armhole (to fit the shoulder pad) meant the sleeves were already wide. The design team was nervous and reluctant to make them wider, especially as the emerging trend was for smaller shoulder pads and narrower sleeves. Remember back then there was no anthropometric data on the Australian TP and the idea of testing was not anything the company management understood. Both the design and the patternmaking teams were uncertain about what to do.

The patternmakers had two choices. Either do what was asked or ignore the request. Arguments flowed both ways around the pattern room and in the design, production, and management teams. On the one hand, the company wanted to please the customers and do the right thing to improve the product. On the other hand, everyone, including retail, was happy with the existing base size sleeve. It looked good and they had been getting a steady stream of approvals and subsequent sales.

If the company made the base size sleeve bigger it would look wrong. This would have introduced a considerable risk of reducing the number of sales in the most frequently selling sizes which would not have been offset by any increase in sales in size 16. In addition, it would have caused a cascade of the technical problems described above. Utmost in people's minds was how any change would play out in the mass production setting. Clearly, more ease in the sleeve head would be unacceptable, as the bigger the ease allowance, the more skilled the sewing machinist needed to be, as this was a more difficult sewing task. The production manager would say no immediately. More ease meant more fails at the quality control point and the need for a higher-paid, more skilled workforce. The company needed to manage all these factors. The patternmaker was the person overseeing, doing, and coordinating pattern changes, and was ultimately responsible for any increase in the cost of production. Juggling all these factors the patternmaker needed a very good argument at hand to justify any major changes with cost ramifications. To say that they were "told to do it" would not have been good enough. So, with all this in mind, after much discussion the pattern room personnel told retail they would look at it, which they did; and then the final decision was to *do nothing*. The only downside was it did not solve the reported problem of the sleeves being too tight in the size 16 garments.

Fast Forward 30 Years…

What measurements, tests and analysis methods could be used in the current environment with current knowledge to solve this problem? Are there any extra benefits to using these methods?

It turns out now there was a third choice: check body growth in each size against the base size and grading. This was not an option at that time for the company; even if they had known how to do it, they didn't have any anthropometric data, certainly none that represented the population. But it is an option now. This is what the patternmaker could do now if asked to solve that problem.

Solution: Map the base size and grade against the TP.

First, select a sample to represent the TP then select relevant measurements such as those shown in Figure 5.18 (Upper Arm Circumference and Bust Circumference), and compare the grade to the TP distribution as in Figure 5.19.

Then calculate average arm measurements per size using linear regression as described in Chapter 3. The arm measurements in Table 5.9 were calculated using a linear regression equation y = 0.35x–28.47 where x is the Bust measurement and y is the Upper Arm Circ. For size 10 the Upper Arm Circ. = –28.47 + 0.35 * 900 which equals 286.53 mm (rounded). Extra benefits included being able to identify any missing sizes and determine the tariffs.

The number of subjects in a size was estimated by dividing the sizes into Bust Circumference size ranges and counting them. By examining the number and percentage of subjects per size, as in Table 5.10, we see where potential sales might be. This is very useful for determining the cut ratios. This would need adjustment if the company decided it would not produce sizes 6 and 4, which was true for the historical example used.

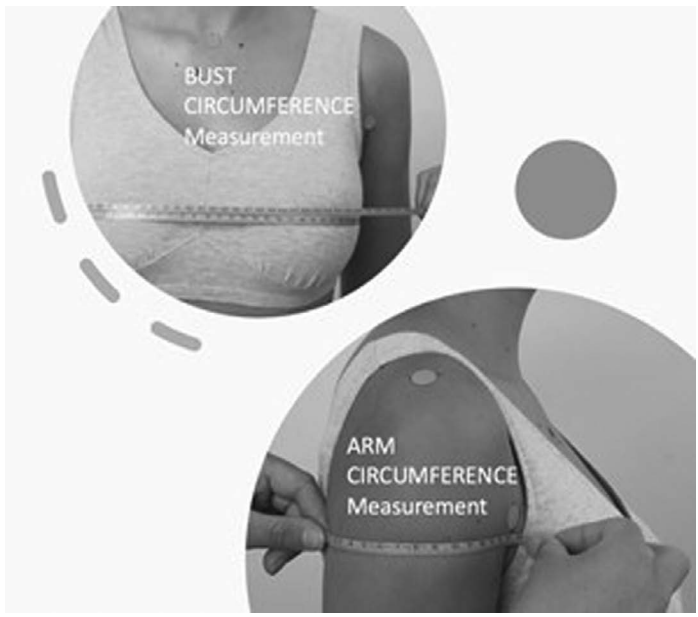


FIGURE 5.18 Two relevant measurements.

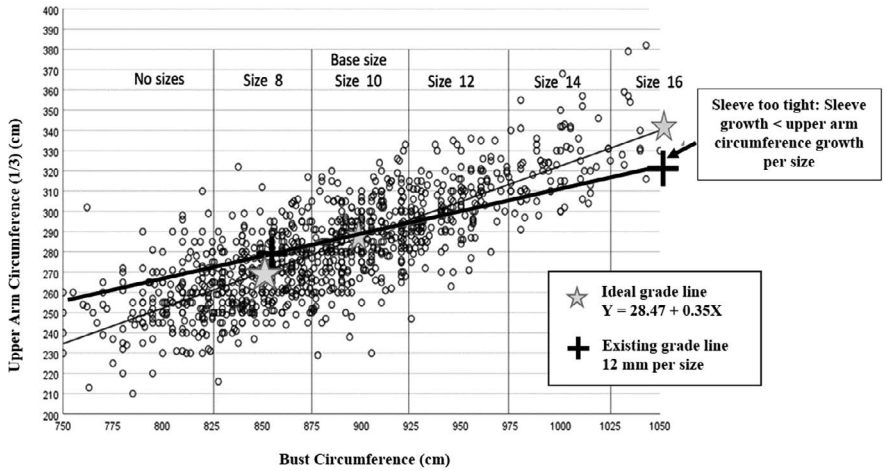


FIGURE 5.19 Existing grade versus ideal grade.

TABLE 5.9
Subjects Sorted According to Bust Sizes to Calculate an Average Arm Girth per Size

AU Size	Bust Mid. (mm)	Bust Min. (mm)	Bust Max. (mm)	No. of Subj.	Actual Arm	Sleeve Growth	Error	Cumulative Grade Error
	750	725	775	22				
6	800	776	825	164	252	–12	5	Slv 11 mm too big
8	850	826	875	328	269	–12	6	Slv 6 mm too big
10	900	876	925	318	287	0	0	0 mm (base size)
12	950	926	975	157	304	12	6	Slv 6 mm too small
14	1000	976	1025	68	322	12	6	Slv 12 mm too small
16	1050	1026	1075	22	339	12	6	Slv 18 mm too small

TABLE 5.10
Percentages in Each Size

Size	Frequency	Percent	Cumulative Percent
4	22	2.0	2.0
6	164	15.2	17.2
8	328	30.4	47.6
10	318	29.5	77.1
12	157	14.6	91.7
14	68	6.3	98.0
16	22	2.0	100.0
Total	1079	100.0	

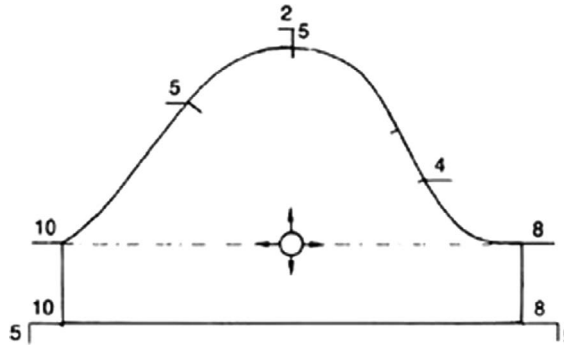


FIGURE 5.20 Sleeve with grading marked in mm.

In the apparel industry, the pattern piece is shown by its shape, and the grade, marked in both x and y directions is shown in mm on the pattern piece in the location of the graded increment. In Figure 5.20, where there is a 10 and 8 marked, they add up to an 18 mm sleeve grade in the x-direction (sleeve width). In addition, there is a 5 mm grade in the sleeve height below the head and a 5 mm grade in the head height making the total grade in the y-direction 10 mm. This summary of the grade allows the garment growth (sleeve width of 18 mm) to be seen immediately.

Grading practices have changed over the years. The use of the computer has allowed grading to become more advanced. This change in the sleeve grade from 12 mm to 18 mm per size is now Australia's most used sleeve grade increment. When grading follows the body growth it is sometimes referred to as allometric or 3D grading. In addition, different companies use different grades, and they are often proprietary and that is part of their fit standard.

Discussion

The correct fix that the technical team missed in the 1990s was a grading problem. Specifically, the sleeve growth mismatch to arm growth made the sleeve progressively tighter for each size we graded. The technical team's decision to do nothing was the best decision they could have made at the time; however, it did not fix the problem. The problem was not a problem of a single size but a systematic problem that affected all sizes. The wide sleeve 80s styling with shoulder pads concealed the grading issue. The problem only emerged in the 90s because the styling was changing, shoulder pads were getting smaller or being removed and the trend was toward tighter, more slim-fitting sleeves.

The base size did not need to change and if they had followed advice from retail and changed the base size that would have caused a cascade of problems and been very bad for business. Current analysis and techniques using data representing the TP allow informed solutions with minimal risk in terms of time or money. The solution was to leave the base size as is and fix the grade. This would have improved sales in the size 16. As an incidental finding, when we mapped the bust sizes against the TP, we learned we should have added a size 6 and potentially picked up an additional 15% in market share.

Although the Company had a "fit standard" that fit standard was never verified. The best practice is to use the SPEED process to fit test best-selling garments in their full size range to establish any issues in individual fit, the range of fit in a size, and the best fit for the TP. When the fit standard is *verified* it becomes a "sustainable fit standard" and leads to improved efficiency for the company such as a cost-effective set of sizes that fit the TP, accurate size selection charts and algorithms, and sustainable tariffs.

REFERENCES

ASTM International. (2015). *Standard Tables of Body Measurements for Juniors, Sizes 0 to 19.* ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959. United States.

Blackwell, S., Robinette, K. M., Boehmer, M., Fleming, S., Kelly, S., Brill, T., Hoeferlin, D., & Burnsides, D. (2002). *Civilian American and European Surface Anthropometry Resource (CAESAR). Volume 2: Descriptions.* Sytronics Inc Dayton OH. https://apps.dtic.mil/docs/citations/ADA408374

NIJ. (2014). *Selection and Application Guide to Ballistic-Resistant Body Armor for Law Enforcement, Corrections and Public Safety: NIJ Selection and Application Guide-0101.06* (Standard Guideline NCJ Number 247281; p. 104). U.S. Department of Justice, Office of Justice Programs, National Institute of Justics (NIJ). https://nij.ojp.gov/library/publications/selection-and-application-guide-ballistic-resistant-body-armor-law-enforcement

Robinette, K. M., Blackwell, S., Daanen, H., Boehmer, M., & Fleming, S. (2002). *Civilian American and European Surface Anthropometry Resource (Caesar), Final Report. Volume 1. Summary.* Sytronics Inc. Dayton OH.

6 Head and Face Wearables

Karen Bredenkamp

ABSTRACT

This chapter describes the wearable product design process for the group of products that are worn on the head and face, highlighting unique challenges and requirements, and describing practical sizing and design approaches through examples. The head and face area are home to all five of the human sensory organs: the eyes, the nose, the tongue, the ears, and the skin. In addition, the head houses the brain, one of the most important vital organs. It is therefore understandable that products designed to protect, enhance, or monitor human sensory performance or status are head and face worn. Furthermore, head and face-worn products are more likely than products in other areas to require being located within a small proximity of the sensory organs, especially the eyes and ears. This means head and face wearable design has special challenges. This chapter discusses the challenges and how to best accommodate them using the sustainable product evaluation, engineering, and design process (SPEED) approach.

BACKGROUND

Head wearables today are much more complex than they were just a 100 years ago. As a result, unlike fashion apparel and footwear, very few consistent or established sizing systems exist for headwear products. A lot of these products are so unique in their applications that their product sizing will need to be done from scratch. In addition, unlike fabric apparel discussed in Chapter 5, many products worn on the head and face have rigid or molded material components. Examples include biking or motorcycle helmets, full- or half-face protective masks, oxygen masks, headphones, eyeglasses, AR/VR headsets, helmets with heads-up displays, etc. Some of these products also include complex and sensitive electronics that need to enhance the human experience or performance by interfacing with the human senses and abilities (most commonly vision, speech, and hearing).

With all the sensory organs located in this area, head wearables must often interface with other head wearables. For example, helmets may have to interface with masks, earphones, glasses, etc. All these items have their specific fit issues and fit testing criteria and sometimes have competing requirements. As a result, fit assessment for many head wearables must include the complex fit of an ensemble of head worn items. Performance metrics for the ensemble and for the individual components can be very important for inclusion into the Concept-of-Fit (COF) metrics. Examples of performance metrics include:

- CB mask leakage (particle count reduction)
- Real-world visibility

DOI: 10.1201/9781003397533-6

- Virtual field of view
- Eye tracking performance
- Noise attenuation
- Speech intelligibility
- Slippage or stability

Most performance metrics cannot be measured without testing with human subjects for two key reasons. First, neither digital human models nor physical manikins have all the physical properties of real people therefore they do not sense things like real people. For example, we can measure temperature, but we cannot determine the temperature that is uncomfortable without testing on people. Also, since we do not know how tight is tight enough and how tight is too tight, the assessment of slippage and stability is impossible without a human subject in the loop. Second, the somewhat spherical shape of the head, the paucity of reliable landmarks in the cranial region, and the large random variability of the landmarks on the face make it impossible to accurately determine the location of the wearable and the other interfacing components in the ensemble on any individual without donning a prototype.

For example, the F-35 helmet must be worn with noise attenuation earcups and an oxygen mask. These items must be positioned correctly to work properly and cannot interfere with the helmet-mounted display. Webster reported that the oxygen mask can be located too close to the visor such that it bends the visor and distorts the display, or it can be too far away resulting in leakage that prevents proper oxygen flow (Webster, 2021). During fitting, pilots are sent to an oxygen tester where technicians identify leaks and ensure proper mask fit when the pilot moves or talks. This type of testing when done early in the design process helps us avoid costly re-designs late in the process.

The SPEED process employs human testing early in the design process with iterative trade studies and design loop testing to ensure the final product meets all performance requirements. We discuss the issues and how to resolve them in this chapter with examples in the case studies section.

Head wearables have constraints that are not directly related to the purpose of the product but may result in impacts on the function of the other sensory organs. As a result, there can be many fit factors to be considered in addition to the ones that satisfy the purpose of the product. This factor can make a fit assessment more complicated. For example, a protective mask may need to be worn with eyeglasses without causing them to fog up. The eyeglasses are not part of the protective mask design, but they should be included as part of the fit assessment. It is important to consider the impact of the entire system when planning head wearable testing.

For protective masks that need to filter out hazardous materials, special equipment to measure leakage might be needed. Case et al. (1989) used a test chamber, the Dynatech Frontier Portable Fit Testing System 1000, for testing a chemical and biological protective mask, the MCU-2/P. This system measured the amount of challenge agent (corn oil mist) that leaked into the masks. A proper seal is an important fit factor in this case. The fit test led to the discovery that the way masks were being issued may have been flawed, by providing masks that were too large resulting in undesirable leakage.

For head wearables, movement, and even facial expressions can impact fit and function, so fit measurements should include activities the wearer might be expected

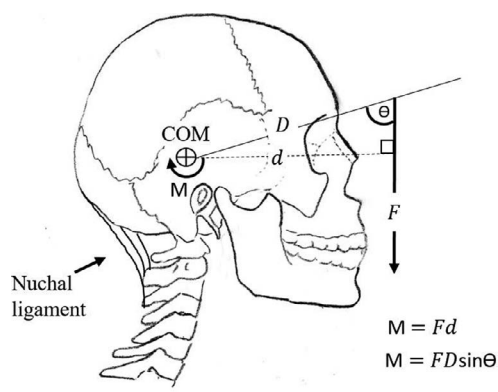
to do while wearing it. For example, in the testing of the MCU-2/P mask (Case et al., 1989) the subjects performed a list of activities while in the test chamber:

- Normal breathing
- Deep breathing
- Walking in place
- Looking up, left and right, while on hands and knees
- Stepping up and down
- Touching toes
- Twisting at the waist
- Rapid side-to-side head movements
- Talking
- Shallow knee bends, and
- Various facial expressions (yawning, smiling, frowning, rotating the chin)

Subjective comfort is very important for the COF as it competes with other fit criteria including slippage and stability, and temperature on comfort. The length of wear time should also be considered because the wearable might be comfortable when first donned but be unbearable after 2 hours.

CENTER OF MASS VERSUS NECK STRAIN

The head rests on the neck and spine, so weight and the location of the weight worn on the head affect the musculoskeletal regions of the neck and Cervical spine. Weight applied far away from the center of mass (COM) of the head causes a larger moment and therefore higher strain on the neck, which could result in neck pain and injury, as illustrated in Figure 6.1. Moment (M) is a function of the amount of vertical force (F) or



Distance (D) [cm]	15	20	25
Force (F) [N]	1	1	1
Angle (Θ) [°]	90	90	90
Moment (M) [N.m]	0.15	0.2	0.25

FIGURE 6.1 Moment around COM on head.

weight applied, and the perpendicular distance (d) to the center of rotation. In this figure, we keep F and θ the same, one Newton and 90 degrees respectively, but we increase the distance (D) between the object and the COM of the head, and show its effect on M. It's clear that by just moving the object further away from the head COM, increases the amount of moment on then neck proportionally. This means if we add weight to the wearable in front of the eyes, such as when we add heavy lenses or displays, it will cause more neck strain than it would if we added the weight closer to the head COM, such as around the ears. The increased moment also increases instability and slippage during movement. Asymmetrical weight distribution can also cause fatigue, pain and/or injury due to uneven loading of the neck. Therefore, head wearables will most likely have to take weight distribution into account in the requirements and COF.

SENSITIVITY TO TEMPERATURE

Another important issue for head worn products is the particular sensitivity to temperature on the head and face. Several studies have suggested thermal discomfort to be a reason for persons not wanting to wear protective headgear (Li et al., 2008; Patel & Mohan, 1993; Skalkidou et al., 1999). A lower rate of motorcycle helmet use has been reported between the north and south of Italy (93% versus 60% respectively), supposedly due to higher rates of thermal discomfort in the warmer southern climate (Servadei, 2003). Options to increase user comfort could include special textiles such as phase change materials (Tiest et al., 2012) or increase ventilation, which enhances both convective cooling and seat evaporation (Alam et al., 2010; Bogerd et al., 2015; Mukunthan et al., 2019a,b). The skin plays a fundamental role in the thermoregulation function to maintain the core body temperature around 37°C. If the whole body or local areas of the body becomes "too hot", blood flow is increased through the dermis (vasodilation) to release heat through the epidermis to the environment. If greater heat loss is required, the surface of the skin is moistened with sweat so that the latent heat of vaporization may be lost through evaporation (Meyer et al., 2002). In warm environments, evaporation of sweat is the main heat loss pathway (Bogerd et al., 2015). Local sweat rates at the head correlated significantly with whole body sweat rate (Bain et al., 2011) and varied spatially with higher sweat rates reported at the forehead compared to the temple, vertex, and rear regions (Cabanac & Brinnel, 2000; Machado-Moreira et al., 2008; Smith & Havenith, 2011). Local differences in thermal sensitivities of the head's surface have furthermore been observed, with the forehead seen to be typically more sensitive than the sides of the scalp (Mehrabyan et al., 2011).

During the design phase, several designs could be evaluated for thermal comfort as one of the objective fit criteria using thermal computational simulation models or physical thermal head forms. The application of anatomically formed thermal manikin head forms provides information about combined heat and mass transfer with and without the use of headgear. Several such simulation models and head forms have been developed. Examples include work published by Abeysekera et al. (1991), Bandmann et al. (2018), Ellis (2003), Hsu et al. (2000), Liu and Holmér (1995a), Mukunthan et al. (2019a,b), Osczevski (1996), and Reid and Wang (2000). Some are commercially available, e.g., the head forms offered by Measurement Technology Northwest (USA), and UCS d.o.o. (Slovenia).

Studies were published using different types of thermal manikin head forms focusing on impaired heat and mass transfer for industrial helmets (Abeysekera et al., 1991; Hsu et al., 2000; Liu & Holmér, 1995b; Reischl, 1986) as well as sports headgear (Bogerd et al., 2008; Brühwiler et al., 2004, 2006; Bruyne et al., 2012; Pang et al., 2013; Reid & Wang, 2000). Thermal manikin head forms are equipped with heating elements and temperature sensors. The surface temperature of the head form is typically regulated at a fixed temperature, and the total power needed to maintain this temperature over a steady-state period is recorded. This heating power equals the total net heat transfer, which allows the quantification of combined heat loss by convection, conduction, and radiation. To quantify the individual heat transfer pathways, measurements must be carried out with conditions allowing only a particular heat transfer pathway while preventing the other pathways. The heat transfer pathway is then quantified by calculating the heat transfer differences between the different conditions.

Of course, these devices can measure the temperature, but they cannot tell us what temperatures are comfortable. We need human subjects for that. However, if we have both we can correlate comfort with temperature in a particular headwear ensemble to help us design the product.

HEAD ORIENTATION AND ALIGNMENT

The somewhat spherical shape of the head presents stability, slippage, orientation, and alignment challenges that are not issues for other areas of the body. While the nose and the ears provide some anchor points for a wearable, they are small, soft, and very sensitive to weight and pressure, so it is rare for these to be good anchor points for anything other than very lightweight wearables. As some of us know, even heavy eyeglass lenses can make them painful to wear. For comfort and stability, it is usually better to fit snugly to the cranium rather than the nose or ears. Hair type is an additional variable that influences slippage and stability. Including objective and subjective metrics evaluating slippage and stability in the fit metrics for head wearables are regularly needed.

However, the most difficult challenge is determining the orientation and alignment of the head for measurement, shape analysis, and/or product design. This is predominantly due to the head not having a clear long axis such as the whole body, arms, legs, or even feet. In addition, the neck anatomy enables a large range of motion: flexion, extension, and rotation.

Head orientation is defined as the head rotation around the x-, y-, or z-axes also referred to as pitch, roll or yaw (see Figure 6.2). A challenge for head wearables is that many different head orientations can be chosen and none of them may coincide with the axis system of the wearable (how the wearable is worn on the head).

Under certain conditions, the orientation of the head is not important or applicable. This includes the analysis of 1D measurements which does not incorporate head orientation information, such as Head breadth, Head length, Face length, Face breadth, Bi-Tragion breadth, Nose width, Mouth width, or even linear distances between

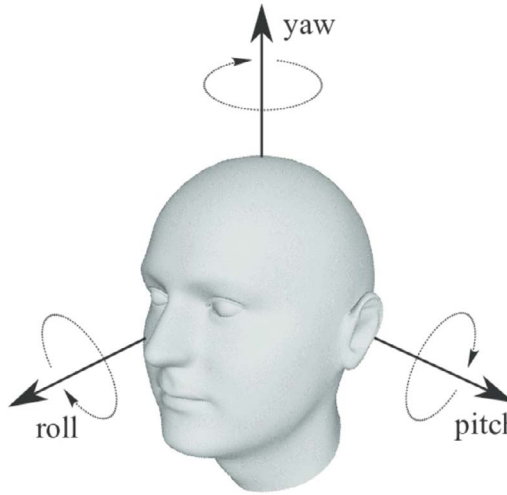


FIGURE 6.2 Pitch, roll, and yaw head rotations.

landmarks such as Glabella to Tragion, or Sellion to Pronasale, etc. However, measurements that incorporate orientation information such as the horizontal fore-aft distance from Tragion to Glabella (only in the z-direction), or Tragion-to-top of the head (vertical measurement, only in the y-direction), etc. are impacted strongly by the head orientation. Therefore, if such measurements are key to the function of the product and must be used for size/shape variance understanding and input to the design, then selecting the correct head orientation becomes very important. This was illustrated in Chapter 2, Figure 2.5, the point-of-view error illustration. To demonstrate this, the effect of head rotation on the measurement values is illustrated in the side views in Figure 6.3. Here we have a head that is measured three times, each in a different head orientation, specifically head pitch.

The *a* measurements are the measurements to the top of the head, and the *b* measurements are to the back of the head. With the three trials overlaid (in the bottom center of the figure), the variance between the location of the infraorbitale point for each orientation is illustrated (see the bottom right of the figure). The position of the point in the vertical or horizontal axes is dependent on the pitch orientation of the head. In this illustration, the location of the landmark differs by more than 40 mm in the vertical direction from trial 2 to trial 3, even though this is the same subject.

For product design, the only correct alignment is the product alignment. This problem is explained with many examples in the chapter on HMD displays by Whitestone and Robinette (Melzer & Moffitt, 1996). Unfortunately, until we have a mock-up or prototype to place on actual people, we do not know what this alignment truly is. Therefore, we must make an educated guess for our first prototype. This educated guess can be improved using quick trade studies on a few live subjects

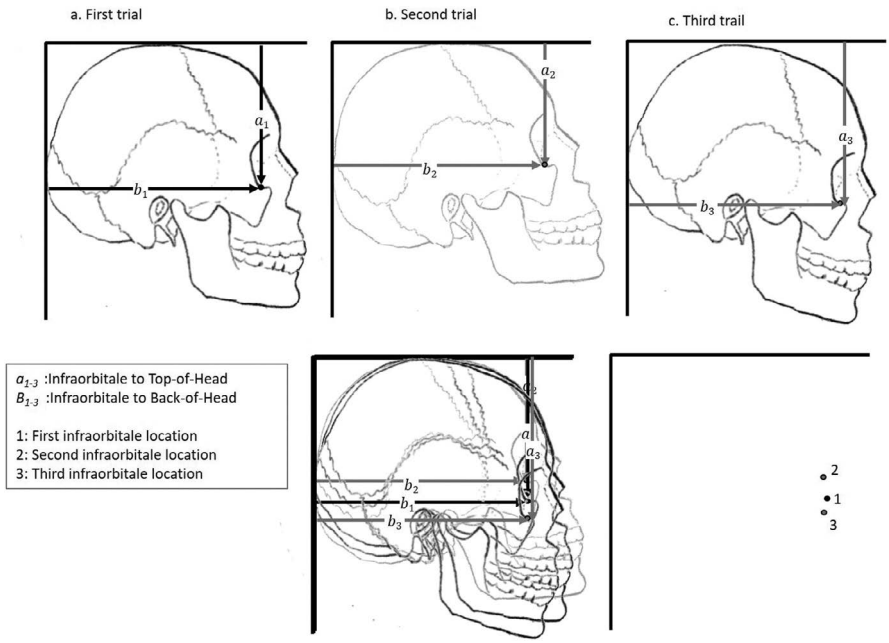


FIGURE 6.3 Three different head orientations and the effect on landmarks and measurements.

using 3D printed mock-ups with weights added to simulate the weight distribution and scans of the product in place on the live subjects. The mock-ups used can be prior versions of a product or other similar products, such as eyeglasses or headsets. The point is to improve the head orientation estimate with respect to the product. As the product develops, additional prototypes can be checked to fine-tune the product orientation.

There are many different alignments that might be used to get us started before we have a mockup. Three common ones are:

- Frankfurt Plane
- Procrustes
- Neutral gaze

While these three alignment methods are good for comparing people in a general way, they are not product alignments and will not reflect the head orientation in the actual product. Therefore, while they might help us get started, they will need to be refined using product mockups.

The Frankfurt Plane alignment was first established at a World Congress on Anthropology held in Frankfurt am Main, Germany in 1884 (Ranke, 1884) where a head orientation termed the Frankfurt plane was proposed. This had to be adapted for living humans and one of the first standard definitions was the Air Standardization Coordinating Committee Air Standard 61/83, (Air Standardization Coordinating Committee (ASCC), 1991), which defined it as "…a standard plane of orientation of

the head. It is established by a horizontal line passing through the right Tragion (the front of the ear) and the lowest point of the right eye socket". This definition has been widely adopted in the field of Anthropometry, such as by the International Standards for Anthropometry Assessment (ISAK, 2001), by the International Standards Organization Anthropometry technical committee in the 3D scanning methodologies for international compatible anthropometric databases Part 1 (*ISO 20685-1*, 2018), the Korean National Anthropometry survey (https//sizeKorea.kr), Japanese head anthropometry survey (Kouchi & Mochimaru, 2004), South African National Defense Force Anthropometry survey (RSA-MIL-STD 127 Vol 1, 2005), and the United States of America Army Anthropometry survey (Gordon et al., 1989, 2012), to name a few. This definition does however leave some space for interpretation. As used by ISAK, the head is viewed from the right side and neck flexed/extended so that the right Tragion landmark at the ear is on the same horizontal plane as the right Infraorbitale landmark. No head yaw or rotation is addressed, therefore assuming that whatever natural asymmetry or yaw is present when the participant originally positions his/her head, would still remain. ISO 20685-1 (2018) defines the Frankfurt plane as the head pitched such that the right Infraorbitale landmark below the eye is on the same horizontal plane as the right Tragion landmark, and that the head is yawed such that the right and left Tragion landmarks are on a horizontal plane. However, as we learned when we began measuring in 3D, the right and left Tragion landmarks are not necessarily symmetric, therefore aligning the head with right and left Tragions on the same horizontal plane, could cause other points on the face to be out of alignment for areas important to the specific product being designed for. For instance, the right and left eyes might now not be on the same plane which could be a requirement for AR/VR or other products designed around the eye. The midline of the face might not be aligned to the vertical, which could be a requirement for products designed around the nose and mouth.

During the SizeChina-Hunan head scanning project, after data collection, the head scans were digitally processed during which each head scan was rotated and translated into the Frankfurt plane. With the x-axis running through the left and right Tragion landmarks, the origin midway between the left and right Tragion, and a positive direction towards the right Tragion. The X/Y plane passed through the left and right Tragion and left Infraorbitale landmarks. Positive y-direction anteriorly, and positive z-direction upwards (Luximon et al., 2015; Wang et al., 2018).

Park et al. (2020) aligned 180 3D head scans to create a parametric adult head model with representation of scalp shape variability. These head models were aligned in the Frankfurt plane for statistical analysis. The coordinate system was defined with the origin midway between the left and right Tragion, the y-axis running through right and left Tragion landmarks. The X/Y plane passed through the left and right Tragion and left Infraorbitale landmarks. The z-axis is a cross product of x- and y-axes, running vertically.

Niezgoda and Zhuang (2015) analyzed the shape variance of approximately 4000 head scans of the United States Civilian population to create head forms for ISO Eye and Face Protection Standards. To analyze the variance of the head and face shape and sizes, Niezgoda and Zhuang (2015) aligned all heads in the Frankfurt plane. Similarly, Yu et al. (2011) developed digital 3D headforms representative of Chinese

workers for use in designing personal protective equipment for the Chinese user. These head forms were also aligned in the Frankfurt plane (horizontal plane) and vertical plane passing through three midpoints, between right and left Tragion, right and left Zygion, and right and left Ectocanthus (Yu et al., 2011).

Creating a parametric model using the Frankfurt Plane is not recommended, however. It leads to a very odd shaped model that is not representative of anyone.

Badawi-Fayad and Cabanis (2007) took advantage of 3D imaging and digitization techniques and recommended using *3D Procrustes superimposition* as a method of determining head orientation and alignment. Procrustes superimposition is an iterative least-square adjustment of all the figures after size normalization. It includes three phases: scaling, translation, and rotation. The skulls are scaled to have the same size, they are translated to have their geometrical centers fit exactly with one another, and finally rotated to minimize the gaps between anatomical points. Badawi-Fayad and Cabanis used the x, y, and z coordinates of 33 anatomical landmark points as input to the Procrustes alignment. While this was better than using the Frankfurt Plane it still results in oddly shaped cranial regions that are not representative of real people because nearly all the reliable anatomical landmarks are on the face and they are not evenly distributed.

Generalized Procrustes Analysis (GPA) was employed as head orientation alignment for the development of Statistical head models for facial animation by researchers such as Dai et al. (2020) and Li et al. (2017). Since most landmarks that are palpable or recognizable are on the face this has a strong bias for the face. For facial animation or for facial recognition this may be a reasonable approach. However, for product design it does not reflect how the product will be worn and the product will have different location biases.

For an AR/VR headset, where the product alignment in relation to the eye and line of sight is important for product fit and function, a head orientation deemed the "Neutral gaze" was used. The *Neutral gaze* pitch angle for each individual was determined during the 3D head data capture (scanning) phase. During the 3D head data capture phase, each individual's head was positioned in an approximate neutral and relaxed orientation. The procedure included: each user was positioned on a chair without a backrest, with both feet resting comfortably approximately hip distance apart on the floor and knees flexed at approximately 90 degrees. Each participant was instructed to sit upright with their shoulders relaxed back and down. They were then instructed to stretch the neck by looking up at the roof, and down at their chest, then looking right, looking left, and bending the neck first to the right (as if attempting to touch the right ear to the right shoulder) and then to the left. Then finally, participants were requested to sit upright again with shoulders relaxed back and down, and to look at themselves in a mirror set up perfectly vertical approximately 1.5 m in front of them. The problem with neutral gaze positioning is repeatability, as there is no control for the participants to position their head in exactly the same position each time. Reliability testing indicated that the pitch angles could range as much as ±6 degrees within one individual positioning his/her head repeatedly (Schnieders et al., 2023).

Almost similar to the neutral gaze head alignment, during the SizeChina-Hunan 3D head anthropometry data collection, the head orientation of participants being

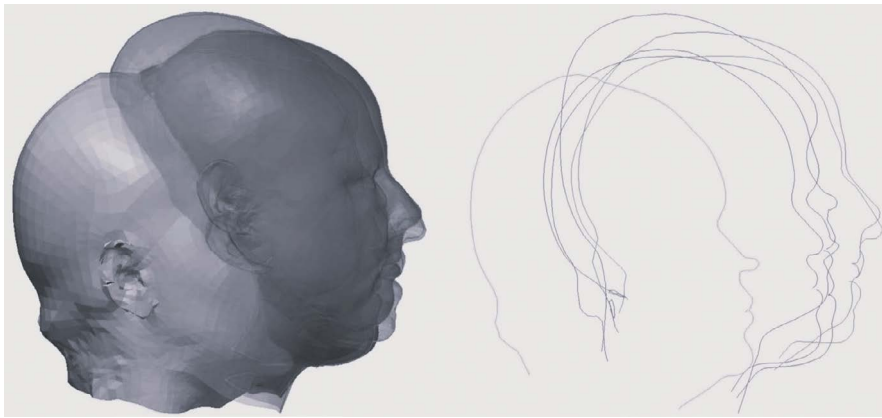


FIGURE 6.4 3D head scan origins for data collected with the same scanning system on different days.

scanned was defined by participants being instructed to look at a fixed point marked in front of them (Du et al., 2006; Luximon et al., 2015; Wang et al., 2018). A similar head orientation protocol was used when capturing 3D head scans for Australian bicycle helmet design (Perret-Ellena et al., 2015).

The origin is another aspect of head alignment that needs to be standardized when combining 3D head data for further analysis. When 3D head scan data is collected, the scanner coordinate system will determine the origin of the data. The 3D scan origin, zero point for all three coordinate axes (x, y, and z), is normally determined by the scanner calibration. Therefore, for some scanner systems, the origin could even vary between different data collection days on one system. Figure 6.4 illustrates head scans that were captured with the same head scanner system, but on different days (with the system calibrated daily).

In the absence of one consistent head coordinate system origin, Lee et al. (2016) investigated multivariate statistical shape variances for several different coordinate system origins. For all three alignment frameworks, the head rotational orientation (pitch, roll and yaw) was the same, with the main point of difference being the origin of the Cartesian system. The head scans were oriented with the x-axis passing through the right and left Tragion landmark points, and the y-axis passing through the Sellion and Supramenton landmarks. The different coordinate system origins included (1) Sellion landmark, (2) Pronasale landmark, and (3) midway between right and left Tragion landmarks. For the origin located at the midpoint between the right and left ear, the largest percentage of variance was described by the 3 PCs (89.7%), compared to the other two origin points (76.3% and 76.9% respectively). Different results are expected if the head scans were orientated in the Frankfurt plane.

In a study during which 3D face scan data was used for the analysis of the facial size and shape, Lee et al. (2017) aligned all the faces with the origin at the Sellion landmark, and the y-axis passing through the Sellion and Supramenton landmark, and the x-axis parallel to a line passing through the left and right Tragion landmarks.

Robinette (2007) explored three different anatomical alignment frameworks for helmet applications. These included: (1) the Principal Axis System (PrinAx), (2) an approximate corneal plane alignment (Eye), and (3) a top-of-head alignment (TopHead). For all three alignment frameworks, the head rotational orientation (pitch, roll, and yaw) was the same, with the main point of difference being the origin of the Cartesian system. The PrinAx method used all head surface points to define surface triangles. The area of each triangle was used to approximate mass distribution (assuming uniform density). Then, the center of mass and principal axes of inertia were calculated for each 3D head scan. The center of mass for all the triangles was used as the origin and the principal axes to define the x-, y-, and z-axes, or the three directions. The second and third methods begin with PrinAx to register all subjects together and establish the x, y, and z directions. The 3D head data was then translated to have an origin for all subjects at the midpoint between the right and left pupil for the second method (Eye) and at the top of the head for the third method (TopHead). The PrinAx method provides an orientation that uniformly distributes variability and is centered around an estimate of the center-of-gravity (cg) of the head which was believed to be a balance point for the head for biomechanical purposes. The Eye method was intended to simulate a pupil location restriction such as might occur for a helmet-mounted display. The TopHead method was intended to simulate a helmet resting position directly above the cg. These three alignment methods are demonstrated through side view and top-down cross-sections for 10 subjects in Figure 6.5.

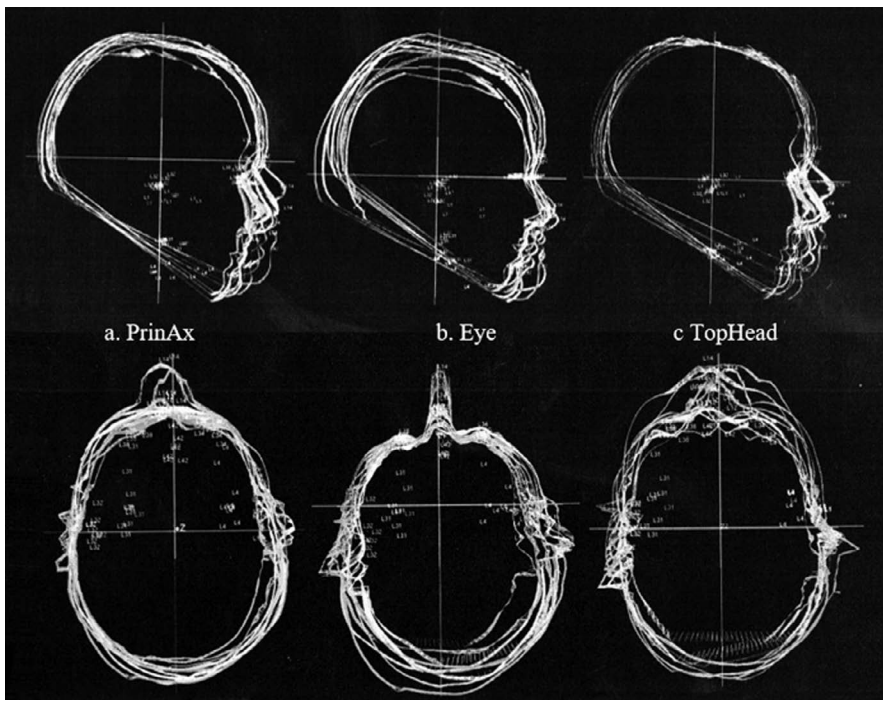


FIGURE 6.5 Three different head orientation methods for 10 subjects.

From the image above, it's evident that the PrinAx method most evenly distributes the variability throughout the head and face. This is similar to the result with a Procrustes Alignment of all the triangles and we can see that the faces are not well aligned.

The Eye method aligns the eyes and nose best but increases the variability on top and at the back of the head. The TopHead method has less variability at the top and back of the head but the most variability in the face, both vertically and horizontally. Even though the three methods all begin with the same basic alignment there are substantial differences that will impact fit.

A study was conducted to arrive at a head model for a Navy flight helmet used a sample size of 747 Navy pilots. They were all healthy and physically fit, and when they were selected for pilot training, they had no vision or hearing problems or physical anomalies that might cause them fit or functioning problems. They also had to meet Stature and Sitting Height standards to ensure they could fit into and fly an aircraft. At the time the head data was collected there were no female Navy pilots, so the sample is all male. Since the US population at that time was only 1% Asian and many of those had Sitting Heights, Seated Eye Heights and Leg Lengths that were too short to operate aircraft controls there were few Asians who made it through flight training. In other words, this was a uniform population compared to the civilian population. We can expect much more variability in these orientations with a more diversified and/or civilian sample.

With the "Neutral gaze" example described above, the 3D head data was manipulated into a Neutral Gaze Vector Coordinate system with the origin defined relative to the heads. For this, the heads were rotated (rolled and yawed) so that the right and left pupils were aligned on the same horizontal plane (since the headset will ideally be aligned to the eyes on a horizontal plane), with the x-axis passing through both pupils. The y-axis was pointing vertically upwards, and the z-axis towards the front. All 3D head data was then translated so that the origin of the coordinate system was at a point midway between the right and left pupil, 3 mm in front of the pupil position (to approximate the Corneal Apex position). This head orientation is demonstrated by the cross-sections (Sagittal and Transverse) for 10 subjects in Figure 6.6.

Out of all these head orientations and origins, which head alignment is the best? Only the product-based axis system. There is no generic axis alignment system that will align the head the way the product will be worn. A universally correct head orientation and alignment doesn't exist. Why is this important? It means we cannot effectively design for multiple people at one time without a product prototype, because we do not have enough information to overlay them properly. Different head models will be required for different head and face products. A "universal" head model in one set head alignment will most likely provide a false sense of head and face variance, since inevitably head orientation artifacts, such as forward-backward rotation (pitch), would not be applicable to how the product is worn.

So, we begin the design process by making educated guesses regarding how the product will fit on the head of one or a few human cases and creating a mock-up to test on the live subjects and adjust our product orientation. A careful

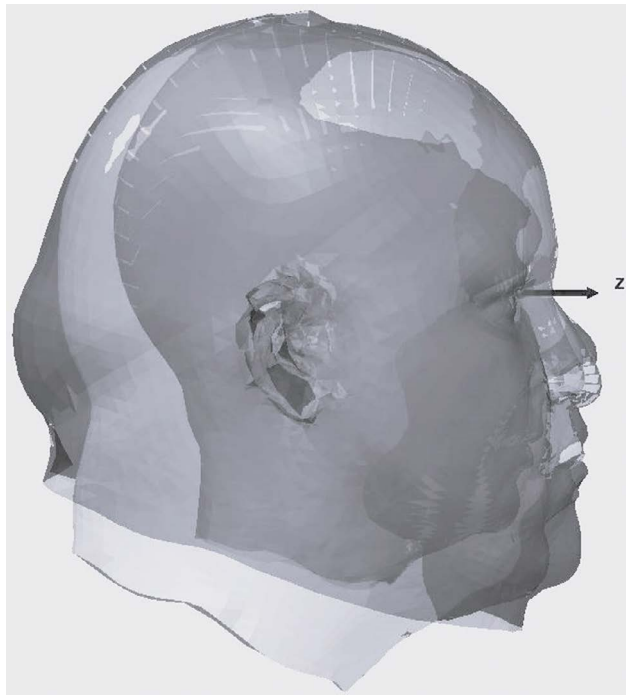


FIGURE 6.6 Alignment of 10 head scans in "Neutral Gaze" around the eye for AR/VR type headset.

understanding of how the product is expected to be worn is formulated before the head orientation is selected. This understanding could be formulated through an informal, small-scale fit test of similar products. With the use of widely used software tools such as MATLAB or Excel, or software programming, 3D head data can be re-oriented to the desired head orientation and alignment with relative ease. After testing mock-ups on human subjects CAD software with reasonably accurate 3D representations of the human head can help us not only re-orient the product to the head but can help us track and document the changes as well to help us avoid repeating mistakes.

For technology products, such as audio/communication headphones or AR/VR headsets, CAD is fundamental to the design process due to the complexity of internal components included in the product design. For head-worn products requiring expensive manufacturing processes, such as molding, CAD furthermore enables the use of computer-controlled manufacturing or prototyping processes. Since the manufacturing processes are usually costly and time intensive, optimizing the design to the human in the CAD space can help save time and money. The CAD file *alone* will not provide a reliable location for how the device *actually* fits on the head, but when used together with mock-up and prototype assessment on human subjects it is a very powerful tool.

Due to time constraints in the product development cycle, product management teams regularly put a lot of pressure on human factors specialists and user testing teams to reduce the time for actual fit testing. In the absence of these teams, the importance of fit testing is often deprioritized and "replaced" by CAD or physical head and face model fitting. This is a mistake that can result in the need for a complete re-design very late in the development process when changes are expensive and more time consuming. We will further highlight the risks associated with this approach and provide examples of how to overcome some of these challenges.

HEAD ANTHROPOMETRY

Since CAD is so valuable for the design and development of head wearables, we need to have good 3D anthropometry of the human head to import into our CAD systems. We also need good 1D anthropometry for characterizing, grouping, and analyzing fit data, as well as for selecting good cases.

It is important to keep in mind that, as we explained in Chapter 2, the digital version of the head that is compatible with CAD software is not an exact copy of the human head. Some file transformation is required from the original scan point cloud file to a format that can be imported into the engineering CAD software used for design. The modifications needed will depend on the engineering CAD software used.

Most of the time, the file size of the original scan (could be as large as 20 Mbytes) must be reduced to enable importing several head cases into a CAD file. File sizes can be reduced through decimation of the polygon structure or using software such as RS WRAP to create a uniform or homogenous polygon file. Files as small as 450 kB, with good enough resolution for use in design, are obtainable through this process. However, it is important to note that through the raw 3D scan cleaning and reduction process, the surface does not exactly match the original points.

Another area typically not well represented in the digital version of the human is the surface under hair. Although the differences between the digital version and the actual head could be reasonably small and insignificant, the differences on the scalp due to the hair are most likely very significant, especially for individuals with long hair and/or large hair volume. When capturing a 3D head scan, normally the user is prepared for the scan by placing a hair compressing cap over his/her hair. For persons with long hair, one method of scanning process regularly used is to have the participant tie the hair up in a high bun, place a hair compression cap over his/her hair, and tie off the hair bun with an elastic so that it's easier to identify and cut from the 3D data during post-processing. Even with the hair compression cap, differences are still observed between the surface captured and the skull. Figure 6.7 illustrates cross-sections of 3D head scans together with an ellipsoid comprising the head length and head breadth measurements for that subject, illustrating the size difference between a surface captured over hair (with hair compression cap) and actual skull size.

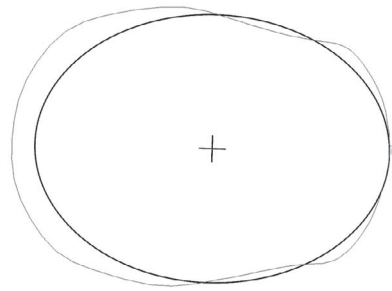


FIGURE 6.7 A cross-section of a typical head scan versus ellipsoid drawn from Head length and Head breadth.

For this reason, several researchers have worked on building head anthropometry databases that more accurately represent the scalp. Park et al. (2020) presented a database consisting of 80 bald scans of US Army males and 100 3D head skull data for US Army females. The US Army female head data included scalp points probed with a FARO Arm touch probe (coordinate digitizer) through four to five hair parts on each side of the midsagittal plane for a total of ~200 data points. The scalp point data were interpolated as a surface using the thin plate spline technique. The scalp surface was aligned with the 3D scan of the head and face based on the four digitized landmark points (right and left Tragion, Pronasale, and Sellion) and combined to obtain a full head surface. Li et al. (2022) developed a scalp probing rig which can be used to capture the actual 3D surface of the scalp in spite of users having hair (accuracy M = −0.2749, SD = 1.0153 mm). At the time of their presentation, they had collected 3D head data for 68 participants.

The short history of head wearables also means the availability of data for representing TPs is a bigger issue than for fashion apparel. Even though 3D head scanners have been available and in use since the mid-1980s (Robinette, 1986), most sample data readily available on the head and face are (1) traditional 1D measurements, (2) adults, (3) from military populations, and (4) contain only a handful of measurements on the head and face. Many companies may have collected head data, but they protect it as proprietary intellectual property. What data there is available, can still be useful for the Starting TP sample as an aid for choosing the first design cases. For example, the CAESAR survey (Blackwell et al., 2002; Hudson & Robinette, 2003; Robinette et al., 2002) collected data on adults ages 18-65 from North America and Europe and included only eight 1D measurements (taken with traditional tools), and nine 3D landmarks (taken from 3D scans) on the head and face. The measurements were:

- Bigonial breadth
- Bizygomatic breadth (face breadth)
- Face length (menton-sellion length)

- Head breadth
- Head circumference
- Head length
- Inter-pupillary distance (IPD)
- Sellion-supramenton length
- The landmarks were:
- Gonion, left and right: a point on the jaw; the lateral point on the corner of the mandible
- Infraorbitale left and right: lowest point on the inferior margin of the orbit (the bony eye socket) marked directly inferior to the pupil
- Nuchale: lowest point on the occiput that can be palpated among the nuchal muscles
- Sellion: point of greatest indentation of the nasal root depression
- Supramenton: point of greatest indentation of the mandibular symphysis, marked in the midsagittal plane
- Tragion, left and right: notch just above the tragus (the small cartilaginous flap in front of the ear hole)

Although the CAESAR survey collected 3D scan surface data, the data included whole-body 3D data only. The whole-body 3D scan resolution (density of the points) was 2–5 mm and, although this is good enough for body-worn items, this resolution provides poor detail around the eyes and ears. For head-worn equipment, this type of resolution is typically not sufficient to provide the surface detail and accuracy required. Most headwear products require sub mm scanner resolution and <2 mm precision on landmark placement and linear head and face measurement.

If there is a need to use 3D scanners for measuring, there are new scanning technologies available focused on scanning the head region only, which can meet this level of resolution and precision. Some of the newer systems, such as the one by 3dMD™, are fast enough to capture facial expressions and motion. This is referred to as a 4D system. This kind of technology can be very helpful for fit analysis, particularly if the scanner is used to capture people wearing the system while talking or moving. It can help identify when gaps occur or when a wearable is limiting motion(s) or slipping.

Most professional traditional anthropometry (manual) measurement tools such as tape measures and calipers can measure at a 1 mm resolution. Measurements taken using traditional anthropometry tools have the added benefit of being comparable to many other traditional head anthropometry databases. This allows the use of other TP data to help find cases and to compare fit data for assessing sizing needs. It is important to note that significant differences have been observed when comparing measurements taken using traditional anthropometry measurement tools, with measurements extracted from 3D scanners (Beaumont et al., 2017; Bredenkamp et al., 2006; Kouchi & Mochimaru, 2011; Simmons & Istook, 2003).

A few 3D scan databases exist that contain specific head scans with a typical resolution of a sub-millimeter. Examples of international head Anthropometry databases are listed in Table 6.1.

TABLE 6.1
Examples of International Head Anthropometric Data

Year	Population/s	Age Range	Type of Data	Size	Reference
2010–2012	United States of American Army (ANSUR)	17–58	1D measurements and 3D head/face scan	6068	Gordon et al. (2014)
2000–2004	South African military (SANDF)	18–65	1D measurements and 3D head/face scan	4000	RSA-MIL-STD 127 Vol 1 (2005)
2006	Chinese population	18–75+	1D measurements and 3D data	1563	Wang et al. (2018); Ball (2011)
2003	US Civilian respirator users (NIOSH)	18–66	1D measurements and 3D head/face scans	3997 & (1013)	Zhuang and Bradtmiller (2005)
2006	Chinese Workers - 1D head and face	18–66	1D measurements	3000	Du et al. (2006)
2003–2004, 2010	Korean civilian population	1–90	1D measurements, 3D head/face scans with landmarks	2076 (2702) [938]	Kim et al. (2017)
2014	Australian population	18–80+	3D head/face scans with HTO	222	Perret-Ellena et al. (2015)
2015–2018	United Kingdom Civilian (The Headspace dataset)	1–89	3D head data	1518	Headspace dataset (https://www-users.york.ac.uk/~np7/research/Headspace/); Dai et al. (2020)
2009	United States of American Civilian	22–75	3D face scans (some landmarks)	105	Texas 3DFRD (http://live.ece.utexas.edu/research/texas3dfr/index.htm); Gupta et al. (2010)

The product requirements, specification, function and intended fit of the product should give some indication about which variables, landmarks, and demographics are important for a given product. As far as possible, it is recommended to stick to established head and face landmarks, used in other anthropometry or head and face product design studies. This will enable comparison and potentially help to compare with other resources.

Sources for landmark ideas can be found in anatomy and physiology, and physical anthropology textbooks such as Gray's Anatomy (Gray & Goss, 1973). One helpful source for the head and face is the classic text on Human Osteology by William Bass

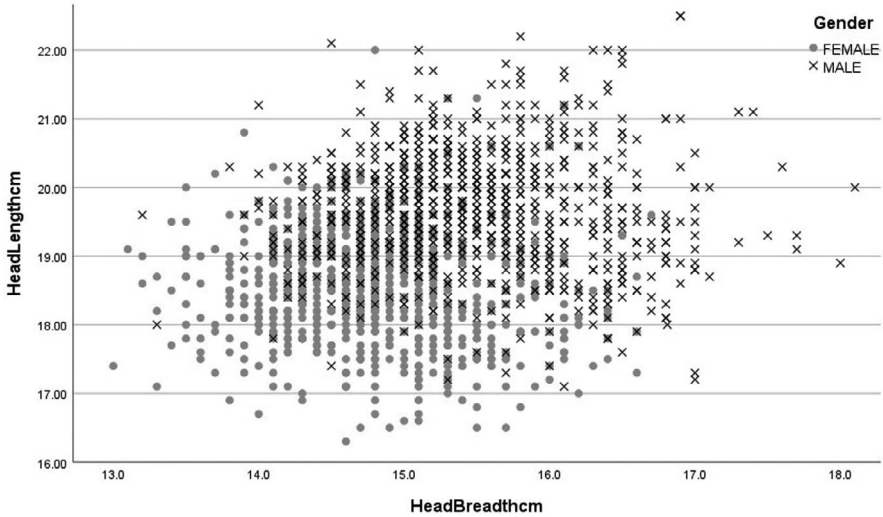


FIGURE 6.8 Illustration of differences between genders for head length and breadth.

(Bass, 1971). Designed as a laboratory and field guide it has excellent illustrations and descriptions for bony landmarks. While originally published in 1971, new editions continue to be published and it is used by colleges and universities throughout the world. White had a more up-to-date osteology book indicating bony landmarks on the head and face (White et al., 2012). Although it is recommended to keep to established and widely used head and face landmarks, deviations might be needed to best suit the fit of the specific head and face worn product. If new landmarks are created for a specific product use, these should be clearly defined and documented to ensure accuracy in repeated studies.

For head wearables, it is very important to be aware of differences due to gender and ethnicity when confirming representation of the TP, since the differences in these groups on head and facial characteristics are substantial. They will need to be considered when selecting cases and when doing testing. An example of gender differences in two head measurements is seen in Figure 6.8.

There are also substantial differences related to ethnicity. Using the same sample, we plotted the head and face measurements for three different ethnic groups by gender in Figure 6.9. In Figure 6.9, we see that Asians tend to be larger for Head Breadth and smaller for Head Length than either White or Black populations. In other words, they have a different head shape. This may not be a factor for stretchy headwear like headbands but is likely to be an issue for more rigid headwear such as helmets or eyeglasses. It will be important to use Asian subjects to design and test headwear intended for an Asian TP. It is important to note that aspects such as eye depth, eye height in relation to the nose, and nose angle are affected by the orientation of the face versus the wearable. If a specific orientation was chosen for the selection of cases as input to the design or selection of test cases, variances in the fit of the device

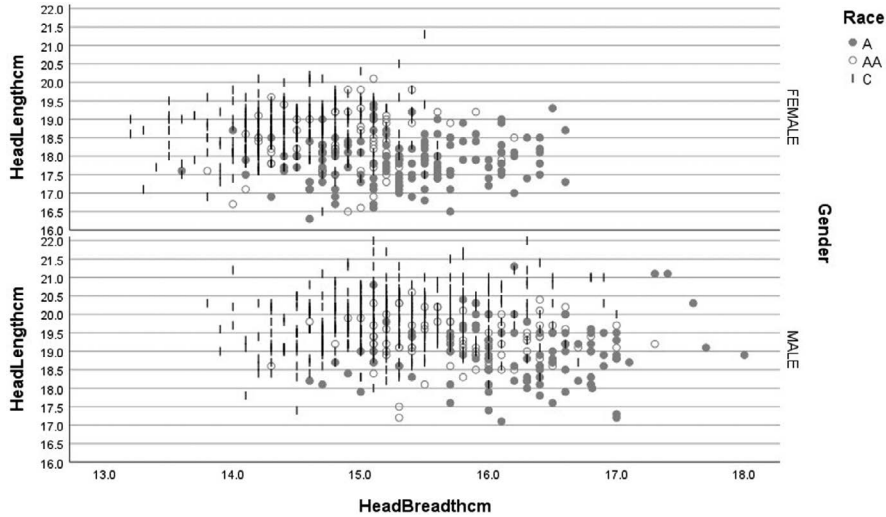


FIGURE 6.9 Head length and breadth for different races.

in relation to the orientation of the head could be expected. Therefore, final fit testing must be completed to verify the actual product fit.

It is furthermore important to note that the sample used in Figure 6.9 as one collected in North America and the Asian subjects were predominantly raised in North America. They may be different than people born and raised in other countries including China, Japan, or Korea, because diet and environment play a role in human physical development. This may suffice as the Starting TP sample, but if the TP is not a North American population it will be important to collect fit test samples and Full TP samples from the target region.

CASE STUDIES

The following case studies are presented as examples of applying the Sustainable Product Evaluation, Engineering, and Design (SPEED) process to head wearable product design. Each case study will follow the same format. Firstly, the aspects demonstrated in the case study will be listed, followed by the example of those aspects.

Case Study 1: Design Loop Testing of a Head Wearable to Demonstrate Use of Product Based Head Orientation

This case study demonstrates the typical test preparation and test flow and procedures applied during a design loop test. This test furthermore demonstrates the use of a product based head orientation and alignment system to identify feature envelopes as input

to the design. Feature envelopes are described in Chapter 2 (see Figure 2.14). Aspects demonstrated in this case study includes:

- Design loop test preparation
- Design loop test flow and procedures
- Product based head orientation and alignment

The purpose of the test was to assist in the design of head-mounted displays and other interfacing equipment such as hearing protection and oxygen masks. An existing helmet system was used as a prototype for establishing the location of eye, ear, nose, and mouth features with the helmet in place. This example is an excerpt and summary from Whitestone et al. (1998) with some modifications to remove jargon, clarify the procedures, and hone in on the parts of that study that are relevant to head wearable testing and data collection.

Test Preparation

Before the data collection, the test procedures were pilot-tested to validate them and determine the number of team members and data collection stations needed to process each subject in less than one hour. The subjects used for the first pilot test were the potential team members themselves. Four stations and five team members were determined to be necessary.

All team members were trained in all procedures, as were several people who were to serve as backups in case someone became ill or needed to be away during data collection. Being trained in all procedures, enabled team members to change stations when necessary to avoid having to delay data collection, and to reduce fatigue. Test subjects were recruited to participate in the study. Some of these were employees of the organization and some were paid contract subjects. The training ensured all team members were measuring and assessing consistently and permitted the optimization of tasks at each test station. In this respect, the training also served as additional pilot testing. After team training, the tasks and number of members at each of the four data collection stations were established.

Test Flow and Procedures

The first station was manned by one team member. Subjects were briefed on the reasons for collecting anthropometric data, as well as on the safe use of the measuring systems. Subjects were then asked to read and sign a consent form and fill out a brief biographical form with their demographic information.

The second station was manned by one person. The anatomical landmarks were located by palpation or visual inspection and marked on the subjects with an eyeliner pencil. Some of these landmarks were used as measuring points for manual tool measurements while others were used as reference points in the scans. Additional landmarks, such as pupils, were later located by visual inspection of the scanned image.

The third station was manned by two people. Data were collected using manual anthropometric tools. This station was staffed by a measurer and a recorder. The recorder entered the data into a laptop computer as the values were called out

by the measurer. The recorder checked the data for obvious errors by comparing the subject's percentile values with the percentiles from other large databases. The recorder also assisted in measuring and positioning the subjects. After the traditional anthropometry data were collected, stickers were placed on the subject's face over the marked landmarks to make them more visible and consistent in size in the 3D scans. To ensure all of the landmarks had a sticker in place, the roll of stickers was pre-divided into strips with the exact number of stickers needed. This helped speed up the process.

The fourth station was the scanning station, which was manned by one person. At this station each subject was properly positioned for the scan. A skull cap which reflects light and compresses the hair was placed on the subjects' heads. This step results in clearer images in the scan and a better representation of the head surface.

After the regular head scan was made, two additional scans were made of each subject:

1. One with the chin raised to get the surfaces under the chin and on the top of the head that might not be visible in the first scan
2. One with the subject wearing his or her personal helmet and oxygen mask

Data were collected on 365 subjects from the target population of Air Force flight crew members. There were four data collection sites, and the same procedures were used at all four locations. Since these were active-duty flight crew members, they each had their own helmets and oxygen masks that had been previously issued to them with the aid of a skilled flight equipment technician. These were the items scanned. It was assumed that these items fit them.

The 3D scanner used was the Cyberware 4020-PS 3D Digitizer, which was one of the first automated 3D scanners used for anthropometric data collection. It collected over 131,000 three-dimensional data points in approximately seventeen seconds. The resolution of the scanned points varied depending on the radius of the object being scanned but was approximately 1.5 mm vertically and 1 mm horizontally over the entire head and face.

This version of the scanner did not have a color camera to make flat-colored landmark stickers visible in the scans. This feature was added in later versions of the scanner (Hoffmeister et al., 1996). For this study's landmarks green felt stickers were used. These stickers absorbed light, making uniform round holes in the scans over the landmarks. These holes had to be filled so that the landmark location could be identified and recorded.

After data collection, the scans were individually processed and edited. To identify landmark locations, create a polygonal mesh, and a solid watertight surface model a special software tool called Integrate was used (Burnsides et al., 1996). All 3D scans were run through an editing procedure where scans were displayed and landmark locations, indicated by the uniform circular holes from the markers were visibly inspected, and any landmarks found to be incorrectly identified or placed were corrected. The location of the center of each filled hole was recorded and linked to a named landmark.

This procedure resulted in the following files:

- Cleaned and edited surface subject scan without a helmet (unencumbered scan)
- Associated unencumbered file with 3D coordinates for named landmarks
- Cleaned and edited surface subject scan with the helmet (encumbered scan)
- Associated encumbered file with 3D coordinates for named landmarks

Product Based Head Orientation and Alignment

Once all the scans had been thoroughly processed and checked, a procedure was performed to create what was referred to as "feature envelopes" for the pupils and the Tragion points on the ears. Feature envelopes are the point distributions for named landmarks with respect to the axis system of the wearable. In this example, they indicated where the eyes and ears fall when wearing this helmet and oxygen mask. The process used can be applied to any wearable that has relatively rigid materials such as a helmet. It was as follows:

1. Scan a representative helmet (i.e., same helmet model, same size that the subject wore) by itself
2. Select helmet landmark coordinates from the helmet scan using visualization software
3. Create a helmet-based coordinate system for alignment, using the helmet scan
4. Register the representative helmet scan with each subject's encumbered scan by the best alignment of the helmet surfaces
5. Register the two registered helmet scans (with and without the subject) to the subject's unencumbered scan by best alignment of visible head and face surfaces
6. Transform the subject's landmarks coordinates into the helmet-based coordinate system
7. Repeat for all subjects & generate bivariate plots and summary statistics for each landmark in the helmet-based coordinate system

Dimples on either side of the helmet were used to define the x-axis, with the positive direction pointing to the left of the helmet. The z-axis runs through the midpoint of the front edgeroll and is orthogonal to x. The y-axis is orthogonal to the x- and z-axis. This is illustrated in Figure 6.10.

After registering the three scans together and transforming the landmarks for each subject into the helmet-based axis system bivariate plots (two axes at a time) of each landmark were viewed in the helmet-based system. The plots for the right and left pupils are shown in Figure 6.11.

In summary, these helmets were situated on each subject's head in the way they were worn in flight. Therefore, this spread of pupil locations is a reasonable way to begin designing a helmet-mounted display for this helmet and this TP. It indicated the type of variation that can be expected in each direction. The spread in the x direction is not as great as the spread in the y and z directions, for example.

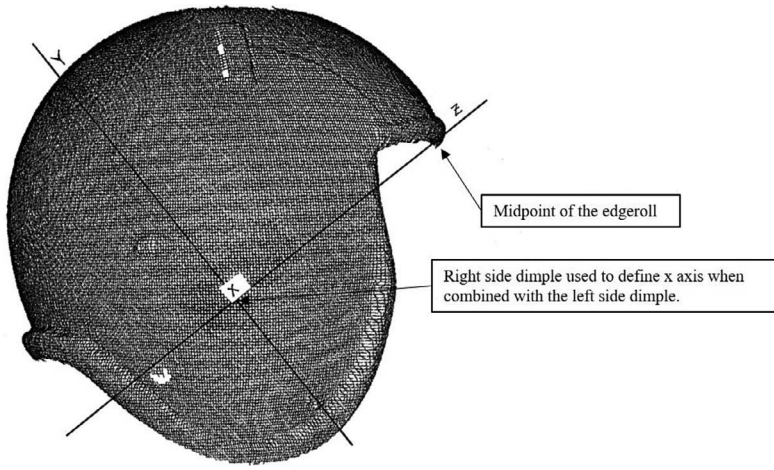


FIGURE 6.10 Helmet-based axis system.

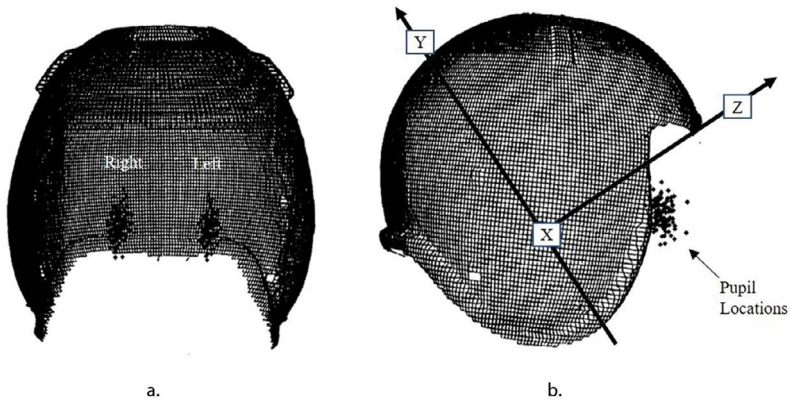


FIGURE 6.11 Pupil distributions in the helmet-based axis system. (a) Front view of helmet with right and left pupil locations. (b) Side view of helmet with the distribution of both pupils overlaid.

Case Study 2: Design Loop Testing with a Non-Functioning Mockup Including Early COF

This case study demonstrates the use of non-functioning prototypes to do the first design loop tests. Early design loop tests are used to start formulating an understanding of how a product will fit, as well as to evaluate and update the ranges of adjustments provided in the design. Aspects demonstrated in this case study includes:

- Early prototype COF metrics, including methods for simulating non-functional aspects of a device for early accommodation and fit evaluation
- Sampling method
- Examples of design aspects informed by early prototype testing

FIGURE 6.12 Early prototype nosepieces used in the design loop fit tests.

An AR/VR headset was the head wearable item being tested. This headset when fully functional would have optics to project the virtual content in front of the users and have eye-tracking cameras to track the eye movement. The headset was designed to fit around the head like a headband with a nosepiece resting on the nose and a forehead pad resting on the forehead. The prototype used in this case study represented the final product in physical size, shape, weight, and center of mass (COM), but had no functional optics or other electronic components. The device was intended to have one size that would fit all persons, with adjustability afforded in the temple arms and the headband. There was further positioning adjustment with the availability of different nose pieces and forehead pads as seen in Figures 6.12 and 6.13.

COF Metrics

During the COF metrics compilation phase, factors to be included in the fit metrics were largely informed by previous experience with similar products. Since the product fit was affected by functional performance in addition to physical size, shape and weight, metrics needed to be compiled to evaluate the expected functional performance, despite that functionality not being present in the prototypes yet. The full list of COF metrics is listed in Table 6.2. The COF metrics were grouped into the following three categories:

1. Physical fit
2. Visual registration
3. Slippage and stability

The COF for *physical fit* had two opposing fit aspects: on one hand, the band must be able to open large enough to easily fit over and around their head, and on the other, it must adjust small enough in respective areas to have contact in designed contact areas including a) the forehead, b) the nose, and c) around the sides and back of the head.

FIGURE 6.13 Early prototype forehead pads with three different thicknesses.

TABLE 6.2
Example of COF Metrics for Early Design Mockups

No	Fit Criteria Description	Measurement Method	Acceptable Fit
		Physical Fit	
1a	Head circumference fit wrt device	Visual assessment of how device fits around the head when donning	Device must open wide enough to fit around head and that it can be donned easily
1b	Forehead pad contact on forehead	Visual assessment of how device fits on face (use pen light when inspecting gaps)	Small gaps allowable, but forehead pad must make some contact with forehead
1c	Nose piece contact on nose	Visual assessment of how device fits on face (use pen light when inspecting gaps)	Small gaps allowable, but nose piece must conform to the nose bridge width and shape
1d	Temple pads contact on head	Visual assessment of how device fits on face (use pen light when inspecting gaps)	Temple pads must have contact with all sides of head
		Visual Registration	
2a	Vertical/height location of display	Participant wears visual registration prototype. Proctor views at prototype eye slots at participant eye level.	Proctor sees participant pupils through prototype slots (see Figures below)
2b	Horizontal distance between display and eye	Participant wears visual registration prototype. With participant eyes closed, push distance measure through the prototype slot to softly contact eyelid.	Distance measure touches eyelid and is fully inserted into provided slot (±3 mm)

(*Continued*)

TABLE 6.2 (*Continued*)
Example of COF Metrics for Early Design Mockups

No	Fit Criteria Description	Measurement Method	Acceptable Fit
		Slippage and Stability	
3a	Slippage at nose landing point	Mark landing point on nose before and after activities: • moving the head in roll, pitch, yaw movement • Standing up from chair and sitting down • Walking • Completing number identification and block removal on a Jenga tower Measure distance between marked landing points	Any two positive indications of slippage: > 3 mm between marked nose landing points > 5 mm reduction in vertical Tragion-temple arm distance and/or visual identification of clear reduction in the vertical Tragion-temple arm distance on side photo's "Yes" to participant subjective question on slippage "Yes" to proctor subjective question on slippage
3b	Slippage at temple arms	Take side photo before and after activities: Measure vertical distance between Tragion landmark on the ear and bottom of the device temple arms before and after activities: • moving the head in roll, pitch, yaw movement • Standing up from chair and sitting down • Walking • Completing number identification and block removal on a Jenga tower	
3c	Participant identified slippage	Subjective question to participant: "Did you notice the device slip at your nose, temples or back of the head during the activities?"	
3d	Proctor identified slippage	Subjective question for proctor: "Did you notice the device slip at the participant's nose, temples or back of the head during the activities?"	

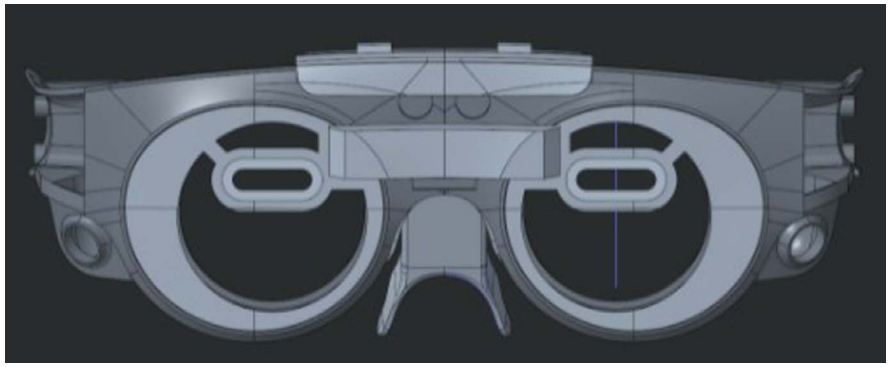


FIGURE 6.14 3D printed mockup for evaluation of height and depth distance to user's eyes.

To ascertain a good fit for *visual registration*, the horizontal, vertical and depth ranges for the position of the pupil in relation to the optics components were theoretically determined. Visual registration included visual performance, eye tracking performance, and visual comfort. The visual performance included the percentage of the virtual display area that was visible to the user. The eye-tracking performance included the ability of the eye-tracking cameras to capture the full pupil, iris/cornea and a minimum number of glints on the iris/cornea. The visual comfort included the position of the virtual content in relation to the eye. Since the prototype had no functional optics or eye-tracking components, the actual acceptable visual registration of the device could not be evaluated. For this reason, the prototype test device included a 3D printed add-on (see Figure 6.14), which enabled a guide and measure of acceptance to the placement of the optics section of the device in relation to each test participant's pupils. This included a depth (z-distance) measure and acceptance range between the inner optics surface and the user's eyelid and a height (y-distance) measure and acceptance range of the pupil in relation to the optics components. These were both included as part of the COF metrics.

Figure 6.15 illustrates the use of the 3D insert for evaluation of vertical (height) placement of the device in relation to the pupil. The vertical slots are used to line up the vertical height of the wearable with the user's eyes by using a taller or shorter nosepiece. During the evaluation, if the middle of the pupil could be seen in the provided slots, the height of the device was deemed acceptable. Since parallax would play such a vital part in the correct judgment of the vertical position of the device, a camera was used for the judgment. The camera was set up on a tripod and the camera lens was aligned horizontal to the participant's pupil height. The "Neutral Gaze" protocol (see the "Head orientation and alignment Paragraph in Chapter 6) was used for Head orientation and to position the participant. The participant's vertical eye height was measured, and the camera tripod height was adjusted to match the floor to center of lens height to the participant's vertical eye height. Figure 6.15a demonstrates when the wearable fitted too high on the face and was adjusted lower on the face by using a shorter nosepiece. Figure 6.15b illustrates a wearable height that was correctly placed in relation to the participant's eyes. If no higher or lower nosepiece

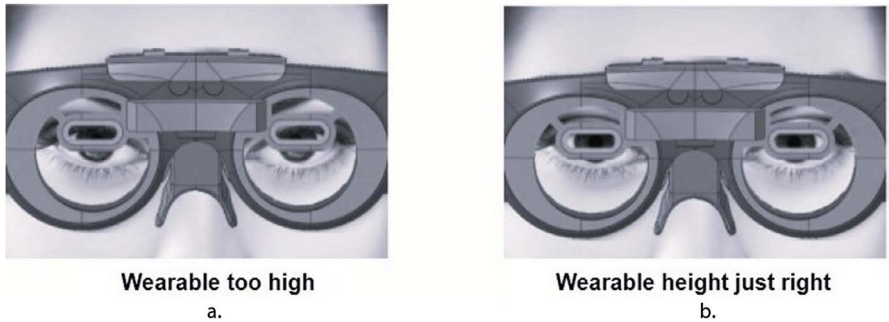


FIGURE 6.15 (a) Wearable too high on face and (b) wearable height just right on face.

adjustments were available to obtain an acceptable wearable height for a participant, the fit was indicated as unacceptable for the participant.

Figure 6.16 illustrates an insert used for the evaluation of the depth (z-distance) between the wearable and the eye. Once the device was in the correct height (vertical) position on the participant's face, the participant was asked to close their eyes, and the insert (see Figure 6.16a) was used to gauge the horizontal distance between the user's eye and the optics by sliding the insert through the slot provided (see Figure 6.16b,c). If the depth distance was at the bold line ±3 lines (mm), the depth was deemed acceptable. The intended pupil-to-device optics depth was modified by 3 mm to accommodate the addition of the approximate eyelid thickness to the theoretical optimum pupil-to-optics distance. If the depth was too close, the forehead pad was exchanged for a thicker forehead pad. If the depth was too far, the depth was decreased by exchanging to a thinner forehead pad. If no thinner or thicker forehead pads were available to decrease or increase the depth respectively, a fit failure was recorded in the fit evaluation sheet.

The COF metrics for slippage and stability included a battery of functional movements selected to represent physical activities that (1) product users are expected to undergo while wearing the device, and (2) previous product testing indicated are likely to invoke slippage and/or instability of this type of headset. Since the device hardware (optics and other electronics) and software were not functional, users could not perform typically expected tasks and activities while wearing the headset.

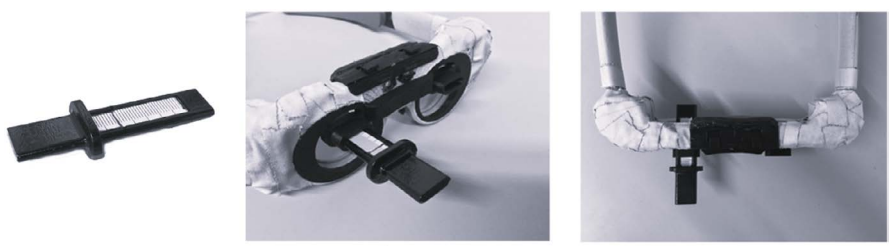


FIGURE 6.16 (a) Insert used to test the depth between the device and the eye, (b) demonstration of a depth that is too shallow, and (c) demonstration of a depth that is correct.

Therefore, expected tasks and activities were broken down into typical expected activities and body movements, including:

- Standing up from a chair and sitting down
- Walking
- Standing in one place and rotating the neck in pitch, yaw and roll rotational movements (see Figure 6.2). For pitch head movement, participants were instructed to "Nod their head 'yes'", for yaw head movement participants were instructed to "Shake their head 'no'", and for roll, participants were asked to move their head side-to-side as if to touch their ears to their shoulders
- Flexing the trunk forward while extending and rotating the neck. These actions were simulated in a Jenga game. A large Jenga tower was set on a low (< knee height) table. Participants were asked to search, identify and remove 6 marked Jenga blocks and place them on top of the tower.

During slippage COF pilot testing, the measures for how to identify slippage were further refined. These included:

- Slippage at the nose landing point
- Slippage at temple arms, including a) visual identification of changes in temple arm location from photos and b) a reduced Tragion-Temple arm measurement
- Participant's subjective identification of slippage
- Proctor's subjective (visual) identification of slippage

To evaluate slippage on the nose, the landing point of the nosepiece on the nose was marked by drawing a line on the nose (using an eyeliner pencil) at the front edge of the nosepiece where it landed on the nose. After completing the slippage activities, another line was drawn on the nose (using a different colored eyeliner pencil) at the final resting point of the nosepiece (see Figure 6.17). At the end of the study protocol,

FIGURE 6.17 Landing points on the nose before and after slippage, marked using different colored eyeliner pencils.

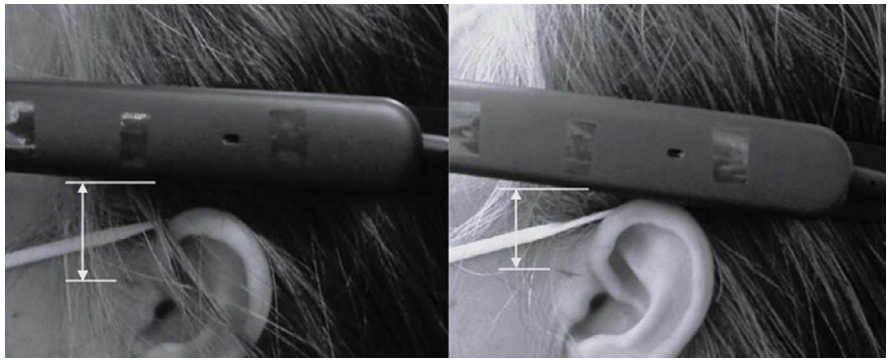


FIGURE 6.18 Side pictures indicating distance between Tragion landmark and bottom of temple arms before and after slippage activities.

when the headset was removed, the distance between the top (highest point) of the two lines was measured using a Caliper. If the distance exceeded 3 mm, it was marked as a fail.

To evaluate slippage at the temple arms, the Tragion landmark was drawn on the participant using an eyeliner pencil. The vertical distance was measured using a Caliper between the Tragion landmark and the bottom edge of the wearable temple arms (see Figure 6.18). In addition to the Tragion to Temple arm vertical distance, side photos were taken before and after all slippage activities to visually assess if the temple arms slid down relative to the ear on the participant's head. If the vertical Tragion to Temple arm distance was reduced by more than 5 mm and the slippage was visible on the side photos, Temple arm slippage was indicated as positive.

Finally, a question was asked to the participant about whether slippage of the back temple arms was experienced during any of the activities to gauge the participant's subjective identification of slippage. Similarly, the proctor was asked to give his/her subjective input on whether he/she observed slippage during any of the activities. Finally, a minimum of two out of the four metrics had to be a positive indication of slippage before slippage was recorded for an individual.

Sampling Method

Due to cost limitations, a relatively small sample size had to be used for this early design prototype test. However, a reasonably good representation of head and face size and shape variance needed to be required to best test accommodation and fit. In this example, a panel of approximately 1500 participants, representative of the TP was available from which participants could be recruited. For each participant in the panel, head and face measurements as well as a 3D base Anthropometry head scan with face landmarks, were available. All the heads in the database were orientated in what was considered a "Neutral gaze" head orientation, and the head scans were all aligned with origin midway between the right and left pupil, 3 mm in front of the pupil (x-axis runs through the pupils, y-axis runs positive vertically upwards and z-axis runs positive anteriorly). The key variables to be represented in the small

sample size were based on the fit of the product as well as previous knowledge of head shape variance in the population. These variables included:

- Head breadth
- Head length
- Inter-pupillary distance (IPD)
- Depth position of the eyes in relation to the forehead (represented by Glabella-to-pupil depth distance)
- Height position of the eye in relation to the nose's preferred landing area (bony region) (represented by Rhinion-to-pupil height distance)

Figures 6.19–6.21 illustrate the distribution of these variables, the 98% boundary ellipse as well as target recruitment cases. Thirty-one target cases were identified to meet the representation of the intended distribution. Participants were selected from the existing target pool database to represent these cases as closely as possible during this prototype user test. In addition to the anthropometry representation, participants were selected to ensure representation of the three main race groups identified in the TP, including Asian, Black African and Caucasian, and genders, including male and female. Despite best efforts, all cases were not necessarily well represented due to target participants not being available on the specified test dates. Out of the identified target cases, 23 participants (12 females and 11 males) were used for this study. Cases that were unacceptably represented included overall small and overall large head cases (see Figure 6.19) and large IPD & Glabella-to-pupil depth and small IPD cases (see Figures 6.20 and 6.21). The mismatch between the recruitment target and reality is a common finding in user testing, and unfortunately the practicality of user testing that researchers must

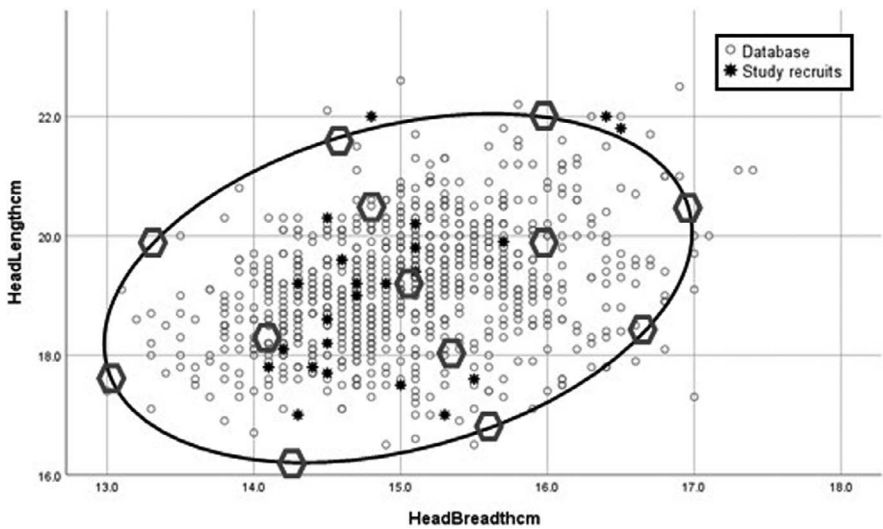


FIGURE 6.19 Bivariate plot with head breadth and head length, 98% boundary ellipse, recruitment target and study participant distribution.

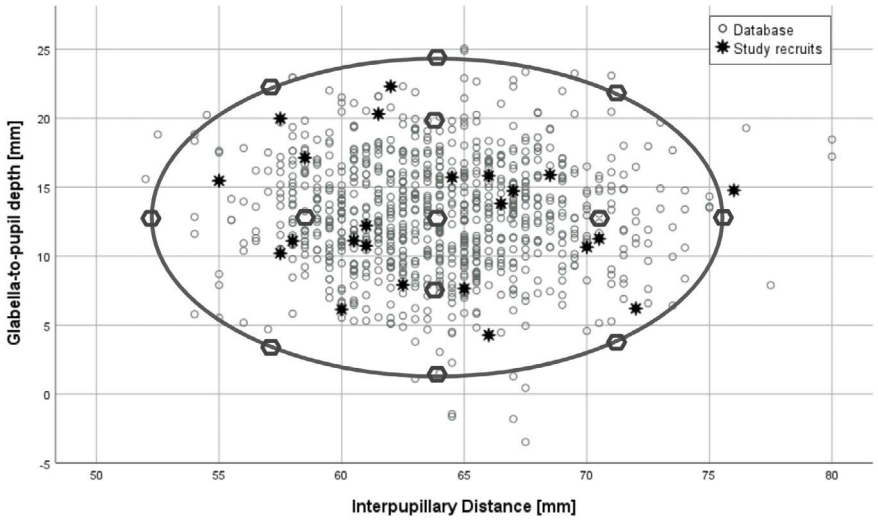


FIGURE 6.20 Bivariate plot with IPD and Glabella-to-pupil depth, 98% boundary ellipse, recruitment target and study participant distribution.

live with. Depending on how large the mismatch is, researchers could decide to continue recruitment until close enough to the targeted participant distribution has been obtained, or this caveat can be noted to the current study findings. In this case, the research team will have to focus future product testing on the underrepresented target demographic.

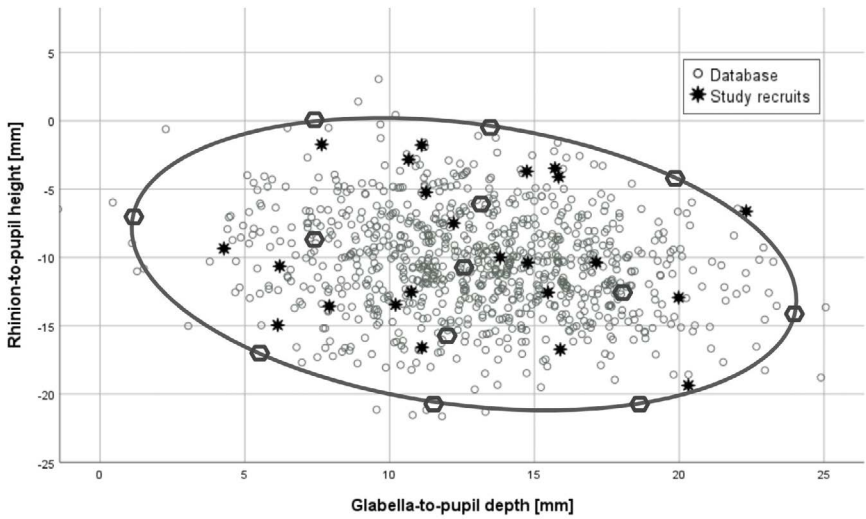


FIGURE 6.21 Bivariate plot with Glabella-to-pupil depth and Rhinion-to-pupil height, 98% boundary ellipse, recruitment target and study participant distribution.

Analysis and Results

The COF metrics, in addition to providing evaluation criteria for fit, also guide the device design aspects that can be informed using the early prototype testing. In addition to the fit aspects included in the COF, user comfort following a wear period can also be evaluated with early prototype testing.

For physical fit, the evaluators looked at (1) does the device fit around the head, and (2) is good contact obtained in the areas that the device is designed to contact against the face including (a) forehead pad, (b) nose piece, and (c) temple pads. The fit testing indicated that the device could be comfortably donned by all test participants, demonstrating that adequate temple band adjustment was afforded in the design. One caveat in the findings was that the test participants did not represent the largest head in the TP and future device fit evaluations will have to target persons representative of large (long and wide) heads.

For device fit in design contact areas, the testing indicated that the forehead pad and temple pad for all participants had acceptable contact at all the pad surfaces. For the nose piece contact, however, the evaluation highlighted that the nose piece was too narrow and didn't conform well to the nose shape variances to be accommodated. Figure 6.22 illustrates some examples of where poor nose piece accommodation was identified.

Although undesired cheek contact was not one of the original COF metrics for physical fit, it was identified as an issue during fit testing. When doffing the device after a wearing period of 30 minutes, an indent at the cheek area was identified for some participants. This is illustrated in Figure 6.23. Although this contact was not deemed uncomfortable by affected participants. The contact was noted for future considerations, especially to include this area for thermal comfort and safety assessment.

Visual registration was evaluated during fit testing as well as post-fit testing through analysis of collected data. During the fit test, the COF metrics for visual registration were recorded and evaluated. This included the vertical and depth

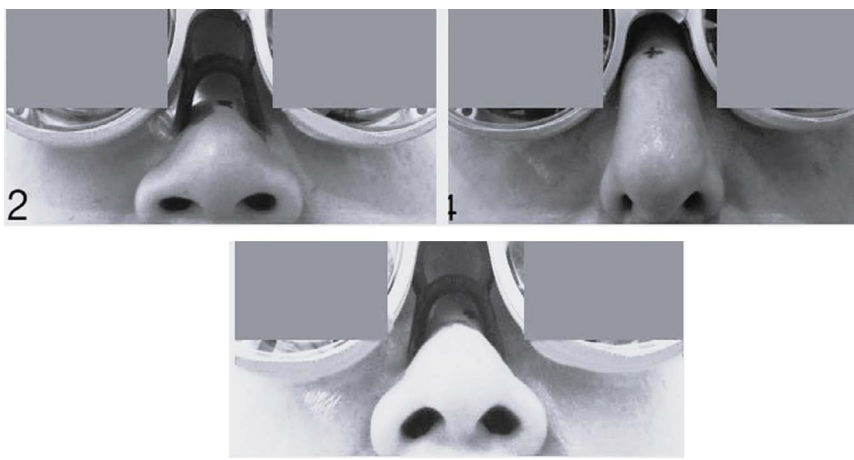


FIGURE 6.22 Poor nose piece accommodation identified during fit testing.

FIGURE 6.23 Undesired cheek contact identified during fit testing.

alignment of the device in relation to the eye position. The test results indicated that all participants could be fitted within the specified height and depth ranges.

During the fit testing, a fit scan was also captured. Figure 6.24 illustrates the 3D head scanner used for capturing fit scans during the study.

During the scan data analysis, the fit scan was aligned to the base anthropometry scan using least-squares point alignment of a selection of 10–20 points on the face. Thereafter, the headset CAD file was aligned to the fit scan using the same method. Finally, the fit scan was removed leaving the headset CAD aligned to the base anthropometry scan (Figure 6.25 demonstrates the process). With this alignment, several different scan analyses could be performed. All scan alignment and analysis were performed using Polyworks Inspector™.

Through this alignment process, the actual fit position of the device eyepiece could be positioned in reference to each participant's pupil point. The difference between ideal/design and actual device fit (right and left eyepiece center points) in relation to the participant's eye position could be calculated. These values for this

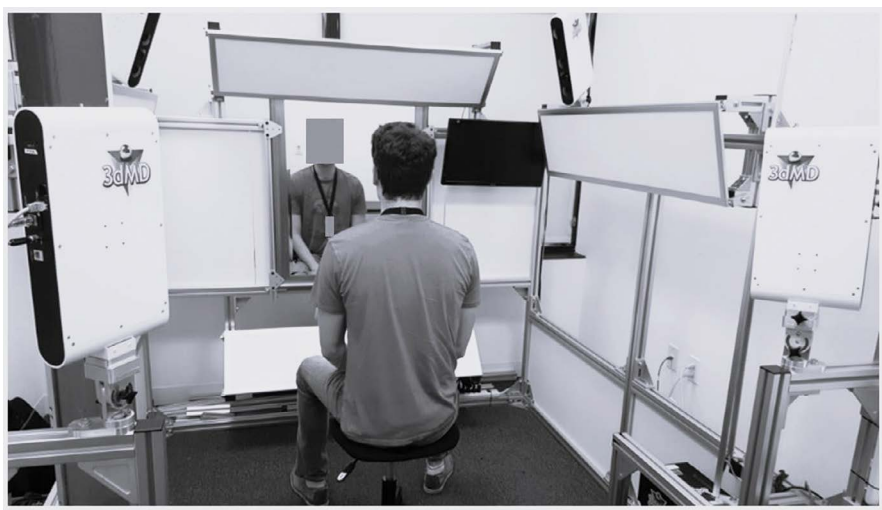


FIGURE 6.24 3dMD laser scanner used for capturing 3D head anthropometry and fit scans.

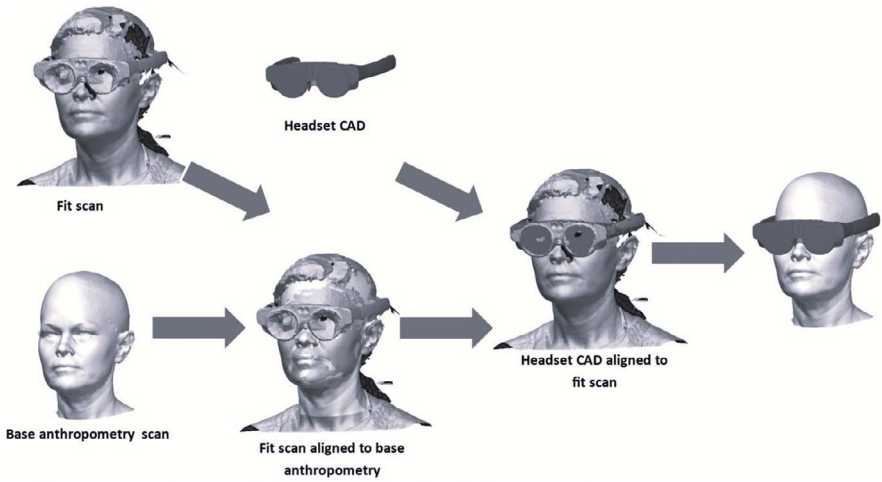


FIGURE 6.25 Scan analysis alignment process.

test is illustrated for the right eye in Figure 6.26. Also included in this figure, is a theoretical visual registration fit volume that was calculated to indicate acceptable visual registration of the device, therefore acceptable percentage of virtual content visibility, placement of virtual content and acceptable pupil and iris view by the eye tracking cameras. From Figure 6.26 it is evident that the placement of the device fell outside the theoretical visual registration fit volume for 4 of

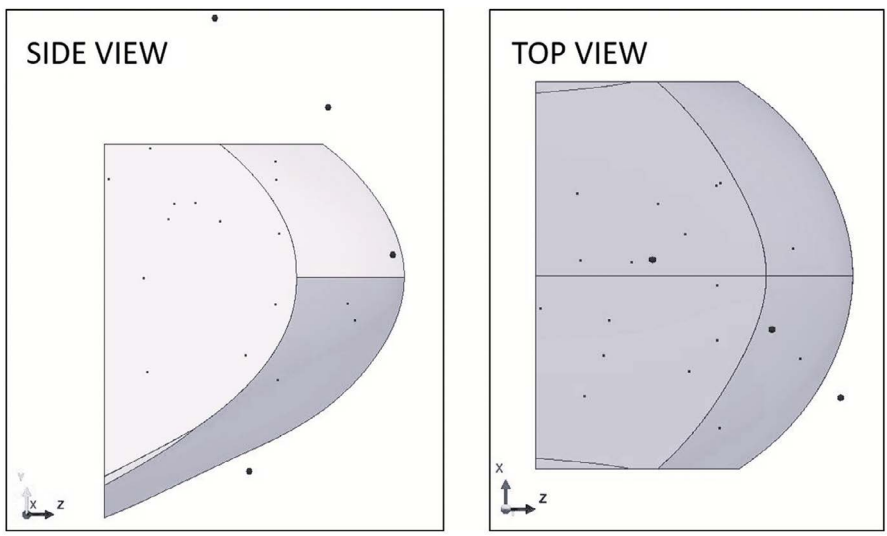


FIGURE 6.26 Error in eye position from ideal/design eye position plotted in relation to the visual registration boundary.

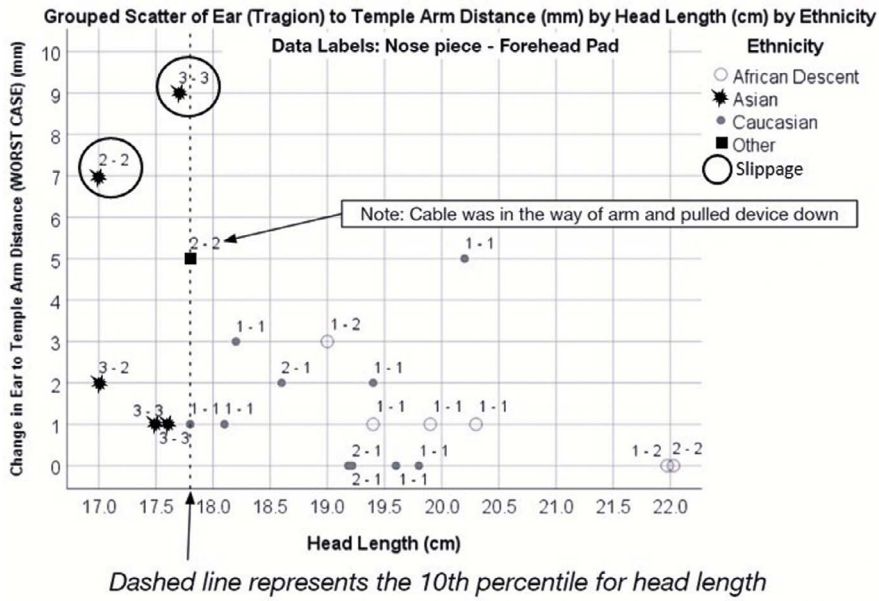


FIGURE 6.27 Change in Tragion-Temple arm measurement before and after slippage activities, participants with slippage marked.

the participants. For these participants, further investigations could be conducted to understand what modifications could be needed to ensure better accommodations. However, in this instance, the deviation from the theoretical fit volume was so small, that the team decided that no modifications would be made to the fit components at this stage, and that visual registration with a functional display and eye tracking system would be confirmed before further design modifications would be made.

For slippage and stability, the results indicated that two out of 22 participants had the device slowly sliding down while completing the slippage activities. Figure 6.27 indicates the change in Tragion-Temple arm measurement from before to after the slippage activity (y-axis) plotted against participant head length (x-axis). The two participants that were identified with slippage issues are highlighted in the figure. It was also identified that both cases had head lengths below 17.8 cm (10th percentile), indicating that head length could be one of the contributing factors to slippage. It was however, also noted that not all participants with a head length <17.8 cm had slippage, indicating that other factors were contributing causes to slippage in addition.

This prototype test included a user test that provided valuable information regarding fit accommodation and requirements to improve the fit interface for the cheeks and nose as well as identified issues of head size variance that contributed to unacceptable slippage.

Case Study 3: Design Loop Updates in COF

This case study demonstrates how the COF metrics can change as the product design matures and specific design fit issues are identified through fit testing. This is a follow-on from Case Study 2 using a more mature prototype in the fit testing. Aspects demonstrated in this case study include:

- Updated prototype COF metrics, including methods for testing fit with a functional prototype

Modifications to the prototype included modifications to the housing and nose piece fit components to improve fit based on earlier fit evaluation findings. This prototype included most of the internal electronics, including a functional eye-piece stack and eye tracking. At this point in the product development cycle, the display and eye-tracking software algorithms were still under development. This meant that the combined binocular virtual image could not be projected, and the eye-tracking algorithm determining eye-tracking success wasn't running yet. To represent the lack of these features for virtual display, an image was projected monocularly (to one eye at a time), and by calculating the visible area for each display individually, the visibility of the virtual display area by both eyes was calculated. For evaluation of the eye tracking performance, the eye tracking cameras were streamed and eye images recorded. Eye images were post-processed and evaluated for acceptance criteria. Another aspect added to the fit metrics included undesired device contact. With the addition of internal electronics, it was determined through simulation and testing that the surface temperatures at the projector bumps could exceed the safety limit for 8-hour skin contact. As a result, the physical fit metrics had to be updated to include undesired contact between the projectors and the skin. The weight and COM were representative of updated projections for the final headset. The device did not include a computer and battery pack yet but was powered from a loose-standing printed circuit board. In addition to the three categories included in the Case Study 2 COF list, an additional category for Comfort was added. Due to improvements to the design, it was mature enough to start evaluating aspects of comfort to enable design modifications focused on wear comfort. The updated COF included the following four categories:

1. Physical fit
2. Visual registration
3. Slippage and stability
4. Comfort

Physical fit metrics predominantly remained the same, with the addition of two metrics for nosepiece fit and one for undesired contact. Previous fit testing highlighted the potential for poor placement of the nosepiece affecting fit and safety in relation to the eye and breathing. If a forehead pad was used that pushed the device forward on the face, it could push the nosepiece too far forward on the nose, causing the nosepiece to no longer fit on the bony region (Maxilla bone), but on the soft tissue region causing it to restrict breathing to a degree. Also, if a higher nosepiece was used with an especially lower forehead pad, the nosepiece angle could cause the

TABLE 6.3
Concept-of-Fit Metrics

No	Fit Criteria Description	Measurement Method	Acceptable Fit
		Physical Fit	
1a	Head circumference fit wrt device	Visual assessment of how device fits around the head when donning	Device must open wide enough to fit around head and that it can be donned easily
1b	Forehead pad contact on forehead	Visual assessment of how device fits on face (use pen light when inspecting gaps)	Small gaps allowable, but forehead pad must make some contact with forehead
1c	Nosepiece contact on nose	Visual assessment of how nosepiece fits on nose (use pen light when inspecting gaps)	No gaps at the top/bridge of the nose
1d	Nosepiece poking into the eyes	Visual assessment of how device fits on face	Nosepiece not poking into eyes
1e	Nosepiece interference with airways	Subjective question: yes/no question	No interference with breathing
1f	Temple pads contact on head	Visual assessment of how device fits on face (use pen light when inspecting gaps)	Temple pads must have contact with all sides of head
1g	Projector contact	Visual assessment of IF the projector bumps contact any part of the user's skin	Projector bumps not contacting skin
		Visual Registration	
2a	Perceived % of virtual display visible	Participant visually identifies the outermost letters for which letter and circle is fully visible on each radial line, from 1 to 8 (see Figure 6.28). This is repeated monocular for right and left eye. The % of virtual display is calculated for each eye. Also, the ambinocular FOV (content seen by either or both eyes) is calculated.	>85% ambinocular FOV
2b	Pupil vertical position	Visual assessment of center of pupil vertical position visible in ET camera eye images. Two ET images are available per eye (total of 4 ET images)	Middle of pupil must fall within middle half of ET image for at least 1 image for EACH eye
2c	Pupil and cornea appearance as per eye images (RIGHT and LEFT eye) when gazing at center gaze point	Visual assessment of pupil and cornea appearance visible in ET camera eye images. Two ET images are available per eye (total of 4 ET images)	Whole pupil and cornea visible in camera view in at least 1 image for EACH eye
2d			Aspect ratio of pupil < 3:1 (height: width)

(*Continued*)

TABLE 6.3 (*Continued*)
Concept-of-Fit Metrics

No	Fit Criteria Description	Measurement Method	Acceptable Fit
		Visual Registration	
2e	LED glint count per eye image (RIGHT and LEFT eye)	6 LED glint pattern is projected on each eye. Number of glints are counted on each eye image. The glint quality is also taken into consideration. Glints that are too dim or split are not considered good glints.	At least 3 good glints must be visible on at least 1 image per eye
		Slippage and Stability	
3a	Slippage at nose landing point	Mark landing point on nose before and after activities: moving the head in roll, pitch, yaw movement Standing up from chair and sitting down Walking Completing number identification and block removal on a Jenga tower Measure distance between marked landing points	Any two positive indications of slippage: >3 mm between marked nose landing points >5 mm reduction in vertical Tragion-temple arm distance and/or visual identification of clear reduction in the vertical Tragion-temple arm distance on side photo's "Yes" to participant subjective question on slippage "Yes" to proctor subjective question on slippage
3b	Slippage at temple arms	Take side photo before and after activities: Measure vertical distance between Tragion landmark on the ear and bottom of the device temple arms before and after activities: moving the head in roll, pitch, yaw movement Standing up from chair and sitting down Walking Completing number identification and block removal on a Jenga tower	
3c	Participant identified slippage	Subjective question to participant: "Did you notice the device slip at your nose, temples or back of the head during the activities?"	
3d	Proctor identified slippage	Subjective question for proctor: "Did you notice the device slip at the participant's nose, temples or back of the head during the activities?"	
		Comfort	
4a	Comfort acceptance	Subjective question to participant after 30 min of wear: "How would you rate the overall comfort of the device on your head?"	95% of participants should rate 3 or better on 5 point Linkert type scale

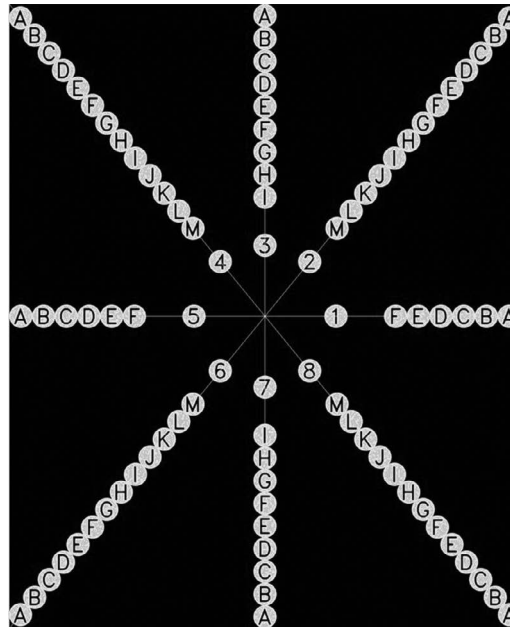


FIGURE 6.28 Left eye image for evaluation of virtual FOV (see COF metric 2a in Table 6.2).

ends of the nosepiece to fit too close/on the inner corners of the eye (Endocanthus). The metric for undesired projector contact was furthermore added. Fit metrics were added to the Physical fit COF to prohibit these poor fit results.

The *visual registration* metrics changed entirely with the increased maturity of the device. The visual registration included two COFs: (1) virtual field of view (VFOV) and (2) eye tracking (ET) performance. For the assessment of VFOV, an image was projected monocularly to each, first the left and then the right displays respectively (see Figure 6.28). Eight lines are projected from the center of the image radially with line number 1 on the medial side. For the right display, line 1 was in the middle left and the numbers increased clockwise to number 8. For the left eye, line 1 was in the middle right of the display and the line numbers increased counterclockwise to number 8. Participants were asked to read the outermost number for which they could see the full number and circle (no part of the circle may be cut off). Each circle represented 2 degrees in VFOV. The visual registration data collection Graphics User Interface (GUI) included a calculation of the percentage of the display visible monocularly, as well as ambinocularly (visible with any one or both eyes). If the ambinocular VFOV exceeded 85%, the fit was deemed acceptable.

For the assessment of ET, the two ET camera images per eye were streamed to the visual registration GUI. The GUI included lines indicating the acceptable vertical location of the pupils in each image. An acceptable fit meant that (1) the center of the pupil fell within the middle half of the ET image for all the ET images, and (2) for at least one of the two images per eye, all the following criteria were met: (1) the pupil and iris/cornea not cut off, (2) the aspect ratio of the pupil <3:1, and (3) more than two good glints visible on the iris/cornea. The eye images for a good fit are illustrated in

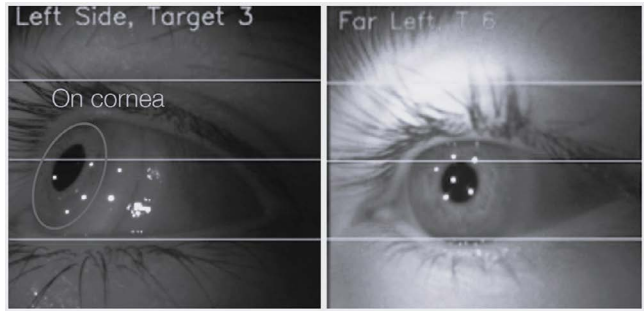


FIGURE 6.29 Good fit ET eye images.

Figure 6.29. Figure 6.30 illustrates two examples of poor fit, the first show the pupil and cornea being cut off in the ET image, and the second, a pupil aspect ratio <3:1 and <2 glints on the pupil.

The slippage and stability metrics in the COF remained unchanged from the previous design loop prototype testing.

An additional metric, a Subjective Comfort was added to the COF metrics. During testing, each participant had to wear the prototype for 30 minutes continuously. During this time, participants were engaged in activities, some for entertainment and to reduce their focus on the device fit and potential discomfort. The activities included walking to and participating in being 3D head scanned, completing all the slippage protocol actions (see those listed in Case Study 2 in this chapter), and then watching a choice of documentary or short films until the 30-minute duration had expired. On completion of the wear duration, participants' subjective comfort was solicited through the question: "How would you rate the overall comfort of the headset on your head?" Further comfort questions were asked about specific regions on the face and head using the same rating scale. Only overall comfort was however included in the COF to ascertain acceptance of fit. Participants were given a printout of the scale to indicate their answer to the question.

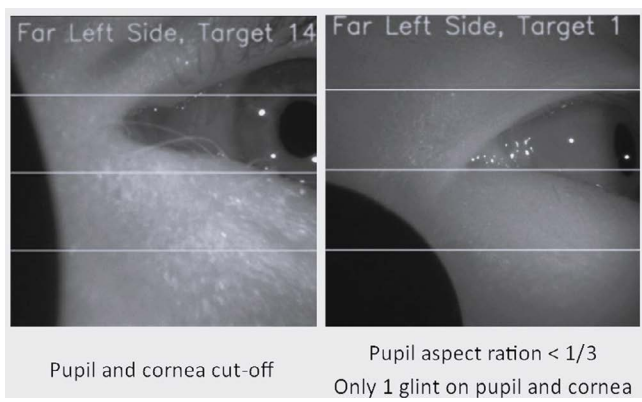


FIGURE 6.30 Poor fit ET eye images.

In summary, this case study demonstrated how the COF metrics, including the means of measuring fit, were improved as the test prototype matured. In Case Study 2, fit measurement relied on the simulated placement of the device since no functional electronics were available. In this case study, fit was evaluated by measuring what portion of the virtual display is visible to the user and ensuring that it was higher than the design intent. Also, an image from the eye tracking cameras were streamed to confirm the correct placement of the device in relation to the user's eyes. These improved methods of measurement of fit ensure more accurate evaluation of fit as the product nears its final design iterations.

Case Study 4: A Trade Study to Investigate Temple Band Closing Force

This case study involves a design loop trade study to determine the best spring setting affecting the closing force of a temple band. Aspects demonstrated in this case study includes:

- Study design and sampling method
- Data analysis method
- Results and discussion of results leading to a final decision

For this headset (see Figure 6.31), an internal spring force is used to auto close the temple bands. This force must be strong enough to close the temple bands when the

FIGURE 6.31 Headset for which temple arm closing torque was investigated during a trade study.

headset is not worn, and to retain the position when a user closes it to a comfortable tightness on their head. This means when in the correct fitting position, it must not loosen over time while being worn. At the same time, the force must not be so tight that it causes pressure discomfort to the user.

Background

During a prototype fit test, the prototype design included a temple band closing torque setting of 50 N mm. During this test, the percentage of participants that experienced the temple band slipping on their head, (13%), was slightly higher than intended for the design. One of the design changes suggested for implementation was an increased temple band closing torque. However, the design team was not confident which torque setting would provide the intended temple band stability, while not causing pressure discomfort around the head. In addition, the design and engineering teams were not satisfied with the way in which the temple band closed to its default closed position. Some of the prototype devices were not closing to the default temple arm setting. A larger temple band closing torque was suggested as a potential solution for this issue as well. As a result, a follow up fit test was recommended where two larger temple band closing torque settings, 65 N mm and 80 N mm, would be tested. The goal of the test was twofold: (1) to determine if a temple arm closing torque increase would improve stability or an improvement in the return of the arms to the default closed position and (2) to determine if the temple band closing torque increase would result in increased discomfort.

Study Design and Sampling Method

For the follow-up testing, the order of the 65 N mm and 80 N mm conditions were randomized in a repeated measures study, with each participant testing both conditions. Since we wanted to know if these torques (65 and 80) improved the situation we found in the first test, we also compared the follow-up test results to the original 50 N mm test. A different sample of participants was used in the first fit test with 50 N mm than the second repeated measures test with the 65 N mm and 80 N mm, so repeated measures analysis was not possible for this comparison. The same protocol was used during all tests. For the first temple force condition, 50 N mm, N = 54 randomly selected participants were used. For the follow-up test, N = 40 randomly selected participants were randomly fitted with both devices

To determine which temple band closing torques would be best, two opposing fit criteria were identified for evaluation: (1) the slippage and stability, and (2) the 30-minute wear comfort of the device. The device was set up to its optimal fit for each participant by the investigator and the device operating temperature was maintained throughout the study.

An overview of the study flow is provided in Figure 6.32. To ensure that the device temperature was stabilized at operating temperature, the device was switched on at least 30 minutes before the study started. Before the participant entered the room, the proctor measured the device surface temperatures at predetermined locations using an Infrared thermometer (FLIR®). The measured values were compared to the expected device operating surface temperatures and recorded in the questionnaire. On arrival of the participant, the proctor provided a briefing of the study and

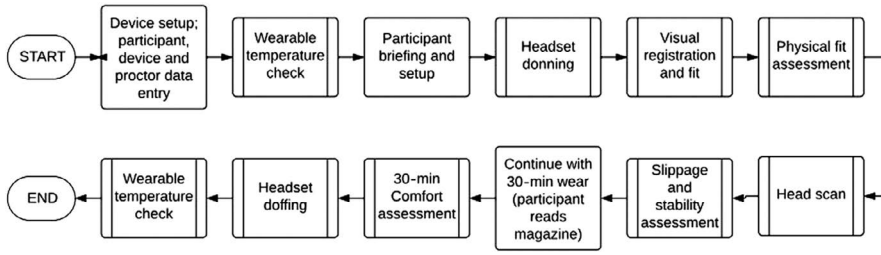


FIGURE 6.32 Study flow during the temple band closing torque trade study.

explained the process that was to be followed to fit and optimize the headset for the participant.

To optimize the fit of the headset for the participant, the device was set up in its default configuration, which is with nose piece # 1 and forehead pad #1, to start. The participant was handed the headset, and the proctor explained how the device should be worn. The participant was asked to put the headset on themselves. All participants put the device on themselves since a proctor cannot guide slight micro-adjustments that might be needed to ensure that the device feels balanced on the face and head and that the optimal positioning on the nose, forehead, sides and back of the head for a comfortable distribution of the load and pressure. After the headset was donned, the proctor went through a process to decide which fit components, nose piece and forehead pad, provided the most optimum fit per participant. This process is described in more detail in the paragraph "Determining best fit adjustment" below. A physical fit assessment was performed followed by a 3D surface capture of the participant wearing the headset. The stability and slippage of the device were assessed after completing the head scan. The detail of activities and metrics included in the stability and slippage is described in more detail in the paragraph "Assess stability". Following the slippage and stability assessment, the participant was asked to sit comfortably on a chair while reading a magazine for the remainder of the time up to 30 minutes. After the 30-minute wear time, the proctor asked a series of comfort questions. These are described in more detail in the paragraph "Comfort assessment questions". On completion of the comfort assessment, the participant removed the device for the end of their study participation. After the participant was escorted out from the study room, the proctor repeated the measurement of the surface temperatures to verify that the device was still at the intended operating temperature range, and the temperature values were recorded in the questionnaire. If the temperature values fell outside of the intended operating thermal range, the participant's comfort results were excluded from the analysis.

Determining Best Fit Adjustment

The headset design had a selection of 5 nose pieces and 2 forehead pads, to provide the optimal fit allowing for the best visual registration for the participant. The nosepieces were overall the same shape, with the primary difference being the length of the stem, starting from #1 being the shortest to #5 being the longest. Nosepiece #5 also had a slight inward curve. The forehead pad had two thicknesses, with #1

FIGURE 6.33 Five nosepieces and two forehead pads used to adjust fit per participant.

being the thinnest and #2 being the thickest. The nosepieces and forehead pads are presented in Figure 6.33.

To identify the optimal fit configuration (nosepiece and forehead pad), each participant was processed through a fit sequence. Starting with nosepiece #1 and forehead pad #1, then nosepiece #2, then nosepiece #3 and so on. The fit result for each fit configuration was recorded. During the fitting, an image was projected on the display of the device binocularly to the user (see Figure 6.34). The user was asked to stare at the grey center cross on the display. A second, colored box indicated whether the device position was correct in height and depth placement in relation to the user's eyes. An acceptable device position was indicated by a box color turned yellow or green. In addition, if the correct number of good quality glints on the eyes were registered by the eye tracking system, a LConf and RConf would indicate Medium or High. An acceptable visual registration (device position and glint count) is illustrated in Figure 6.34. A green box color and High LConf & RConf was preferred to yellow box color and/or Med LConf or RConf value. Therefore, the iteration of fitting with nosepiece and forehead pads were continued until a green box color and High LConf & RConf was obtained, or all nosepieces and forehead pads were fitted. If no green box color and LConf or RConf value was registered, a yellow box color and/or Med LConf & RConf was accepted. A nosepiece change resulted in a vertical, upwards shift in the colored box over the grey box, and a forehead pad change from #1 to #2, reduced the size of the colored box. The box color turned green when the height and size of the colored box corresponded to the grey box position and size. An unacceptable placement of the

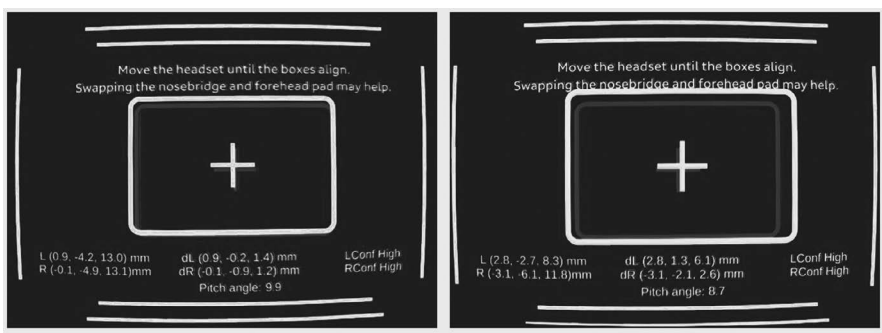


FIGURE 6.34 Good (left) and acceptable (right) visual registration.

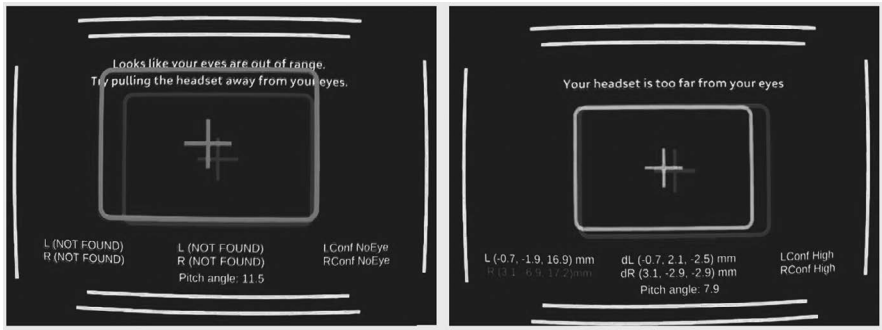


FIGURE 6.35 Unacceptable visual registration.

device in relation to the user's eyes and/or unacceptable glint count is illustrated in Figure 6.35.

Assess Stability

To assess slippage of the device, each participant was requested to complete a series of activities, previously identified to induce slippage. These included:

- Standing up and sitting down on a chair
- Walking between study rooms
- Completing a Jenga game activity, where the participant had to search for and retrieve 6 blocks marked. (This activity induced neck rotation and flexion, and trunk flexion movements.)

To identify slippage, the metrics included subjective proctor identification of slippage, subjective participant identification of slippage, and a Tragion-to-temple arm vertical measurement. At least 2 out of the 3 measures had to indicate slippage for the case to be accepted as slippage.

Assess Comfort

To assess physical comfort, a subjective questionnaire was administered after each participant had worn the headset in its best-fit configuration for that participant, for a duration of 30 minutes. A 10-point Likert scale was used to assess comfort in all the different interface areas of the device, with 1 being bad and 10 being good. The areas for which comfort was evaluated included:

1. Overall
2. Forehead
3. Nose
4. Side of the head (above the ears), right and left
5. Side of the head (further back behind the ears), right and left
6. Back of head

Data Analysis

To investigate the significance between differences observed between the slippage findings (categorical data of "Y" and "N"), a z score test for two population proportions was run. The z score test for two population proportions was used instead of the t-test, since the slippage results are categorical data ("Y" and "N"). In order to investigate the difference between comfort results for the different conditions, a comparison of means was used since ordinal data (10-point Likert scale results) were compared. Data collected during two different studies were included in this trade study; therefore, two different types of tests were used to investigate how the comfort results compare between the data samples. First of all, an independent t-test was used to compare the data collected during the first prototype test, 50 N mm, to the data collected during the second tests, 65 N mm or 80 N mm. An independent sample t-tests was used here since data was collected on two different groups of people during the first and second studies. Secondly, a paired-sample t-test was used to compare the datasets collected for the 65 N mm to 80 N mm force conditions. The paired-sample t-test could be used here since the same people repeated both the 65 N mm and 80 N mm force condition tests. t-Tests were used instead of the non-parametric alternatives, the Mann-Whitney U-Test or Wilcoxon Signed Rank Test, due to the relatively large sample size ($N > 20$) and the t-test having greater statistical power. The datasets collected during the two different studies could be compared to each other only due to the same protocol being used during both studies. If all the studies were run with the same participants, a repeated-measures ANOVA could have been run to compare all three to each other. All statistical tests were performed using SPSS analysis software and an on-line calculator for the z score test for two population proportions (https://www.socscistatistics.com/tests/ztest).

Results

The slippage findings are illustrated in Figure 6.36. No or slight slippage was deemed acceptable, but a positive Slippage result as per the fit guide was deemed unacceptable. Between the 50 N mm (9% Slippage), 65 N mm (7.5% Slippage) and 80 N mm (5% Slippage) conditions, a slight reduction was observed in slippage as the temple band force increased. Therefore, a higher force would be preferred to reduce the prevalence of slippage.

The differences in slippage (with "Y" = "Slippage" and "N" = "No slippage" or "Slight slippage") were investigated between the condition with the most slippage (50 N mm) and least slippage (80 N mm) to determine if the differences observed were deemed significance. A $z = 0.7777$ (with two-tailed $p = 0.4354$) was calculated for slippage between 50 N mm and 80 N mm. Since $p > 0.05$, this difference observed in slippage was not deemed statistically significant. Since the best and worst slippage results indicated no statistically significant difference, the slippage for 65 N mm will similarly indicate no statistically significant difference.

To conclude, the slippage results between the conditions were not statistically significant. Therefore, no matter which temple arm force would be selected, no statistically significant difference in slippage will be expected.

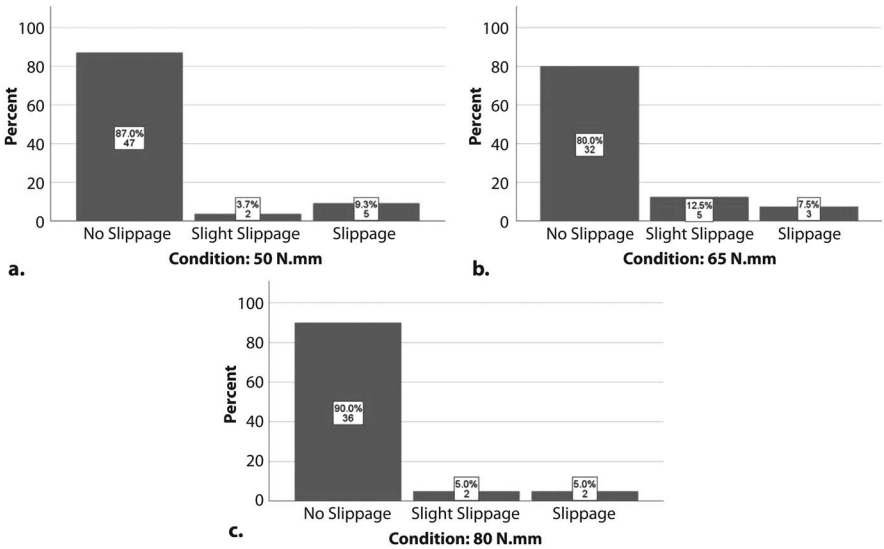


FIGURE 6.36 Slippage distribution for the (a) 50 N mm, (b) 65 N mm, and (c) 80 N mm torsion band closing torque conditions.

The results from the paired samples t-test indicated no statistically significant difference between 65 N mm and 80 N mm for overall comfort, nose or forehead comfort. A statistically significant difference was observed in comfort on the right and left sides of the head (above the ears and further back). The results for the paired-sample t-test between 65 N mm and 80 N mm is presented in Table 6.4. Looking at mean differences (Column "Mean" under "Paired Differences" in Table 6.4), even though the difference was found statistically significant, the mean difference is less than 1. On a 10-point comfort scale anything less than 1 point is for most people not perceivable. A difference in 1 point between findings is therefore not practically significant. Therefore, even though the differences in right and left side pressure statistically detectable (found statistically significant), the research team deemed the magnitude of the difference too small to be practically significant.

To compare the comfort results between the 50 N mm and the 65 N mm or 80 N mm conditions, an independent-samples t-test was conducted between the 50 N mm and 65 N mm conditions, and secondly the 50 N mm and 80 N mm conditions. The results from these tests are indicated in Tables 6.5 and 6.6. The results between the 50 N mm and 65 N mm indicated, similarly to the difference between 65 N mm and 80 N mm, no statistically significant difference between overall comfort, nose and forehead comfort, but a statistically significant difference between the right and left side of head (above the ears and further back). Also, similar to the difference between 65 N mm and 80 N mm, the mean differences (see column "Mean difference" in Table 6.5) between the side of head comfort values between 50 N mm and 65 N mm conditions, were less than 1,

TABLE 6.4
Paired Samples t-Test between 65 N mm and 80 N mm

Paired Samples Test										
		Paired Differences					t	df	Significance	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference					
					Lower	Upper			One-Sided p	Two-Sided p
Pair 1	Overall_65–Overall_80	0.000	1.754	0.277	–0.561	0.561	0.000	39	0.500	1.000
Pair 2	Nose_65–Nose_80	–0.225	2.703	0.427	–1.090	0.640	–0.526	39	0.301	0.602
Pair 3	Forehead_65–Forehead_80	–0.350	1.703	0.269	–0.895	0.195	–1.300	39	0.101	0.201
Pair 4	Left Side of the Head (above ear)_65–Left Side of the Head (above ear)_80	0.550	1.934	0.306	–0.069	1.169	1.798	39	0.040	0.080
Pair 5	Right Side of the Head (above ear)_65–Right Side of the Head (above ear)_80	0.625	1.863	0.295	0.029	1.221	2.122	39	0.020	0.040
Pair 6	Left Side of the Head (further back)_65–Left Side of the Head (further back)_80	0.375	1.390	0.220	–0.070	0.820	1.706	39	0.048	0.096
Pair 7	Right Side of the Head (further back)_65–Right Side of the Head (further back)_80	0.400	1.446	0.229	–0.063	0.863	1.749	39	0.044	0.088

TABLE 6.5
Independent-Samples t-Test between 50 N mm and 65 N mm

Independent Samples Test											
		Levene's Test for Equality of Variances		t-Test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
How would you rate the overall comfort (includes the fit, visuals, and heat)?	Equal variances assumed	5.213	0.025	1.239	92	0.109	0.218	0.496	0.401	–0.299	1.292
	Equal variances not assumed			1.274	90.708	0.103	0.206	0.496	0.390	–0.278	1.270
Nose	Equal variances assumed	0.889	0.348	0.954	92	0.171	0.343	0.518	0.543	–0.560	1.595
	Equal variances not assumed			0.961	86.244	0.170	0.339	0.518	0.539	–0.553	1.589
Forehead	Equal variances assumed	1.169	0.282	–0.462	92	0.322	0.645	–0.214	0.463	–1.133	0.705
	Equal variances not assumed			–0.451	76.055	0.327	0.653	–0.214	0.474	–1.158	0.730
Left Side of the Head (above ear)	Equal variances assumed	7.342	0.008	2.001	92	0.024	0.048	0.590	0.295	0.004	1.175
	Equal variances not assumed			2.118	91.278	0.018	0.037	0.590	0.279	0.037	1.143
Right Side of the Head (above ear)	Equal variances assumed	10.232	0.002	2.186	92	0.016	0.031	0.713	0.326	0.065	1.361
	Equal variances not assumed			2.338	89.487	0.011	0.022	0.713	0.305	0.107	1.319
Left Side of the Head (further back)	Equal variances assumed	7.912	0.006	1.770	92	0.040	0.080	0.578	0.326	–0.070	1.226
	Equal variances not assumed			1.885	90.364	0.031	0.063	0.578	0.307	–0.031	1.187
Right Side of the Head (further back)	Equal variances assumed	6.872	0.010	1.780	92	0.039	0.078	0.559	0.314	–0.065	1.183
	Equal variances not assumed			1.883	91.303	0.031	0.063	0.559	0.297	–0.031	1.149

TABLE 6.6
Independent Samples t-Test between 50 N mm and 80 N mm

Independent Samples Test											
		Levene's Test for Equality of Variances		t-Test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
How would you rate the overall comfort (includes the fit, visuals, and heat)?	Equal variances assumed	2.173	0.144	1.211	92	0.115	0.229	0.496	0.410	–0.318	1.310
	Equal variances not assumed			1.233	88.939	0.110	0.221	0.496	0.403	–0.304	1.296
Nose	Equal variances assumed	2.086	0.152	1.428	92	0.078	0.157	0.743	0.520	–0.290	1.775
	Equal variances not assumed			1.463	90.131	0.074	0.147	0.743	0.508	–0.266	1.751
Forehead	Equal variances assumed	1.927	0.168	0.288	92	0.387	0.774	0.136	0.473	–0.803	1.075
	Equal variances not assumed			0.279	73.808	0.390	0.781	0.136	0.487	–0.835	1.107
Left Side of the Head (above ear)	Equal variances assumed	0.029	0.866	0.113	92	0.455	0.910	0.040	0.353	–0.662	0.741
	Equal variances not assumed			0.111	78.737	0.456	0.912	0.040	0.359	–0.675	0.754
Right Side of the Head (above ear)	Equal variances assumed	0.218	0.642	0.232	92	0.408	0.817	0.088	0.378	–0.664	0.840
	Equal variances not assumed			0.233	84.494	0.408	0.817	0.088	0.378	–0.664	0.840
Left Side of the Head (further back)	Equal variances assumed	1.175	0.281	0.574	92	0.284	0.567	0.203	0.353	–0.499	0.904
	Equal variances not assumed			0.589	90.349	0.279	0.557	0.203	0.344	–0.481	0.887
Right Side of the Head (further back)	Equal variances assumed	0.385	0.537	0.459	92	0.323	0.647	0.159	0.347	–0.529	0.848
	Equal variances not assumed			0.465	87.664	0.321	0.643	0.159	0.342	–0.521	0.840

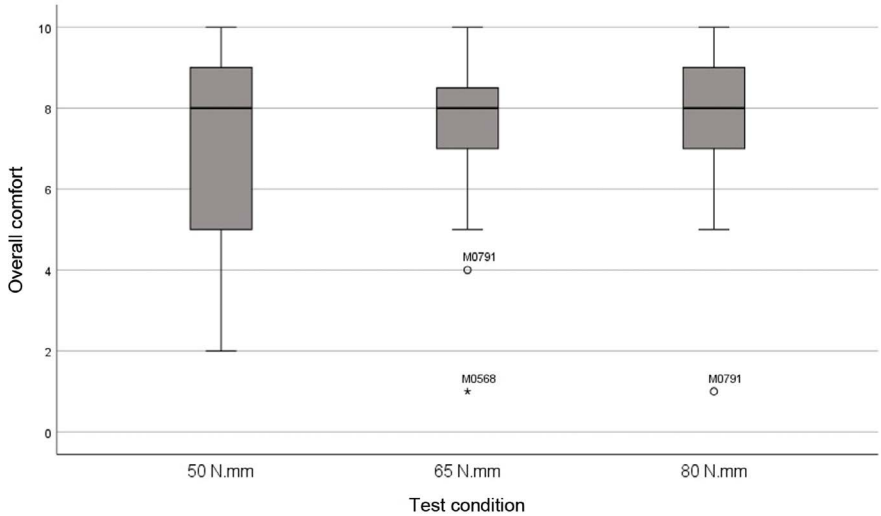


FIGURE 6.37 Overall comfort distribution per torque condition.

and therefore not practically significant. Looking at the differences in comfort between the 50 N mm and 80 N mm conditions, no statistically significant differences were observed for any of the comfort metrics. The box-and-whisker plots in Figures 6.37–6.40 illustrate the comparative distribution of the comfort results for the different conditions tested.

The box-and-whisker plots provide some further insights into the marginal differences in comfort between the conditions. For *overall* comfort and *nose* comfort, although no significant differences were observed and no noteworthy differences in median values, the higher temple band closing torque condition (80 N mm) had fewer persons with lower (<6) comfort scores compared to the lower force conditions. The nose comfort could be explained by a higher temple force resulting in better weight distribution around the temples, and therefore less weight on the nose.

For the rest of the comfort scores, a consistent pattern was not prevalent from 50 N mm to 65 N mm to 80 N mm.

To conclude on the comfort findings, no practically significant differences were observed between overall, nose, forehead, and side of head comfort between the difference torsion band conditions. The overall slippage and comfort findings showed no significant difference between the different torsion band forces. Since the largest torsion spring force, 80 N mm, ensured the best result of returning the torsion band to its default closing position it was therefore selected as the final torsion band force setting.

In summary, this case study demonstrated a typical approach used in a design loop trade study to inform design decisions and direction. This case study demonstrated an example for sampling method and study protocol, application of statistical

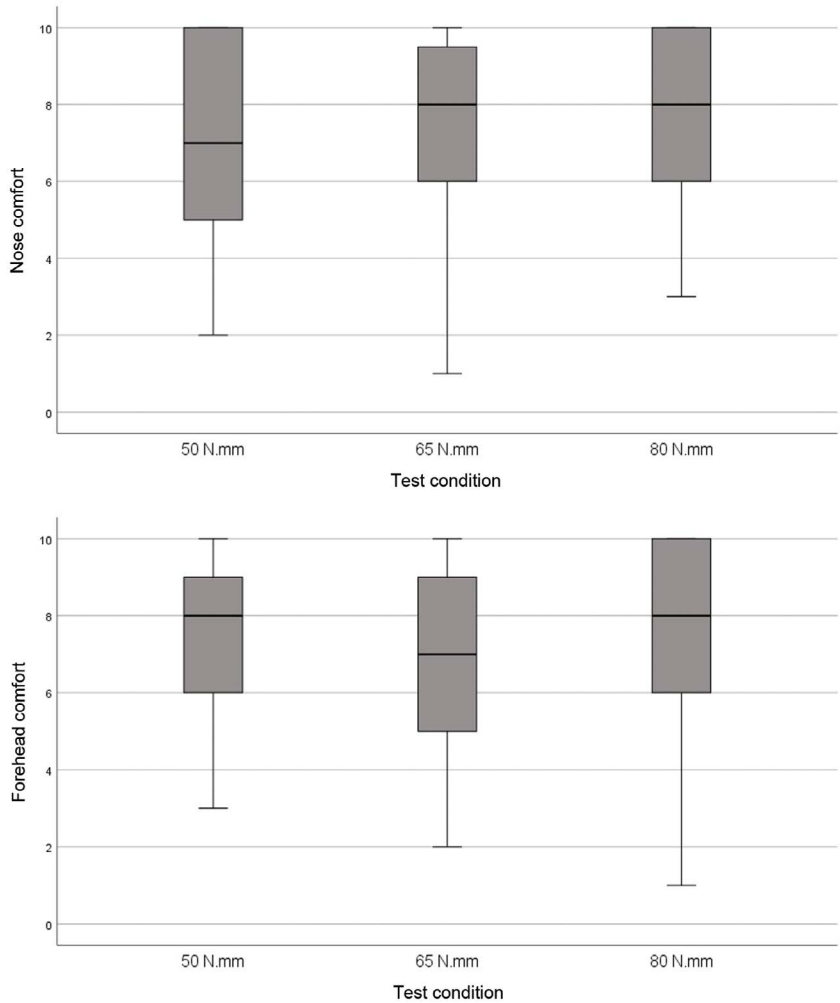


FIGURE 6.38 (a) Nose comfort distribution and (b) forehead comfort distribution per torque condition.

analysis tools typically used in trade studies, and discussion of results and how they inform the decision of design direction.

Case Study 5: Design Loop Evaluation Tools Highlighting Inputs to Design Changes

This case study goes into further detail on a design loop test, the tools used to investigate fit issues and the insights gained from those. Aspects demonstrated in this case study includes:

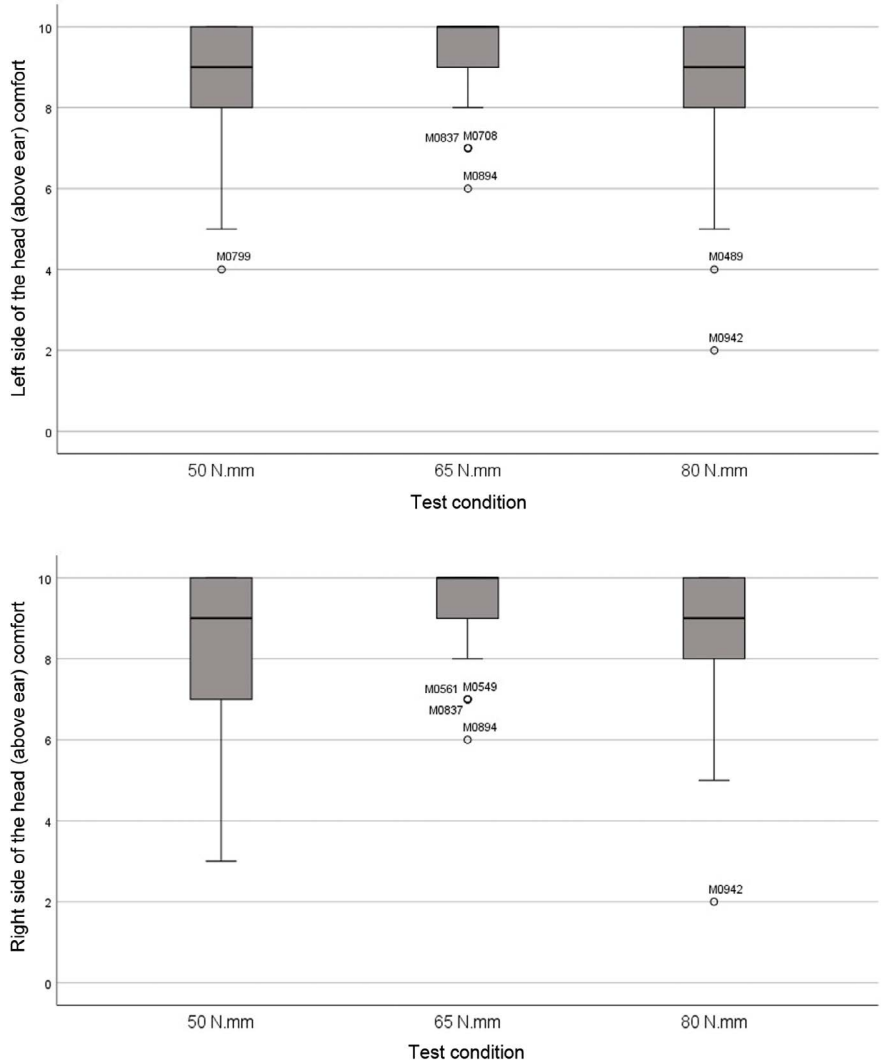


FIGURE 6.39 (a) Left side of head (above ear) comfort distribution and (b) right side of head (above ear) comfort distribution per torque condition.

- Statistical analysis methods used to investigate fit issues
- Scan analysis methods used to investigate fit issues
- Results and discussion of findings highlighted through statistical and scan analysis

During the previous case study (Case Study 4), the issue of slippage for this type of headset was identified and potential solutions were explored. This case study does not describe trade-offs between different potential solutions but rather demonstrates

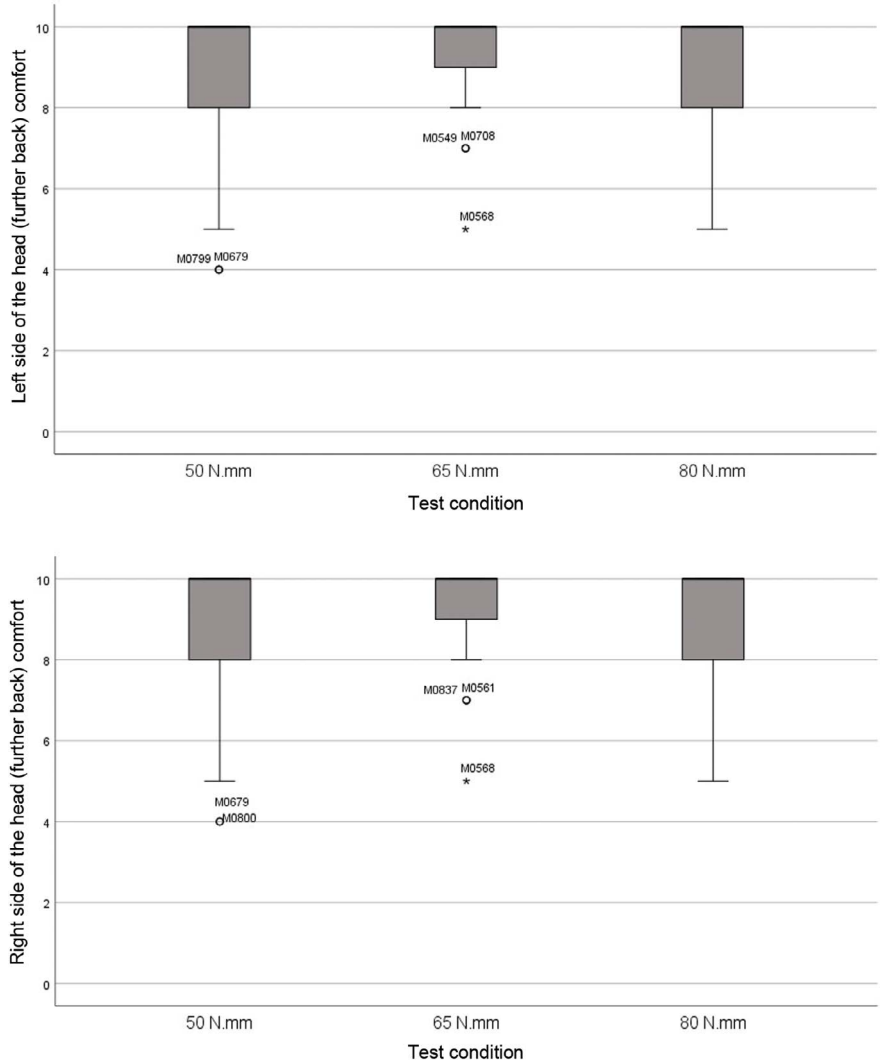


FIGURE 6.40 (a) Left side of head (behind ear) comfort distribution and (b) right side of head (behind ear) comfort distribution per torque condition.

tools that can be used to dive deeper into understanding the root cause of the problem. Since design loop tests have been discussed in other examples, this example will not describe the planning phase of the test (such as recruitment strategy), fit metrics (COF) or test tools. This case study will focus more on the tools employed for design and fit evaluation for the sole purpose of understanding aspects of poor fit. Two main sets of results analysis tools were used in this test: statistical tools and fit scan analysis tools. Within these sets, the following tests will be described in more detail in this example:

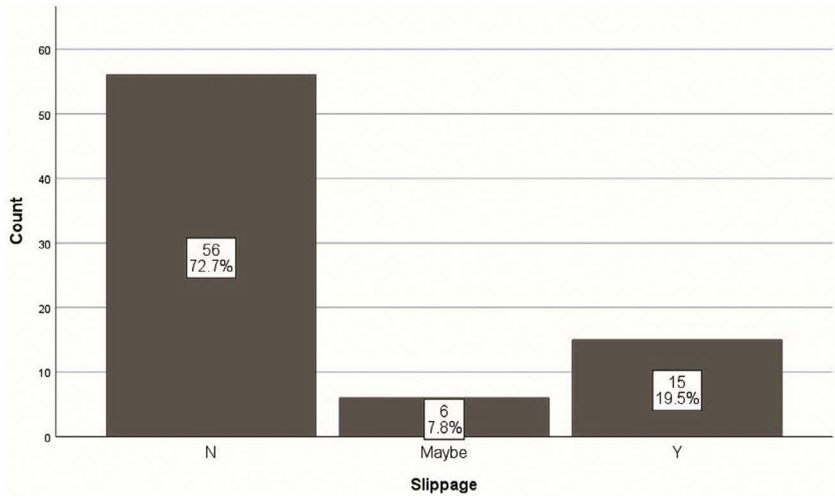


FIGURE 6.41 Slippage distribution for the headset.

Statistical Tools

1. Stepwise discriminant analysis
2. Plot failures on bivariate plots with anthropometry variables
3. Distribution of failures by race and gender

Fit Scan Analysis Tools

1. Sectional cuts
2. Area difference maps

Study Design and Sampling Method

During this case study, a new headset prototype was evaluated for fit. The device was evaluated for comfort after 30 minutes of wear, as well as for slippage and stability. The test was conducted on a randomly selected sample of 78 persons (42 females and 36 males) consisting of Asian, Caucasian, African American and Other race groupings.

Statistical Analysis Methods and Discussion of Results

As initial look into the study results slippage was classified as "Yes" (definite slippage observed), "Little" (slippage was very little and considered marginal) and "No" (no slippage identified). The frequency plot of slippage is indicated in Figure 6.41 and shows that the combined incidence of yes or little slippage was 27.3%.

To investigate the potential causes of slippage, *a stepwise discriminant analysis* was run including all the anthropometric variables available for the participants, relevant to the headset fit. This included:

1. Interpupilary distance
2. Head breadth

3. Face breadth
4. Max frontal span
5. Bi-Tragion breadth
6. Head length
7. Sellion width
8. Nose width max
9. Sellion-rhinion length
10. Face length
11. Head circumference
12. Hair type (Straight = 1, wavy = 2, curly = 3, kinky = 4 based on the rate of curliness of the hair)
13. Hair volume sides (Thin = 1, medium = 2 and thick = 3)
14. Hair volume back (Thin = 1, medium = 2 and thick = 3)

The "Little" slippage category was not deemed a very reliable classification with the potential that someone identified with "Little" slippage, could have been a slippage or non-slippage case. As a result, the discriminant analysis was run on different combinations of categories. Firstly, it was run with Yes, Maybe and No as dependent variables. Secondly, Maybe and Yes were combined into a "Slippage" and No into a non-slippage group and these two groups were selected as dependent variables. Thirdly, "Maybe" was combined with the "No" group as the non-slippage group versus "Yes" as the slippage group. And finally, the slippage group comprises all "Yes" and the non-slippage group comprises all "No". The first discriminant analysis ("Yes", "Maybe" and "No" dependent variables) indicated Head length as the discriminator (see Table 6.7).

The second discriminant analysis (Slippage ("Yes" and "Maybe") versus non-slippage ("N") as dependent variables) indicated Head length and Bi-Tragion width as discriminators (see Table 6.8). This discriminant function is a contrast, where breadth is contrasted with length. This means that slippage versus no slippage is observed in cases with either short wide heads versus long narrow heads. From experience with head Anthropometry, we know that these head shapes are indicators of Asian versus non-Asian participants.

The third and fourth discriminant analyses (Slippage ("Yes") versus non-slippage ("N" and "Maybe") and "Yes" versus "No" as dependent variables respectively) indicated hair type, nose width (max) and face length were significant discriminators

TABLE 6.7
Discriminant Analysis Results with "Yes", "Maybe", "No" as Dependent Variables

Wilks' Lambda					Standardized Canonical Discriminant Function Coefficients	
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.		Function 1
1	0.904	7.480	2	0.024	Head Length cm	1.000

TABLE 6.8
Discriminant Analysis Results with Slippage ("Yes", "Maybe") and Non-slippage ("No") as Dependent Variables

Wilks' Lambda					Standardized Canonical Discriminant Function Coefficients	
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.		Function 1
1	0.867	10.549	2	0.005	Bi-tragion Breadth cm	–0.693
					Head Length cm	1.105

(see Table 6.9 for "Yes" versus "No" and "Maybe"). This is also a contrast with Nose Width contrasted against the other two. In other words, as Nose Width gets larger the other two get smaller and vice versa.

Based on the findings of the discriminant function (also observed during previous studies on similar products) slippage could be identified for predominantly Asian participants. The distribution of the discriminant function variables was investigated by looking at *bivariate and distribution plots.* Asian participants were furthermore included in the plots to observe further trends. Head length and Bi-Tragion breadth, with Asian versus non-Asian point label, are illustrated in Figure 6.42. This plot illustrates that, for both males and females, a relatively large percentage of slippage ("Yes" and "Maybe") cases were observed for persons with wide, short heads. A large number of these were Asian. In addition, apart from two participants, all other slippage cases (including "Yes" and "Maybe") had head lengths shorter than 19.5 mm.

The distribution of face length versus nose width, with hair type as point label, is shown in Figure 6.43. Slippage was observed for persons with a face length shorter 12.5 cm and nose widths wider than 3.2 cm. The majority of persons who had slippage had straight hair (hair type = 1).

The distribution of slippage is illustrated for hair type in Figure 6.44. This figure indicates a clear higher incidence of slippage for people with straight hair. Of the people who had slippage, 62.5% had straight hair.

TABLE 6.9
Discriminant Analysis Results with Slippage ("Yes") and Non-slippage ("No", "Maybe") as Dependent Variables

Wilks' Lambda					Standardized Canonical Discriminant Function Coefficients	
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.		Function 1
					NoseWidthMaxcm	–0.984
					FaceLengthcm	0.615
1	0.821	14.478	3	0.002	Hairtype#	1.142

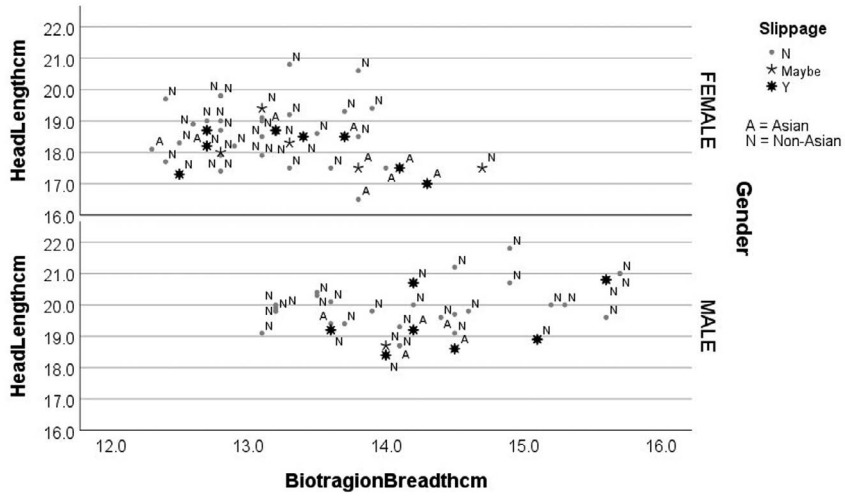


FIGURE 6.42 Head length versus bi-tragion breadth.

The statistical analysis highlighted key factors that contributed to slippage. These included Head Length versus Bi-Tragion (head) width relationship. With shorter heads relative to their width, they have more incidence of slippage. Another relevant variable identified was hair type, with a higher incidence of slippage observed for people with straight hair. In addition, other factors were also observed to influence the prevalence of slippage included face anthropometry variables. This included persons with shorter face lengths and wider nose widths. These two anthropometric

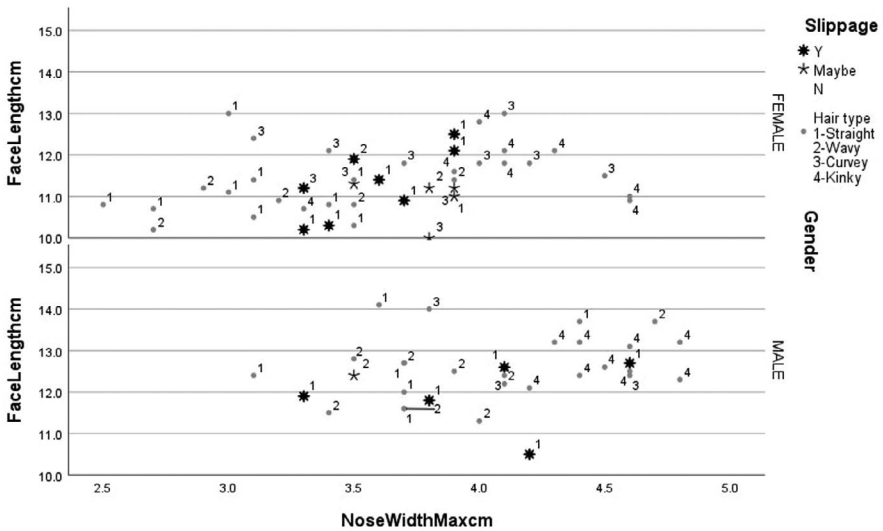


FIGURE 6.43 Face length versus nose width (max).

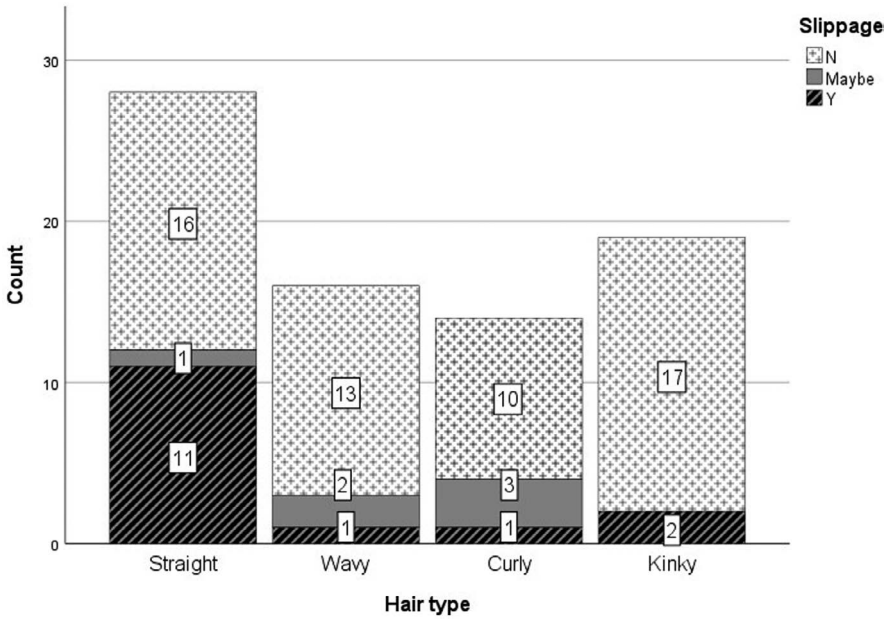


FIGURE 6.44 Distribution of hair type broken down by occurrence of slippage.

variables are not directly related to the fit of the headset but could be investigated further to identify demographic groups who could see more incidences of slippage.

Scan Analysis Methods and Discussion of Results

During the study, a 3D scan was taken of the participant wearing the headset, termed a fit scan. During the data analysis, the fit scan was aligned to the base anthropometry scan using the closest point alignment of a selection of 10–20 points on the face. Thereafter, the headset CAD file was aligned to the fit scan using the same method. Finally, the fit scan was removed leaving the headset CAD aligned to the base anthropometry scan (Figure 6.25 demonstrates the process). With this alignment, several different scan analyses were performed. All scan analysis was performed using Polyworks Inspector™.

First of all, a *cross-section* was taken through the midway of the headset (see Figure 6.45). This cross-section provided insight into how each individual's head curve shape compared to the headset design fit.

The cross-sections were compared for persons who experienced slippage versus persons who did not experience slippage. Looking at the cross-sections for persons who experienced slippage (see Figure 6.46), the first characteristic that was visually observed was that the heads were noticeably wider in relation to length compared to the non-slippage persons (see Figure 6.47). On further investigation, it was observed that the flexible temple band did not closely follow the curve of the head, especially in the area where the flexible band departed from the rigid temple arms. This observation led to the hypothesis that the rigid temple arms were too long in relation to

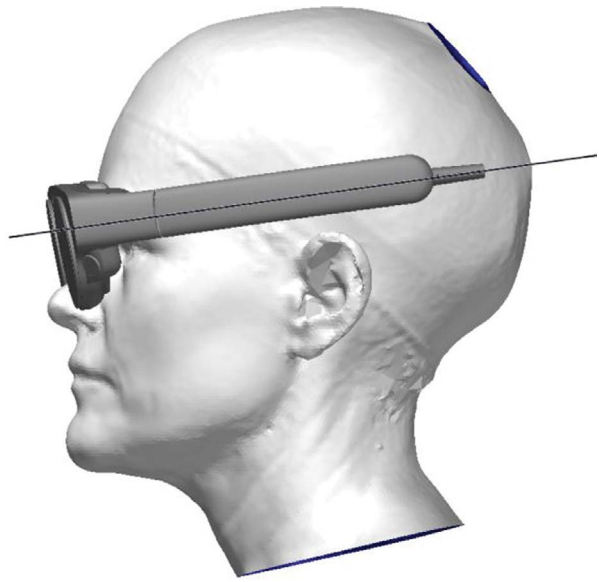


FIGURE 6.45 Cross-section of head midway through headset temple arms.

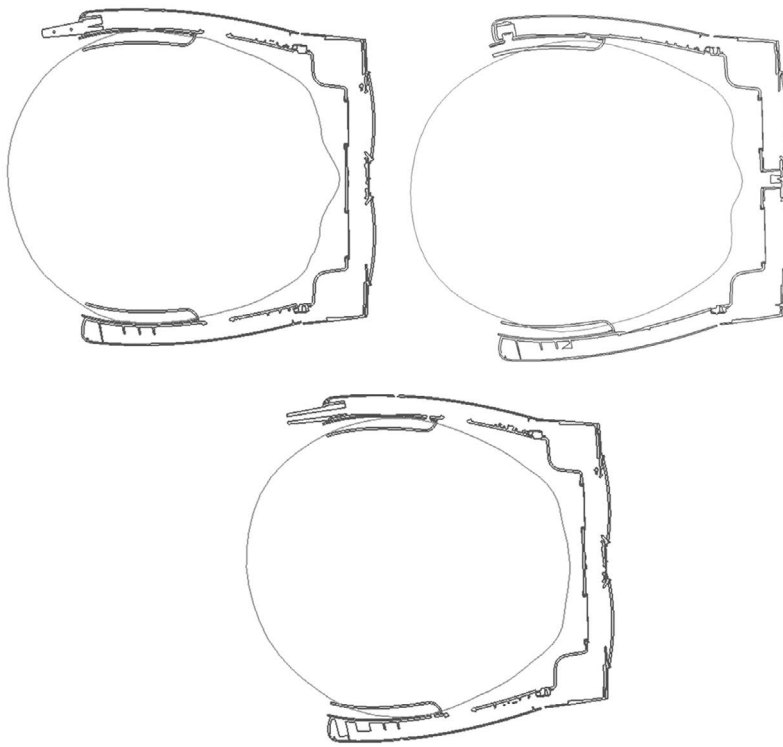


FIGURE 6.46 Headset versus head cross-section for persons with slippage.

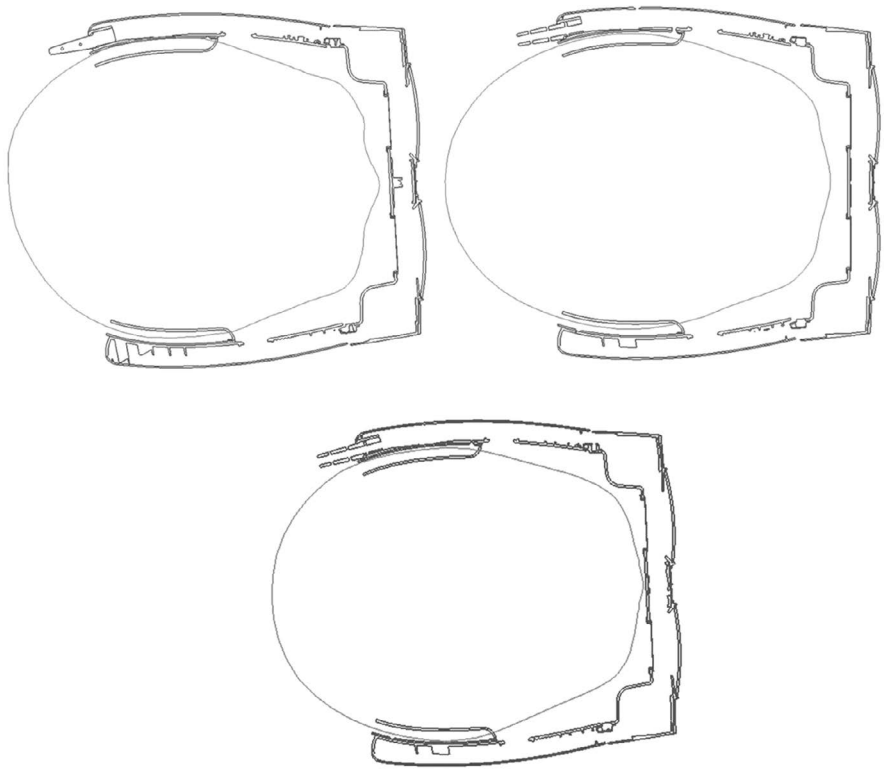


FIGURE 6.47 Headset versus head cross-section for persons with no slippage.

shorter heads, and when the band had to make an acute bend due to the large head breadth versus head length for these participants, it would not allow for good band contact needed to keep the device secure on the head. The design options discussed to overcome this fit issue included (1) shortening the rigid temple arms or (2) allowing the flexible band to break away from the rigid temple arms sooner.

Secondly, *a surface difference map* (termed a "Color map" in Polyworks Inventor™) was obtained between the base head scan and the headset components in contact with the user, which included the nosepiece, forehead pad and side temple pads. Since the CAD components are rigid bodies and cannot conform (compress or bend) in the way that the real objects can, the color map will indicate larger distances between surfaces in areas where the real-life deformable wearable objects would compress or deform the most. From the user's perspective, the areas where these wearable components contact the face, would include areas with limited soft tissue such as on the maxilla bone in the nose region, forehead and brow ridge areas, and sides of the head above the ears. The thin, soft tissue typically observed in these areas would be approximately uniformly distributed. The wearable compression or deformation would normally be paired with higher levels of pressure. Therefore, the color maps served as an indication of areas where higher pressure could be expected to occur due to compression of the soft headset material.

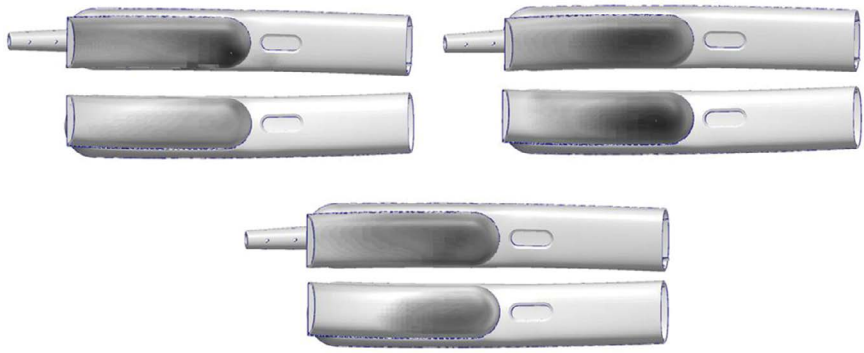


FIGURE 6.48 Difference map for persons with slippage.

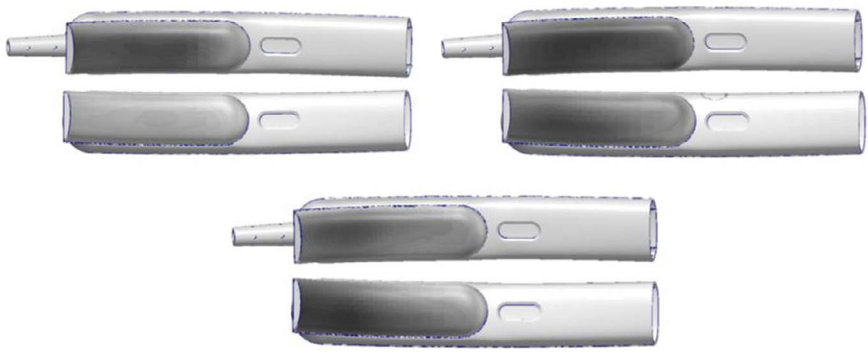


FIGURE 6.49 Difference map for persons with no slippage.

The different color maps for the persons with slippage (such as that seen in Figure 6.48) were compared to persons with no slippage (such as that seen in Figure 6.49). Less apparent differences were observed between the color maps for these two groups of persons. For several of the "slippage" persons, the color maps pointed towards more pressure on the more anterior part of the side temple pads, whereas for the "non-slippage" persons, the pressure seemed to be more uniformly distributed on the pad. Higher surface differences (expected to equate to pressure) is indicated as darker in shade on Figures 6.48 and 6.49. The trend of more pressure to the front of the temple pads was also observed to be more prevalent for persons who experienced discomfort on the sides of the head. This finding furthermore pointed out that the side temple arms needed to follow the curve of the head better, and temple padding needed to provide better, more equal pressure distribution, especially for persons with wider (and typically relatively shorter) heads.

In summary, the consolidated findings of this investigation guided the design team to understand that the current design does not allow for adequate accommodation of the different head shapes typically observed between Asian and non-Asian

users. The scan analysis highlighted that the band did not follow the more acute head curve posterior to the ears toward the back of the head, as typically observed in the Asian population. For users where slippage was observed, this investigation highlighted that more compression towards the front of the side temple pads was observed instead of uniform contact or an increase in contact towards the back of the side pads (such as for no-slippage cases). This observation supports the finding from the cross-sections that the temple arm shape of the wearable does not follow the (fit) the curve of the heads.

Case Study 6: Importance of Fit Testing to Predict Sizing Numbers for Purchasing

This case study presents a summary example of a typical sizing loop tests. Although the objective of the sizing loop test was to inform purchasing decisions, poor sizing quality and poor fit resulted in inputs for product design optimization, before the purchase decisions would be useful. Aspects demonstrated in this case study includes:

- Poor sizing quality regularly observe in existing product lines
- COF metrics for hearing protection product (summary)
- Conclusions of findings and design recommendations

This study and its results were first reported in a technical report by Robinette (Robinette, 2007). It was one of a series of studies by the Air Force Research Laboratory (AFRL) and Naval Air Systems Command (NAVAIR) with the goal to provide military aviation crews with effective, affordable, reliable, easy-to-use hearing protection that allowed safe, extended exposures in up to 150 dB of aviation related noise. This case study is a summary.

In this study the effectiveness of a passive hearing protection earmuff and cloth helmet assembly was measured using a sound attenuation measurement system called the Microphone-in-real-ear (MIRE) system. The earmuff had earcups to reduce the amount of sound coming into the ear and it was worn with a cloth helmet, referred to as a cranial, that enabled impact protective panels, called shields, to be attached. An illustration of the entire ensemble is shown in Figure 6.50.

The earmuff came in only one adjustable size. The cloth cranial helps to hold the earcups in place, and it came in four sizes: 6.75, 7, 7.25, and 7.5. The goal of this fit test was to determine the range of accommodation within one cloth cranial size. Knowing the range of fit in one size would allow the range to be mapped against the full TP sample to determine what cranial sizes were needed and if any could be dropped.

Poor Sizing Quality

It was assumed that since this product was in use it was accommodating the TP. This was intended to be a sizing loop study to determine purchasing numbers per size. However, the test revealed the design modifications were necessary to accommodate the TP.

Prior to testing the soft cranials and their patterns were measured. This led to the unexpected discovery that three of the four sizes (6.75, 7, and 7.25), had little or no actual difference between them. Inspection of the specifications revealed that the

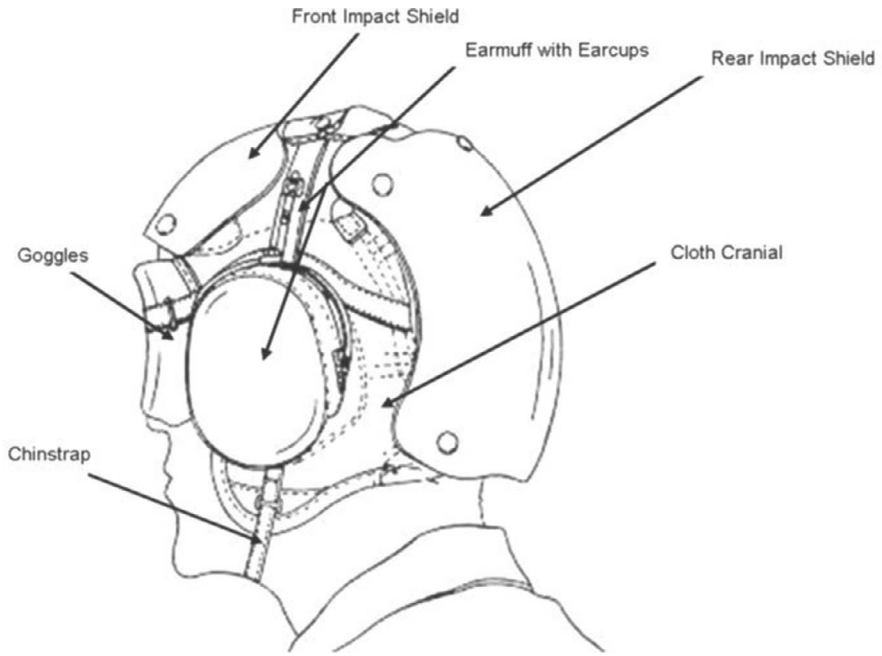


FIGURE 6.50 Earmuff with complete cranial assembly.

impact shields that attach to the cranials for all sizes are identical (interchangeable). The same is true of the cloth side panels that hold the earcups in the cranial. Due to the rigid material in the impact shields and the ear cups, the soft cranial sizes were adjusted to accommodate them. This made them identical where the impact shields and ear cups attached, which included all the important sizing areas. The size 7.5 was made larger. There were small differences (5 mm or less) in the cloth cranial in areas that didn't really impact fit including the length along the center seam and the widths at the front edge, the sleeve snap, the back center, and the back bottom. These measurements of the cloth cranial are shown in Table 6.10.

COF Metrics for Hearing Protective Device (Summary)

The COF for the prototype fit test was that the assembly should provide a level of sound attenuation while simultaneously being comfortable with little or no slippage

TABLE 6.10
Cloth Cranial Measurements

Size	Front Edge (cm)	Sleeve Snap (cm)	Back Center (cm)	Back Bottom (cm)	Length (cm)
6.75	15.5	17.5	16	12.5	38.5
7	15	18	16.5	12.5	39
7.25	15.5	18	16.5	13	39
7.5	16.5	19	19	13.5	39

during normal work activities. These conditions interact so simultaneous fit for all criteria was important. For example, the most comfortable cranial may be the loosest, but the best sound attenuation is achieved with a tight fit. A questionnaire was used to assess comfort and slippage, with the subjects asked to perform a set of movement tasks.

Study Method

Prior trade studies had indicated that the size 7.5 provided worse sound attenuation than the other sizes. It was assumed that this was due to a looser fit over the ears. Therefore, the size selected for testing was size 6.75. Any of the three smallest sizes would be essentially equivalent so the smallest of the three should provide as good or better performance than the other two.

The test sample consisted of 30 men and 30 women, each tested in the size 6.75. The fit test process was as follows:

- Each subject was measured manually for a series of 1D head measurements and a 3D scan of their head was done.
- Assembly was donned, and the earmuff and chin strap were adjusted to achieve the best comfort and stability possible.
- Performed a series of movements and tasks and filled out the comfort and stability questionnaire.
- Moved to a special MIRE facility where the sound attenuation was evaluated.

Conclusions and Recommendations

MIRE attenuation testing results indicate that males consistently achieve a higher attenuation score than females by 5–10 dB and the addition of the cranial did not add significant hearing protection for them. The different response for females was determined to be due to the earmuff adjustability and band length. The band was not adjustable for most females, who needed a smaller band but were unable to adjust it to the needed length. The men, on the other hand, were able to adjust the band for their head size.

While lower MIRE attenuation scores and minimal adjustment were the most common among women, those men with Bitragion-coronal arcs less than 340 mm also had the issue. It was determined that allowing the band to reduce in size by 20 mm would provide enough adjustment to accommodate all the females and the few males who required a smaller band length than the current configuration allows.

The band length caused all critical fit failures. It was concluded that with the added adjustment the male and female TP could be accommodated in one size without slippage, without significant discomfort, and with maximal sound attenuation.

However, there was another important fit issue identified that was not part of the COF for the fit mapping. The proper placement of the front impact shield. For many of the subjects, the proper earcup positioning and the proper impact shield location on the forehead were not both possible simultaneously. This meant that the front impact shield was not in the right position, (over the forehead) to provide impact protection when the ear cups were in the right place for hearing protection. The design requires the earcups to rotate off the ear as the impact shield is rotated forward and down over the forehead.

Analysis of the 3D scans with and without the cranial in place indicated a 33 mm forward rotation would be necessary to protect the forehead. However, rotation of the head band by this amount would force the earcups off the ears wrecking the sound attenuation. The proximity of the sleeve snap to the front impact shield only allows for 1 mm forward rotation before the head band contacts the impact shield. Therefore, it was concluded that the front impact shield would need to be adjusted forward 32 mm to accommodate forehead protection while maintaining hearing protection, or additional sizes would be needed that had a lower position for the front impact shield.

In summary, the fit test intended to be a sizing loop fit test, but instead highlighted design and sizing failures in the product. In order to fix the problem, the team recommended that shortening the headband, to a headband that allows for a 20 mm smaller length would accommodate the TP in one size provided the front impact shield location is not required to be in the correct place for all users. If the placement of the front impact shield was important, re-design would be needed to either allow shield rotation or have additional sizes allowing its proper placement. If a fit test was not performed in order to provide accurate size prediction numbers, the sizing chart might have been used resulting in purchasing a range of sizes that did not adequately fit and protect any of the intended users.

REFERENCES

Abeysekera, J., Holmér, I., & Dupuis, C. (1991). Heat transfer characteristics of industrial safety helmets. *Undefined*. https://www.semanticscholar.org/paper/Heat-transfer-characteristics-of-industrial-safety-Abeysekera-Holm%C3%A9r/3aee278dd30f424cdc0b27ffc78a193660617daf

Air Standardization Coordinating Committee, (ASSC) (1991, September) " A Basis for Common Practices and Goals in the Conduct of Anthropometric Surveys", *Air Standard 61/83*.

Alam, F., Chowdhury, H., Elmir, Z., Sayogo, A., Love, J., & Subic, A. (2010). An Experimental Study of Thermal Comfort and Aerodynamic Efficiency of Recreational and Racing Bicycle Helmets. *Procedia Engineering*, *2*(2), 2413–2418. https://doi.org/10.1016/j.proeng.2010.04.008

Badawi-Fayad, J., & Cabanis, E.-A. (2007). Three-Dimensional Procrustes Analysis of Modern Human Craniofacial Form. *Anatomical Record (Hoboken, N.J.: 2007)*, *290*(3), 268–276. https://doi.org/10.1002/ar.20442

Bain, A. R., Deren, T. M., & Jay, O. (2011). Describing Individual Variation in Local Sweating During Exercise in a Temperate Environment. *European Journal of Applied Physiology*, *111*(8), 1599–1607. https://doi.org/10.1007/s00421-010-1788-9

Ball, R. (2011). *SizeChina: A 3D Anthropometric Survey of the Chinese Head*. Technische Universiteit Delft.

Bandmann, C., Akrami, M., & Javadi, A. (2018). An Investigation into the Thermal Comfort of a Conceptual Helmet Model Using Finite Element Analysis and 3D Computational Fluid Dynamics. *International Journal of Industrial Ergonomics*, *68*, 125–136. https://doi.org/10.1016/j.ergon.2018.07.004

Bass, W. M. (1971). *Human Osteology: A Laboratory and Field Manual of the Human Skeleton*. Missouri Archaeological Society.

Beaumont, C. A. A., Knoops, P. G. M., Borghi, A., Jeelani, N. U. O., Koudstaal, M. J., Schievano, S., Dunaway, D. J., & Rodriguez-Florez, N. (2017). Three-Dimensional Surface Scanners Compared With Standard Anthropometric Measurements for Head Shape. *Journal of Cranio-Maxillofacial Surgery*, *45*(6), 921–927. https://doi.org/10.1016/j.jcms.2017.03.003

Blackwell, S., Robinette, K. M., Boehmer, M., Fleming, S., Kelly, S., Brill, T., Hoeferlin, D., & Burnsides, D. (2002). *Civilian American and European Surface Anthropometry Resource (CAESAR). Volume 2: Descriptions*. SYTRONICS INC DAYTON OH. https://apps.dtic.mil/docs/citations/ADA408374

Bogerd, C. P., Aerts, J.-M., Annaheim, S., Bröde, P., de Bruyne, G., Flouris, A. D., Kuklane, K., Sotto Mayor, T., & Rossi, R. M. (2015). A Review on Ergonomics of Headgear: Thermal Effects. *International Journal of Industrial Ergonomics*, *45*, 1–12. https://doi.org/10.1016/j.ergon.2014.10.004

Bogerd, C. P., Brühwiler, P., & Heus, R. (2008). The Effect of Rowing Headgear on Forced Convective Heat Loss and Radiant Heat Gain on a Thermal Manikin Headform. *Journal of Sports Sciences*, *26*, 733–741. https://doi.org/10.1080/02640410701787783

Bredenkamp, K., Zwane, P. E., & De Ridder, H. (2006). Accuracy investigation of the tc2 scanner, the first 3D whole-body scanner employed in South Africa. *Proceedings of the 9th Conference of the Ergonomics Society of South Africa*. Ergonomics Society of South Africa, Pretoria, South Africa. https://drive.google.com/drive/u/1/my-drive

Brühwiler, P., Ducas, C., Huber, R., & Bishop, P. (2004). Bicycle Helmet Ventilation and Comfort Angle Dependence. *European Journal of Applied Physiology*, *92*, 698–701. https://doi.org/10.1007/s00421-004-1114-5

Brühwiler, P. A., Buyan, M., Huber, R., Bogerd, C. P., Sznitman, J., Graf, S. F., & Rösgen, T. (2006). Heat Transfer Variations of Bicycle Helmets. *Journal of Sports Sciences*, *24*(9), 999–1011. https://doi.org/10.1080/02640410500457877

Bruyne, G., Aerts, J.-M., Vander Sloten, J., Goffin, J., Verpoest, I., & Berckmans, D. (2012). Quantification of Local Ventilation Efficiency Under Bicycle Helmets. *International Journal of Industrial Ergonomics*, *42*, 278–286. https://doi.org/10.1016/j.ergon.2012.02.003

Burnsides, D. B., Files, P. M., & Whitestone, J. J. (1996). *Integrate 1.25: A Prototype for Evaluating Three-Dimensional Visualization, Analysis, and Manipulation Functionality* (AL/CF-TR-1996-0095). Crew Systems Directorate, Human Engineering Division, Armstrong Laboratory. https://apps.dtic.mil/sti/citations/ADA330986

Cabanac, M., & Brinnel, H. (2000). Beard vs. Forehead, Ten Years Later. *American Journal of Human Biology*, *12*(4), 460–464. https://doi.org/10.1002/1520-6300(200007/08)12:4<460::AID-AJHB5>3.0.CO;2-H

Case, H., Ervin, C., & Robinette, K. M. (1989). *Anthropometry of a Fit Test Sample Used in Evaluating the Current and Improved MCU-2/P Masks* (Technical Report AAMRL-TR-89-009). Armstrong Aerospace Medical Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/a215173.pdf

Dai, H., Pears, N., Smith, W., & Duncan, C. (2020). Statistical Modeling of Craniofacial Shape and Texture. *International Journal of Computer Vision*, *128*(2), 547–571. https://doi.org/10.1007/s11263-019-01260-7

Du, L., Zhuang, Z., Guan, H., Xing, J., Tang, X., Wang, L., Wang, Z., Wang, H., Liu, Y., Su, W., Benson, S., Gallagher, S., Viscusi, D., & Chen, W. (2006). *Head-and-Face Anthropometric Survey of Chinese Workers*.

Ellis, A. J. (2003). *Development of Fundamental Theory & Techniques for the Design and Optimisation of Bicycle Helmet Ventilation*. RMIT University.

Gordon, C. C., Blackwell, C. L., Bradtmiller, B., Parham, J. L., Barrientos, P., Paquette, S. P., Corner, B. D., Carson, J. M., Venezia, J. C., Rockwell, B. M., Mucher, M., & Kristensen, S.. (2012). *2012 ANTHROPOMETRIC SURVEY OF U.S. ARMY PERSONNEL: METHODS AND SUMMARY STATISTICS*.

Gordon, C. C., Churchill, T., Clauser, C. E., Bradtmiller, B., & McConville, J. T. (1989). *1988 Anthropometric survey of US army personnel: Methods and summary statistics* (Technical Report Natick/TR-89/044; Issue Technical Report Natick/TR-89/044). Anthropology Research Project Inc Yellow Springs OH.

Gray, H., & Goss, C. M. (1973). *Anatomy of the Human Body* (29th ed.).

Gupta, S., Castleman, K., Markey, M., & Bovik, A. (2010). *Texas 3D Face Recognition Database* (p. 100). https://doi.org/10.1109/SSIAI.2010.5483908

Hoffmeister, J. W., McPohlenz, M., Addleman, D. A., Kasic, M. A., & Robinette, K. M. (1996). *Functional Description of the Cyberware Color 3-D Digitizer 4020 RGB/PS-D.* (AL/CF-TR-1996-0069). Human Engineering Division Crew Systems Diroctorate Wright-Patterson AFB. https://apps.dtic.mil/sti/pdfs/ADA478783.pdf

Hsu, Y.-L., Tai, C.-Y., & Chen, T.-C. (2000). Improving Thermal Properties of Industrial Safety Helmets. *International Journal of Industrial Ergonomics*, *26*(1), 109–117. https://doi.org/10.1016/S0169-8141(99)00058-X

Hudson, J. A., & Robinette, K. (2003). *CAESAR: Summary Statistics for the Adult Population (Ages 18-65) of Italy* (Technical Report AFRL-HE-WP-TR-2004-0165). Air Force Research Laboratory, Biomechanics Branch.

International Society for Advancement of Kinanthropometry (ISAK) (2001). International standards for anthropometric assessment. International Society for the Advancement of Kinanthropometry, Potchefstroom, South Africa.

ISO 20685-1:2018. *3-D scanning methodologies for internationally compatible anthropometric databases—Part 1: Evaluation protocol for body dimensions extracted from 3-D body scans*, (2018) (testimony of ISO 20685-1). http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/32/63260.html

Kim, J. Y., You, J. W., & Kim, M. S. (2017). South Korean Anthropometric Data and Survey Methodology: 'Size Korea' Project. *Ergonomics*, *60(11)*, 1586–1596.

Kouchi, M., & Mochimaru, M. (2004). Analysis of 3D Face Forms for Proper Sizing and CAD of Spectacle Frames. *Ergonomics*, *47*(14), 1499–1516. https://doi.org/10.1080/00140130412331290907

Kouchi, M., & Mochimaru, M. (2011). Errors in Landmarking and the Evaluation of the Accuracy of Traditional and 3D Anthropometry. *Applied Ergonomics*, *42*(3), Article 3. https://doi.org/10.1016/j.apergo.2010.09.011

Lee, W., Goto, L., Molenbroek, J., & Goossens, R. (2017). Anlysis Methods of the Variation of Facial Size and Shape Based on 3d Face Scan Images. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, p. 1413). https://doi.org/10.1177/1541931213601836

Lee, W., Yang, X., Jung, H., You, H., Goto, L., Molenbroek, J., & Goossens, R. (2016). Application of Massive 3D Head and Facial Scan Datasets in Ergonomic Head-Product Design. *International Journal of the Digital Human*, *1*, 344. https://doi.org/10.1504/IJDH.2016.10005368

Li, G.-L., Li, L.-P., & Cai, Q.-E. (2008). Motorcycle Helmet Use in Southern China: An Observational Study. *Traffic Injury Prevention*, *9*(2), 125–128. https://doi.org/10.1080/15389580801895152

Li, P., Tashjian, A., & Hurley, M. (2022). Digitizing human scalp shape through 3D scanning. *Proceedings of the 7th International Digital Human Modeling Symposium*, *7*(1), Article 1. https://doi.org/10.17077/dhm.31760

Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Transactions on Graphics*, *36*(6), 1–17. https://doi.org/10.1145/3130800.3130813

Liu, X., & Holmér, I. (1995a). Evaporative Heat Transfer Characteristics of Industrial Safety Helmets. *Applied Ergonomics*, *26*(2), 135–140. https://doi.org/10.1016/0003-6870(95)00010-a

Liu, X., & Holmér, I. (1995b). Evaporative Heat Transfer Characteristics of Industrial Safety Helmets. *Applied Ergonomics*, *26*(2), 135–140. https://doi.org/10.1016/0003-6870(95)00010-A

Luximon, Y., Ball, R., & Chow, E. (2015). A Design and Evaluation Tool Using 3D Head Templates. *Computer-Aided Design and Applications*, *13*, 1–9. https://doi.org/10.1080/16864360.2015.1084188

Machado-Moreira, C. A., Wilmink, F., Meijer, A., Mekjavic, I. B., & Taylor, N. A. S. (2008). Local Differences in Sweat Secretion from the Head During Rest and Exercise in the Heat. *European Journal of Applied Physiology*, *104*(2), 257–264. https://doi.org/10.1007/s00421-007-0645-y

Mehrabyan, A., Guest, S., Essick, G., & McGlone, F. (2011). Tactile and Thermal Detection Thresholds of the Scalp Skin. *Somatosensory & Motor Research*, *28*(3–4), 31–47. https://doi.org/10.3109/08990220.2011.602764

Melzer, J. E., & Moffitt, K. (1996). *Head-Mounted Displays: Designing for the User* (1st edition). McGraw-Hill Professional.

Meyer, B. J., Meij, M. S., van Papendorp, D. H., & Viljoen, M. (2002). *Human Physiology* (3rd ed.). Juta. https://www.goodreads.com/work/best_book/41292373-human-physiology

Mukunthan, S., Vleugels, J., Huysmans, T., Kuklane, K., Mayor, T. S., & de Bruyne, G. (2019a). Thermal-Performance Evaluation of Bicycle Helmets for Convective and Evaporative Heat Loss at Low and Moderate Cycling Speeds.pdf. *Applied Science*, *9*, 3672.

Mukunthan, S., Vleugels, J., Huysmans, T., Sotto Mayor, T., & Bruyne, G. (2019b). *A 3D Printed Thermal Manikin Head for Evaluating Helmets for Convective and Radiative Heat Loss: Volume VII: Ergonomics in Design, Design for All, Activity Theories for Work Analysis and Design, Affective Design* (pp. 592–602). https://doi.org/10.1007/978-3-319-96071-5_63

Niezgoda, G., & Zhuang, Z. (2015). Development of Headforms for ISO Eye and Face Protection Standards. *Procedia Manufacturing*, *3*, 5761–5768. https://doi.org/10.1016/j.promfg.2015.07.822

Osczevski, R. (1996). *Design and Evaluation of a Three-Zone Thermal Manikin Head*. 20.

Pang, T., Subic, A., & Takla, M. (2013). A Comparative Experimental Study of the Thermal Properties of Cricket Helmets. *International Journal of Industrial Ergonomics*, *43*, 161–169. https://doi.org/10.1016/j.ergon.2012.12.003

Park, B.-K., Reed, M., & Corner, B. (2020). A Three-Dimensional Parametric Adult Head Model With Representation of Scalp Shape Variability Under Hair. *Applied Ergonomics*, *90*. https://doi.org/10.1016/j.apergo.2020.103239

Patel, R., & Mohan, D. (1993). An Improved Motorcycle Helmet Design for Tropical Climates. pdf. *Applied Ergonomics*, *24*(6), 427–431.

Perret-Ellena, T., Skals, S. L., Subic, A., Mustafa, H., & Pang, T. Y. (2015). 3D Anthropometric Investigation of Head and Face Characteristics of Australian Cyclists. *Procedia Engineering*, *112*, 98–103. https://doi.org/10.1016/j.proeng.2015.07.182

Ranke, J. (1884). Verstandigung uber Ein Gemainsames Cranio-Metrisches Verfahren (Frankfurter Verstandigung) {Standardization of a Common Head Measurement Method}. *Archiver Anthropologie*, *15*, 1–8.

Reid, J., & Wang, E. L. (2000). A System for Quantifying the Cooling Effectiveness of Bicycle Helmets. *Journal of Biomechanical Engineering*, *122*(4), 457–460. https://doi.org/10.1115/1.1287163

Reischl, U. (1986). Fire Fighter Helmet Ventilation Analysis. *American Industrial Hygiene Association Journal*, *47*, 546–551. https://doi.org/10.1080/15298668691390205

Robinette, K. M. (1986). Three-dimensional Anthropometry-shaping the future. Human Factors Society Inc., Santa Monica, CA. *Proceedings of the Human Factors Society-30th Annual Meeting*, *Vol. 1*., 205.

Robinette, K. M. (2007). *Maximizing Anthropometric Accommodation and Protection* [Technical Report]. Air Force Research Laboratory, Biomechanics Branch. https://apps.dtic.mil/sti/pdfs/ADA478783.pdf

Robinette, K. M., Blackwell, S., Daanen, H., Boehmer, M., & Fleming, S. (2002). *Civilian American and European Surface Anthropometry Resource (Caesar), Final Report. Volume 1. Summary*. SYTRONICS INC DAYTON OH.

RSA-MIL-STD 127 Vol 1. (2005). *Ergonomics Design: Anthropometry and Environment* (RSA-MIL-STD 127 Vol 1). ARMSCOR.

Schnieders, T., Bredenkamp, K., & Daneshyan, M. (2023). Analysis of Variance in Neutral Gaze Head Orientation During 3D Head Anthropometry Data Collection. *Proceedings of 3DBODY.TECH 2023*. 3D Body TECH, Lugano, Switzerland.

Servadei, F. (2003). Effect of Italy's Motorcycle Helmet Law on Traumatic Brain Injuries. *Injury Prevention*, *9*(3), 257–260. https://doi.org/10.1136/ip.9.3.257

Simmons, K. P., & Istook, C. L. (2003). Body Measurement Techniques: Comparing 3D Body-Scanning and Anthropometric Methods for Apparel Applications. *Journal of Fashion Marketing and Management: An International Journal*, *7*(3), Article 3.

Skalkidou, A., Petridou, E., Papadopoulos, F. C., Dessypris, N., & Trichopoulos, D. (1999). Factors Affecting Motorcycle Helmet Use in the Population of Greater Athens, Greece. *Injury Prevention*, *5*(4), 264–267. https://doi.org/10.1136/ip.5.4.264

Smith, C. J., & Havenith, G. (2011). Body Mapping of Sweating Patterns in Male Athletes in Mild Exercise-Induced Hyperthermia. *European Journal of Applied Physiology*, *111*(7), 1391–1404. https://doi.org/10.1007/s00421-010-1744-8

Tiest, W., Kosters, N., Kappers, A., & Daanen, H. (2012). Phase Change Materials and the Perception of Wetness. *Ergonomics*, *55*, 508–512. https://doi.org/10.1080/00140139.2011.645886

Wang, H., Yang, W., Yu, Y., Chen, W., & Ball, R. (2018, October 16). *3D Digital Anthropometric Study on Chinese Head and Face*. Proc. of 3DBODY.TECH 2018 - 9th Int. Conference and Exhibition on 3D Body Scanning and Processing Technologies, Lugano, Switzerland, 16–17 Oct. 2018. https://doi.org/10.15221/18.287

Webster, E. (2021, August 5). An inside look at F-35 pilot helmet fittings. *Air Force News*. https://www.af.mil/News/Article-Display/Article/2719003/an-inside-look-at-f-35-pilot-helmet-fittings/

White, T. D., Black, M. T., & Folkens, P. A. (2012). *Human Osteology* (3rd ed.). Elsevier Academic Press.

Whitestone, J. J., Zehner, G. F., Mountjoy, D. J., Blackwell, S. U., Gross, M. E., & Naishadham, D. (1998). *Summary Statistics and HGU-55/P Feature Envelopes for the 1990 USAF Anthropometric Survey* (Technical Report AFRL=HE-WP-TR-2002-0172). Human Effectiveness Directorate, Crew System Interface Division, Air Force Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA412018.pdf

Yu, Y., Benson, S., Cheng, W., Hsiao, J., Liu, Y., Zhuang, Z., & Chen, W. (2011). Digital 3-D Headforms Representative of Chinese Workers. *The Annals of Occupational Hygiene*, *56*, 113–122. https://doi.org/10.1093/annhyg/mer074

Zhuang, Z., & Bradtmiller, B. (2005). Head and Face Anthropometric Survey of U.S. Respirator Users. *Journal of Occupational Hygiene*, Nov; *2*(11), 567–576. https://doi.org/10.1080/15459620500324727

7 Footwear

Sandra Alemany

ABSTRACT

This chapter describes the use of the Sustainable Product Evaluation, Engineering, and Design (SPEED) process as it relates to footwear, and how it can be used to refine the typical current footwear development process. It also details the complex anatomy and biomechanics of the foot and the resulting complexity of designing effective footwear products. This includes shoe anatomy from a point of view of the fit and functional aspects, and the methodologies to improve the fit of the footwear by following the design loop and the sizing loop. The issues with establishing an effective concept-of-fit (COF) for footwear and related assessment methods to study the physical interaction between the foot and shoe are also discussed. Finally, examples of how to improve the footwear design and sizing are provided in the form of two case studies.

BACKGROUND

Foot anatomy is one of the most complex in the body. The foot supports the body weight, transferring the forces to the ground, and it is responsible for the body movement with efficiency and stability. The anatomical structure of the foot is deformable to enable the proper absorption and distribution of forces through the foot. Thus, foot shape and dimensions are dynamic since they change during body movements such as walking, standing, or running.

The feet are also affected by the footwear. For example, the height of the heel modifies the shape and posture of the foot even for low values. The upper material of the shoe compresses the foot modifying the shape and dimensions while containing soft tissue and restraining the movement of the foot in the shoe. This influence depends on the stiffness level of the upper material and the design of its structure (e.g., pump, sneaker). Therefore, the anthropometry of the foot measured barefoot and in a static posture varies substantially from the foot in real conditions, inside the shoe and during movement. Consequently, the anthropometry of the foot and its variability do not transfer directly to the measurements of the footwear. The creation of footwear that accommodates the anthropometry and biomechanics of the foot requires the development of a form, called a last.

Most current footwear development processes are based on old rules and guidelines that have not been verified and updated according to the specific needs and requirements of the target population (TP). They lack tests and validations with actual people early in the development process. As a result, many footwear products go unsold, must be sold with high discounts, or remain unused because they are uncomfortable or do not function as desired. This wastes materials and money for

DOI: 10.1201/9781003397533-7

both companies and consumers. One of the issues is inconsistent sizing such that customers need a size up or down from their usual size, or the shaping of the size does not work for their foot. This is particularly important in the context of online shopping, and centralized purchasing of footwear (e.g., uniforms, and safety footwear provided by the company). These issues can result when the last and the base size are not validated against the TP.

Footwear is made from components that must be assembled and fit together. These components include the last, the upper material and its pattern, the sole, insole, shank, heel, and often special components such as rigid metal toe caps or sole puncture protection components. Some examples are shown in Figure 7.1. The shoe last is the "central component" that assures the proper integration between all the components. As a consequence, any shoe last modifications and the final approval of the last have implications for the design of all the components.

Another issue is functional design flaws that make it all the way to production. The fit issues are identified too late in the process after the design is nearly complete. This occurs when the evaluation of the final version of the shoe is done as the last step in the process of development of the "base size" instead of the base size being determined by early prototypes. At this point in the process, changes are expensive, time consuming, and fixing the issues may affect several components (e.g., sole, insole). This situation is particularly critical because suppliers of these components are often outsourced or located in other regions.

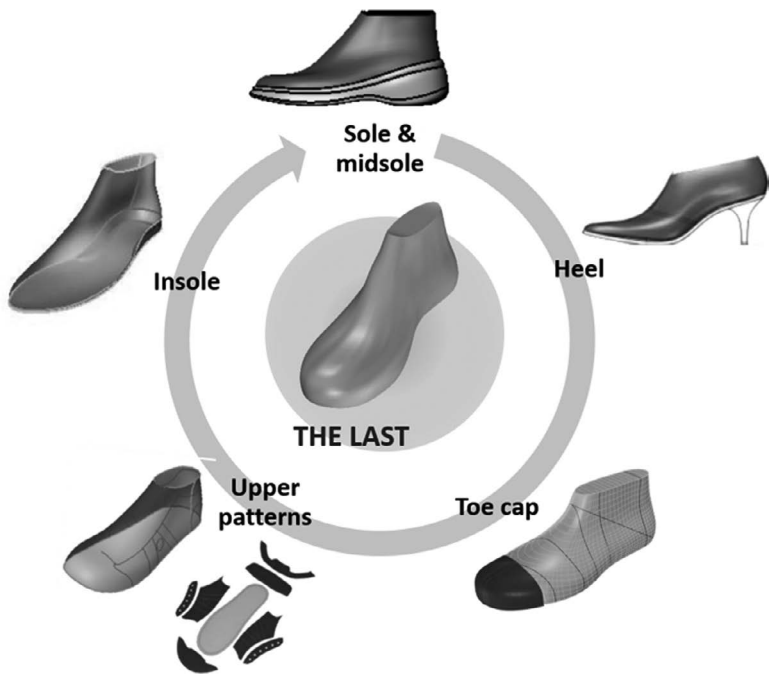


FIGURE 7.1 Footwear components assembly.

Often the requirements of these components change during the development process when partial validations or reviews reveal that design modifications are needed. For effective design and production, these changes must be properly communicated to all the stakeholders involved and adaptations made in a timely and accurate manner. This means that, for footwear, changes made late in the development process are more costly than changes made early. Currently, the decision making is dependent upon the knowledge and experience of the designer or developer with limited data. This is particularly frequent for fashion footwear. With the Sustainable Product Evaluation, Engineering, and Design (SPEED) process, this knowledge is developed into a knowledge database that can be revisited to give the designers and developers more consistent and reliable knowledge for better decision making. The application of the SPEED process to fashion footwear development is explained in more detail in Case Study 1 of this chapter.

The SPEED process provides evidence-based data to designers and developers to help them make more informed decisions. It uses trade studies and prototype testing during design, ensuring that all the components work well together and fit, at a minimum, the base size users before the product is finalized. Any issues are resolved early when changes are less expensive to make, and since the decisions are based on valid data, the decisions are better informed requiring fewer product iterations. It ensures the product accommodates the TP well, minimizing waste and optimizing sales potential. Although it adds some additional steps in the beginning compared to the current process, in the end, we arrive at a better product and often, by refining the product early when it is fast and inexpensive to do so, we save time. The duration of the development process can be controlled by introducing a proper design of the validation test and an optimal integration in the internal development process of the company.

The benefits of the SPEED process are:

- **Optimal fit of the product catalog.** The comfort of the new collections will be improved and can be considered as an identity element of the brand.
- **Consistency of the fit in relation to the sizes.** Better communication with consumers and good consumer confidence in quality and sizing for both online shopping and centralized purchasing.
- **Streamlined design and development process using a fit standard.** Establishing precise criteria for the last dimensions significantly reduces the need for redesign. Then, fit trials could become primarily a verification step, minimizing the required adjustments.
- **Company core knowledge database.** Since the testing and data collection are done inside the company, it forms a knowledge database for future data-based decision making. This improves quality, reduces risk, and reduces time to market for future products.

As in the mass-produced apparel industry, the footwear industry has several good practices, which is why they can sell footwear and make a profit. The SPEED process retains some of these but builds upon them to improve fit quality, sales opportunity, and reduce waste.

First, each season, a new collection starts with a review of the season trends (e.g., material, colors, textures, trends in clothing and accessories) and footwear concepts. This information is then used to define new concepts and aesthetic lines, and preselect components such as the upper materials, ornamental accessories, and heels. All the information collected in this phase is the base for developing the concept and lines of the new collection. This phase establishes the product requirements and is used throughout the industry today. We refer to this as the Inputs phase. The SPEED process adds the development of key performance indicators (KPIs), and the concept-of-fit (COF) to ensure all stakeholders agree before proceeding. Without these processes, the performance and fit criteria can be a moving target leading to continuous changes and needless iterations.

Second, the footwear industry uses cases in the form of shoe lasts development. The last is used to produce the footwear, so rather than being the shape of the wearer's foot, it is the shape of the inside of the shoe or boot. The shape that fits depends on the type of shoe and the complexity of proportioning and shape of the foot of the wearer versus the last.

There are companies that produce and sell lasts to footwear manufacturers, such as Jones and Vining (https://jonesandvining.com/lasts/) that make and sell lasts for all kinds of footwear, Sorrell Notions and Findings (https://sorrellnotionsandfindings.com/product-category/boot-lasts/) that sell lasts specifically for cowboy boots, Crispinians (https://www.crispinians.com/) that make custom wooden shoe lasts using numerically controlled milling, and Podohub (https://podohub.com/custom-shoe-lasts/) that make custom lasts using 3D printing.

The SPEED process selects or creates the base size last using data from the TP and the case selection methods described in Chapter 3, then validate it with trade studies (design loop) to verify and adjust before finalizing it. This ensures the footwear will fit the base size for the TP and will work effectively for the type of footwear and materials to be used. The last is critical to effective fit and fixing any issues early will be cheaper and faster than having to start over with a refined set of lasts. We provide an example of how to do this in the first case study later in this chapter.

The validation process of the last is like the block patterns used in clothing. To validate the fit of the last, a mock-up of the footwear is first manufactured using a similar upper design and materials but using a preliminary sole. A live model, who represents the reference anthropometric foot dimensions of the base size, performs a fit test of the shoe prototype. The feedback gathered is used to modify the dimensions of the last.

Modifying a shoe last is complex. It is an organic shape without geometrical references. The modification in one section must be smoothed to transition to the neighboring sections. Therefore, once the last has been refined and verified, on the fit model, it is important to do trade studies with 3–5 additional subjects to ensure it will work for more than one person. The first iterations do not have to use the final materials. They are a quick check to ensure there are no major problems.

Often the design of the different components is done in parallel, so issues can occur later when the final materials, thickness of the insoles, and other designed features are implemented. Our SPEED process allows for interim checks during the process using trade studies and prototype tests to speed up development, catching

some mismatches early. However, it is also important to validate the fit with prototypes in the final materials and configurations. The prototype test of all the actual components and the complete footwear product should include at least 5–10 subjects representing the full range of fit for the base size. This validates the size.

Sometimes the schedule requirements may put pressure on to reduce or eliminate this fit validation loop. This will increase the risk of a suboptimal fit and footwear performance. Sometimes the results of the tests done in this phase, may suggest the need to refine the design and sometimes to refine the last. A quick test with a few subjects can be reassuring or convincing to management that a more thorough test is needed before proceeding with the other sizes. Once the company has a database of fit and anthropometry and a fit standard with a validated base size last for a TP there will be less need for some of the validation tests.

Finally, the footwear industry has established standard shoe and boot length grading systems (Mondopoint, EU, and UK). Conversion charts are reported in standards (ISO 19407:2023). While these grading systems may not be optimal for every wearable, they are well-established and understood by consumers so it is usually best to use them (provided you validate the base size), unless the footwear is some special type for which a different grade will be beneficial. However, these grades are foot-length grades only and do not address the width or product shaping grade. The footwear developer must determine those aspects, and this is best done with the aid of trade studies and prototype tests in the design and sizing loops to ensure the adequation to the TP. It will also indicate if there is a need for width sizes, in addition to the length sizes.

The SPEED process includes a fit audit in the sizing loop to ensure the graded product will accommodate the TP, verify that the size designations are correct, as well as provide data for the tariff which indicates how many of each size to produce and sell to minimize waste. For a commercial manufacturer it may not be cost effective to produce sizes that will have few wearers. For organizations that must fit everyone, such as the military or firefighters, it may be more cost effective to make use of an off-the-shelf product for most people and a custom product for the few people at the extremes. The fit audit will help make these decisions. The fit audit can also help wearers select their best fit size. This is particularly true if the product comes in multiple widths, or if the best length is affected by the width.

The differences between the current practice and one that has been refined using the SPEED process are summarized in Table 7.1. The SPEED process includes some additional testing and design refinements and has data and testing-based decisions, as well as three additional outputs: (1) a set of sizes based on the TP needs; (2) a foot, last, and fit scores database for its use with future products; and (3) a fit standard for rapid sustainable fit and sizing in the future.

ISSUES WITH DIFFERENT FOOTWEAR TYPES

There are many functional aspects for shoes such as stability, thermal comfort, friction, pressure distribution, flexibility, torsion, and shock absorption, that can be more or less important depending on the type of shoe. These functional aspects affect the fit requirements for instance, a shoe with good stability requires a snug fit in the

TABLE 7.1
Comparison between Current and SPEED Refined Process

Steps	Current Practice	With SPEED
Requirements Established	Yes, usually only for aesthetic aspects	Yes
COF Developed and Approved	No	Yes
TP Sample	No	Yes
Initial Base Size Last Selection/ Creation	Yes	Yes
Base Size Selected Using TP Data	No	Yes
First Prototype/Mock-Up	Yes	Yes
Fit Model Selection Basis	Estimate from previous footwear	Based on TP data
First Last Fit Validation Test	Optional usually only aesthetic validation	Fit model assessment to COF and aesthetic validation
Last Refinement and Interim Prototypes Made	Maybe	Yes
Size, Last, and Component Trade Studies	None	3-5 subjects each iteration
Base Size Produced in Final Components	Yes	Yes
Base Size Fit Validation in Final Components	No	Final Fit Test (5–10 subjects representing full foot range for the size)
TP-Based Size Range and Tariff	No	Yes
Foot Anthropometry, Last, and Fit Scores Database	No	Yes
Fit Standard for Future Products	No	Yes

rearfoot. The multiple variables that affect the footwear fit must be analyzed with a human-in-the-loop approach for effective fit and sustainable fit standards.

Footwear types and styles influence on the shape and performance of the shoe and in the perception of fit and comfort, defining the requirements and specifications of the shoe. This chapter includes only an overview of a few footwear types to illustrate the impact on the requirements and specifications. These requirements should be addressed during the product development process and the design loops proposed. We illustrate some of the issues with examples of three different shoe types.

Dress Shoes and Casual Footwear

Dress shoes and casual footwear are primarily influenced by the style and trends of every season. Some brands may introduce their own elements according to the identity of the brand image. Therefore, most of the design elements are selected or defined following the study of new trends.

While casual footwear are everyday items that are expected to be very comfortable, dress shoes might be shoes that are required in the workplace, or that are only worn for special events or occasions. When worn for special occasions there tend to be additional criteria such as following aesthetic fashion trends, having the perception of high quality, not looking worn, creases usually are avoided, and providing a "perfect look" in combination with the clothing and other accessories. This type of footwear is usually very uncomfortable compared to other shoe types and is conceived to be worn in limited situations and duration. However, this shoe style has been adopted by some companies and professionals as part of the accepted dress code at the office. When comfort is not the priority, the design of the shoe can go against good ergonomic principles. The regular use of such dress footwear may cause many foot problems resulting in painful feet. This is one reason there are many types of pads for the heel, toes or plantar metatarsal heads, developed as accessories to reduce the pain caused by this type of footwear. A challenge of the development process of this type of footwear is to mitigate the discomfort by introducing innovative solutions.

For women, the traditional dress footwear is a pump with a heel height that can range typically from 4 to 10 cm. This can be a source of discomfort and pain. Those with especially high heels (6–10 cm), are usually narrow and pointed to create the desired stylized shape and the upper material used needs to be rigid to avoid creases at the flexion line of the forefoot. The components of a high heel pump are shown in Figure 7.2.

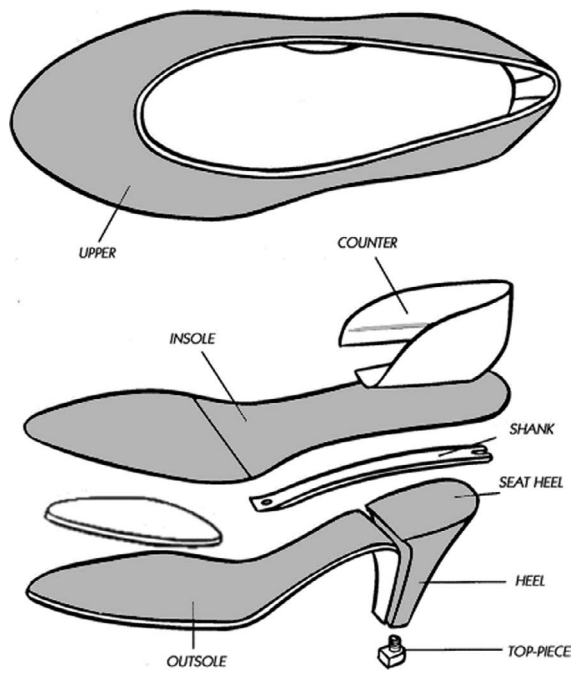


FIGURE 7.2 Components of a high heel shoe.

Some of the most challenging issues are introduced by:

- Heel height and shape
- Shoe last size and shape
  - Wide versus narrow shape
  - Additional thickness to accommodate the removable insole
  - Front to back shape
  - Instep height
  - Toe cap shape
  - Fastener, if any (such as laces or straps)
  - Coordination and modification according to the insole thickness and sole shape
  - For soles provided by a separate sole manufacturer the last is adapted to the sole shape
- Sole
  - Type
  - Material
  - Shape
  - Outline and top surface of the sole should fit the outline and bottom surface of the last
- Insole
  - Materials
  - Thickness (often not uniform, but thicker in the heel area)
  - Coordinated to match the last

Since the upper material is not deformable, the accommodation of a large variety of foot shapes and widths is not feasible, and it is also difficult to adapt to the shape variation during walking. In other words, they can be uncomfortable and difficult to walk in. In addition, they are highly unstable and more susceptible to producing plantar foot pain because they concentrate the pressure under the metatarsal heads and toes region as illustrated in Figure 7.3.

The foot slides forward in these shoes and to limit this slippage, the forefoot area must be tight fitting. As a result, this type of shoe produces a high compression of the toes and the metatarsal arch as shown in Figure 7.4.

The high pressure toward the front of the foot and the compression of the toes create a trade-off for the footwear designer. A wider or longer shoe will provide a better fit and will accommodate wider feet but it will produce more slippage, especially in

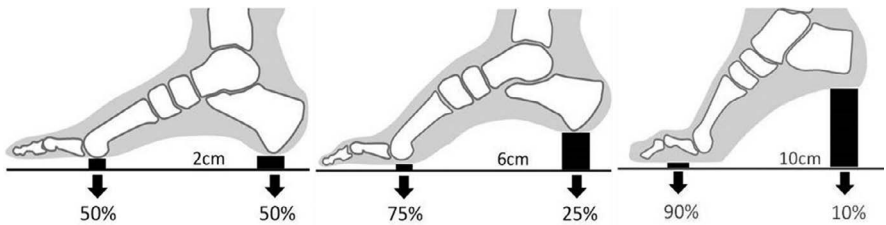


FIGURE 7.3 Pressure distribution under the heel and metatarsal heads due to the heel height.

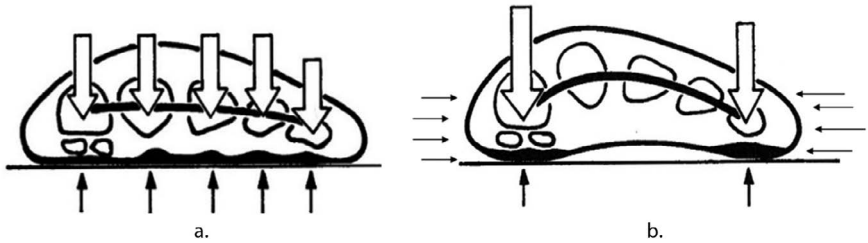


FIGURE 7.4 (a) Barefoot toes front view. (b) Compressed toes in a tight shoe.

narrower feet. Both effects can only be effectively measured with fit testing on live subjects and this kind of preference will influence not only the design but also the accommodation range of foot anthropometry and how many of each size will be needed for a given TP. When establishing the requirements for this type of shoe, the accommodation range and the relationship with the distributions of foot width for the base size must be included. If this information is known during the requirement definitions, a cost-benefit analysis can be included as a result of the fit test and the projections against TP to decide the efficiency of creating additional width sizes.

The sole for a pump is usually thin and is made of hard leather without any capacity for pressure distribution. This is another element of discomfort. It includes a metal part in the shank to provide high rigidity and avoid a break in the central part of the shoe that is not in contact with the floor. To mitigate discomfort, it might be important to allow room for a cushioning insert.

A ballet flat is a type of footwear in between casual and dress styles for women and is like a pump in that it has no instep fastener, such as laces or straps, but it has a very thin heel, 1 cm or less. Like the dress pumps it can have an issue with heel slippage when walking due to the interaction of forces as shown in Figure 7.5.

Since ballerinas are usually made of a softer material, slippage can be a problem unless the length is very precise and tight. A loose fit produces uncomfortable slippage in the heel that is compensated pressing with claw-shaped toes. On the other hand, a tight-fit causes overpressure in the heel and the dorsum of the toes that often produce blisters. The upper material of ballerinas may need some stretch and the last may need to be shorter for a given size than the last of a pump. Comfort and slippage are competing issues that can only be effectively assessed by means of tests with users. This can best be done with a prototype fit test of the base size to establish the range of fit within the size and the suitable sizing of the base size last.

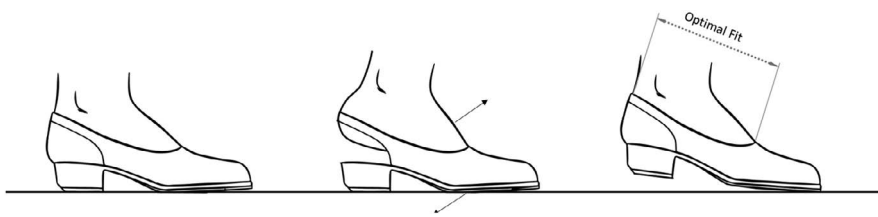


FIGURE 7.5 Interaction forces during walking for a pump without a fastener at the instep.

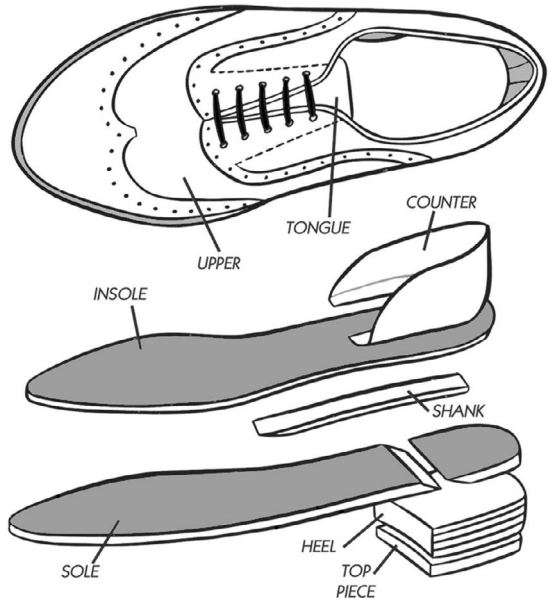


FIGURE 7.6 Components of the traditional Oxford shoe.

A traditional type of men's shoe is the Oxford flat. Originally, Oxfords were formal, laced shoes for men, made of leather. The typical components of an Oxford shoe are shown in Figure 7.6. They have evolved into a range of styles (e.g., derby, monk, blucher) and there are also versions for women. Traditionally, they have a pointed toe shape, and they are made of rigid leather for the upper and sole. As a result, it provides a tight fit and a small range of accommodation. For this reason, this type of shoe is usually offered in extra width sizes.

Typically, the Oxford has a thin sole made of leather that prevents flexion at the metatarsal region. This is required to avoid the generation of anti-aesthetic creases and wrinkles in the forefoot. The lack of flexibility and extra length modifies the person's gait pattern making it more difficult to bend the shoe around the metatarsal arch of the forefoot. In order to support gait movement in the case of rigid soles, some shoes are designed with a toe spring (see Figure 7.7) that creates a "swing

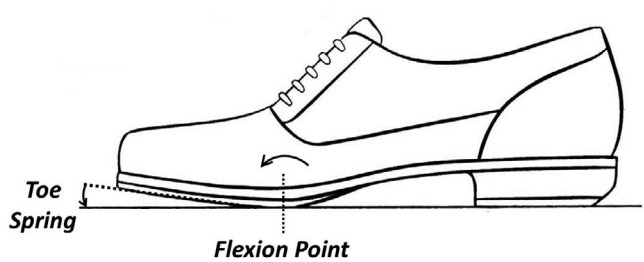


FIGURE 7.7 Toe spring design element used to support the movement of the foot.

effect". Sometimes the sole or the mounting plant has a thin layer of soft material to provide a certain cushioning effect. The mounting insole is used as an intermediate component between the upper material and the sole.

While the Oxford is a traditional style available in the databases of all the last makers, it is important to adapt it to the anthropometry and fit preferences of the TP. Extra widths can be considered to increase the accommodation rate of each size, but all widths may not be needed for all markets. Moreover, the pointed toe requires extra room in the toes, therefore they may need extra length to maintain the functional fit of a given foot length so the length sizes may be needed. Finally, markets with different ethnic mixes may require lasts of different shapes. For example, a Japanese market may require lasts that are shaped differently from front to back than a European market. These things can be verified using a sizing loop fit audit, or full fit test. This will indicate the number and assortment of sizes needed for any market and will indicate if a new last shape is needed for some or all of the new markets. If a new last is needed it will need to be adjusted and tested with design loop prototype fit tests.

Safety, Protective, and Occupational Footwear

This type of footwear incorporates specific features introducing elements which influence the structure of the footwear to protect the wearer from injuries that could arise through workplace accidents. It must accomplish the regulations of each market and some types must be tested and certified by external laboratories. Safety footwear is considered personal protective equipment (PPE) and the requirements related to the protection depends on the category according to the level of protection (ISO 20345:2021 Personal protective equipment – safety footwear). Some categories include:

- **Safety and protective footwear** – includes a toecap designed to ensure protection against impact and compression loads. Usually, they include additional protective elements such as the anti-penetration insole, protection of the ankle, metal protection of instep and metatarsal regions as well as specific properties of the sole (i.e., material and geometry) to prevent slips, trips, and falls.
- **Occupational footwear –** does not include toecaps. These types of footwear are designed for workplaces and activities where both hazards, such as impact and compression, to feet and toes were not identified during the risk assessments.
- **Special application footwear –** requires modified types of footwear characterized in their dedicated standards related to specific activities and hazards; resistance to chain saw cutting, protection against chemicals, electrical isolation, fire resistance, welding tasks, protection against cool environment, etc.

The list of requirements in the case of safety and occupational footwear can be large compared to other types of footwear. Since they are designed to protect against hazards, most of these requirements are solved using metal parts and strong materials

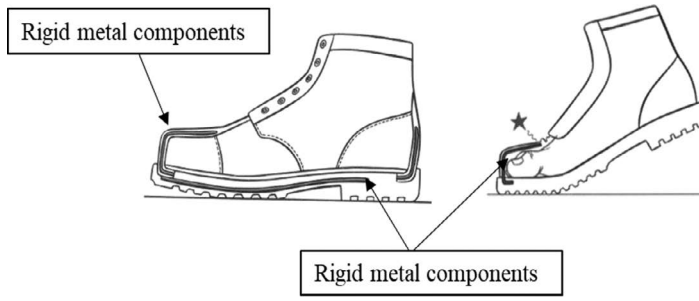


FIGURE 7.8 Effect of metal parts in shoe fit (Ramiro et al., 1995).

that may reduce the ergonomics and comfort of the shoes. Thus, a key challenge of safety footwear is to meet the safety requirements while providing an effective and comfortable fit. Consequently, the footwear specifications are defined after a risk assessment of the work environment, and the work environment conditions must be included in establishing the COF and the test conditions for trade studies and prototype tests. There can be a wide range of variants such as:

- Safety footwear for indoor industrial applications
- Safety footwear for outdoor construction
- Firefighters footwear for extreme heat conditions
- Military footwear for prolonged outdoor wear
- Medical clogs for indoor use with long duration standing

Safety footwear sometimes requires a metal toe cap and an anti-perforation metal insole. This can necessitate extra room at the toe area to accommodate the lack of flexibility of metallic parts as illustrated in Figure 7.8 and minimize discomfort due to rubbing or pinching.

The toe cup introduces an extra volume and it is responsible for the bulky appearance of this footwear. Safety boot upper material covers all the foot and is thicker compared to other types of footwear to ensure appropriate foot protection. The bottom is usually very thick including different layers which may limit the flexibility, adaptation to the movement and proprioception of the user. As a result, safety footwear is often very heavy, robust, and non-deformable. The requirements for these items might include an assessment of the wearer's ability to do his or her job or to accomplish the required tasks.

Safety footwear used at the workplace will be worn all day and used while performing different, often repetitive tasks. It can be important to establish and test the type of postures and activities required by the user in the course of their work. For instance, the requirements will be different for a workplace that requires a standing posture most of the time and a work that requires a combination of different activities and posture (e.g., walking, driving, squats, load handling). In the first case it is better to specify requirements related to pressure distribution of the foot plant in standing posture to reduce foot pain, while in the second case, it will be more

optimal to improve the flexibility of the shoe to adapt to the movements and extreme postures and to reduce the overall weight of the shoe.

Medical clogs are a type of occupational footwear that requires to be easy to put on and take off, be resistant against contaminants, be comfortable for long duration standing and walking tasks, be designed to prevent slips, trips, and falls and be easy to clean and disinfect. They usually include an anatomical insole to reduce overpressures and allow for comfortable standing for long periods of time.

The use of safety footwear is mandatory thus, it should accommodate almost all foot types in off-the-shelf sizes with the option of custom fit sizes for extreme sizes or pathologies. Sizing loop fit testing is important for this purpose and can help in making decisions about which sizes to produce in quantity, how many of each size to produce or purchase, and which to custom-make.

Sports Footwear

The development of athletic footwear is based on the specific needs and requirements for each type of sport (e.g., running, hiking, golf, basketball, soccer). The biomechanics of the movements of certain sports are relevant to optimize performance and to prevent injuries. With this aim, requirements related to the grip of the sole for specific sports surfaces and movement directions, the energy return, the stability, or the shock absorption are some of the priorities for the development of sports shoes. For the adequate performance of these functional aspects of the shoe, fit is also an important factor. For instance, a sole and a midsole designed to achieve good stability will not be efficient if the fit of the shoe at the heel area is too loose.

Some important aspects for establishing requirements include the sports surface, expected sports movement, biomechanical loads, and shoe stability on the foot. The type of sport surface is critical for the design of the sole. The grip on the sports surface is important for providing protection against slips, trips and falls as well as allowing for the required dexterity for a specific sport performance. Some examples are golf shoes (natural grass), soccer (with different soles for natural and artificial grass), running shoes, basketball shoes, tennis (considering also different types of surfaces), and mountain or climbing shoes. The type and material of the sole necessarily impact the last and the overall fit.

Performance of sports movements is key to achieving good results in any sport, and the fit of the shoe can be an important factor. Each sport can include a single or a combination of different sports movements such as straight running, lateral displacement, sprint, change of direction, vertical jump and landing, torso rotation, cycling, and golf swing, and each can require different properties in the shoe. Some examples are illustrated in Figure 7.9. In some instances, the shoe can be required to help prevent certain movements such as ankle pronation, supination, or knee torque.

The level and type of biomechanical loads are relevant to define the requirements for the protection of the body that can be introduced by different components of the shoe. The specifications and solutions to protect the athlete (e.g., energy absorption of impact loads) are sometimes negative to maximize performance. This is the reason why different shoes are often designed for training and for competition, focusing on protection or performance.

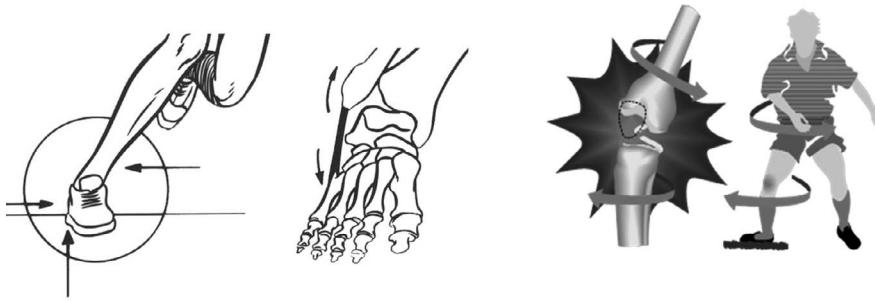


FIGURE 7.9 Forces during sports movement that may require specific fit and optimize the surface grip to avoid injuries.

Prototype fit testing is usually important to maximize the behavior of different components of the shoe (e.g., stability, movement adaptation, thermal comfort). Additionally, there can be some specific sports elements that should be assessed. For instance, in the case of soccer boots, fit is important to have a good control of the ball (Olaso Melis et al., 2016). In these situations, it can be very important that experienced athletes in these sports are used as test subjects.

Running shoes typically include a bulky bottom with a thick midsole made of cushioning and very light material, and a thin sole to provide an optimal surface grip. New trends in running footwear also include a curved carbon plate and thick midsoles of reactive materials that provides a return of energy. Lightness is directly related to performance and the upper material usually needs to be light, breathable and can be adapted to different foot shapes. Sometimes the upper material can be too flexible allowing the foot to shift left and right over the sole, particularly when running over uneven surfaces. Experienced runners will be able to provide immediate feedback about these sorts of issues.

The structure of soccer shoes is determined by the design of the sole. It is a thin plate with a distribution of studs at the forefoot and heel to increase the traction with the turf surface. Leather is the most frequently used material for the upper part of the shoe use to its long-lasting durability. For soccer footwear, the impact and ball control with the foot dorsum requires a high durability and a good perception of the ball touch therefore, the upper material is very fitted to the foot. Only experienced soccer players can give good feedback about these issues.

Sports footwear always requires good properties of breathability and sweat absorption. Additionally, the environmental conditions should be considered. For wet outdoor activities (e.g., hiking, trail running, climbing), shoes include waterproof membranes as well as protection against cool temperatures in the case of winter sports. Therefore, it can be important to set up and maintain the environmental conditions of the sport when doing prototype testing.

KEY PERFORMANCE INDICATORS (KPIs)

KPIs are the defined acceptance or rejection criteria for the product. KPIs for footwear are defined for both the single components and the complete footwear. Some

of these items are not directly related to fit, such as the material quality or the cost of a component. One KPI for fit might be something like, "90 percent of the TP will achieve a comfortable fit". An acceptable fit for an individual is defined and documented in the COF.

KPIs must be documented and measurable. Initially, it is common for the requirements and the acceptance criteria to be set too high and sometimes are difficult or even impossible to meet. During the development process, as we learn what is possible to achieve, the specifications are often revised and updated. KPIs allow us to track these changes. There are also some competing requirements, so it is necessary to prioritize them or establish the points of compromise. Well-documented KPIs help us accomplish this.

There are several types of KPIs related to footwear including:

- **Quality –** these tests are usually performed in the footwear industry. They relate to basic quality specifications and are highly related to the price of the components. Manufacturers of footwear components use KPIs to control the quality of upper, insole and sole materials, laces, and footwear manufacturers also use these tests to control the quality of their products.
- **Safety** – safety footwear should comply with a list of relevant standards such as resistance against an impact load in the toes, anti-perforation of the sole, antistatic or electric isolation.
- **Functionality –** functional KPIs are not very common for most footwear today. Only companies interested in research and innovation include functional tests to ensure a high-level comfort for the wearer of the product. In particular, sport footwear companies are the pioneers in controlling the functional KPI of the shoe. This type of KPI can be part of the COF.
- **Overall fit –** fit KPIs usually indicate the proportion of the TP that must be accommodated in the footwear. This is documented in the COF along with a definition of what constitutes a fit.

KPIs can be measured in many ways. Physical tests can be applied to single components or to the complete footwear. They might include the assessment of aspects such as strength of materials, compression test of insole, midsole and sole, stitching strength, friction of the sole with the ground, wear, or ultraviolet resistance.

Fit and function KPIs usually require the use of human subjects, a complete footwear product as a complete pair. Only one foot might be needed in the first stages of the development when there is just a mock-up to validate the last fit and the upper design. However, footwear performance overall will require both. The KPI must specify what level of fit or function is a pass, and/or what percentage of the subjects must pass.

Biomechanical and physiological tests are used to analyze specific biomechanical properties of the footwear. These can be trade studies with a few subjects to compare different design alternatives, compare to a past design or compare the performance. For a trade study, the KPI might be something like, "the new design performs as good or better than that past design".

There are no thresholds or performance scales for biomechanical and physiological tests thus, the assessment should be done in a relative way comparing different

conditions. However, a company that regularly includes some of these tests in the product development process could generate a database to assess and compare products.

Some biomechanical and comfort tests include:

- Shock absorption measured with accelerometers in the leg and head
- Ground reaction forces
- Pressure distribution on the footplant
- Movement analysis of the ankle and metatarsal joints
- Temperature and humidity in several foot zones

A powerful element to establish thresholds or target values for the biomechanical and physiological tests is to relate them to user perception tests.

Perception and comfort tests are subjective evaluations of different aspects of the product. They should be done performing the type of activity considered in the product concept and they should be done by subjects, regular users of that type of product. At the initial stages of the development, there are usually quick tests done just to discard some options or to validate the shoe last. When the first prototypes of the complete shoe are ready, it is recommended to perform a longer-use test including the assessment of several functional aspects.

The structure of the KPIs and the type of tests can be used to create new knowledge relevant to product innovation by relating "Inputs" (characteristics of the TP and product properties) with objective biomechanical and physiological variables and comfort values. These are then used to establish and prioritize the thresholds or biomechanical scales of goodness. This approach (see Figure 7.10) was proposed by García et al. (2021) based on a three-level framework. When a company creates a database of shoe tests including the user and shoe characteristics, the physical test results, the biomechanical test results, and the subject perception of fit and comfort,

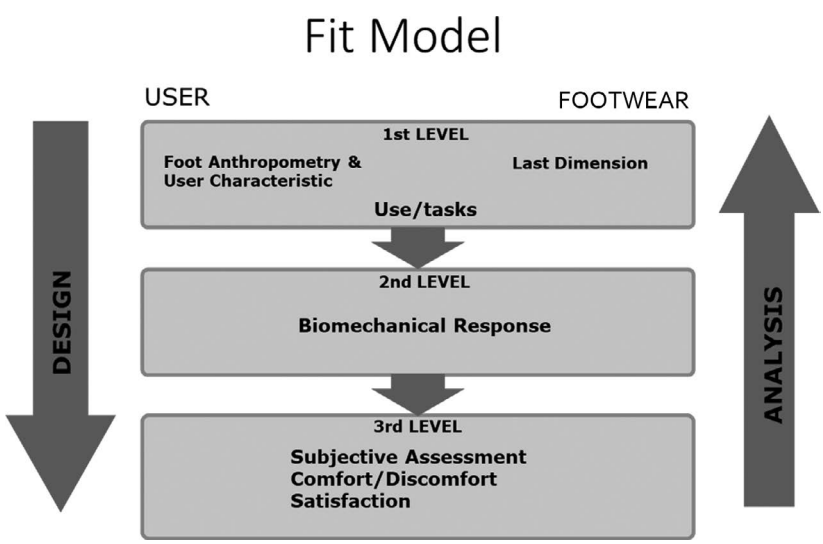


FIGURE 7.10 Structure of variables' relationship to determine design rules and assessment criteria.

it is possible to relate the variables from top to bottom to create design criteria and rules as well as to relate the variables from bottom to top in order to establish assessment thresholds and priorities for each criteria or test. During the product development process, it is common to perform physical tests of the shoe prototypes and tests of single components (e.g., upper material stiffness, cushioning of insole materials). This is important to make good decisions related to the proper selection of material and its thickness as well as to decide between design alternatives. Saving the test results in a structured database can inform future footwear development, improve our test methods, and potentially reduce the need for some tests.

CONCEPT-OF-FIT (COF)

The specification of fit will depend on the type of footwear namely, the intended use and the desired performance. Regardless of the footwear type it is usually important to assess the comfort, tightness, looseness, and stability in different areas of the foot including the length, the heel area, the instep, the ball of the foot, and the toe area.

A set of postures, movements, and performance criteria should also be defined. The subject will perform the movements before answering the fit perception questionnaire. The type of movement can range from a standard walk around in the case of casual footwear or the simulation of extreme postures such as a squat or going on tiptoe. An example is shown in Figure 7.11.

The questionnaire is part of the COF. An example of a footwear questionnaire is provided below in Table 7.2.

Subject Assessment

The technician asks the subject about the fit perception and how they prefer it at each foot region starting from the length. It is necessary to be sure that the subjects understand each foot area. Fit perception is sometimes confused and mixed with personal preferences. This is the reason why both questions are separated as illustrated in Figure 7.12.

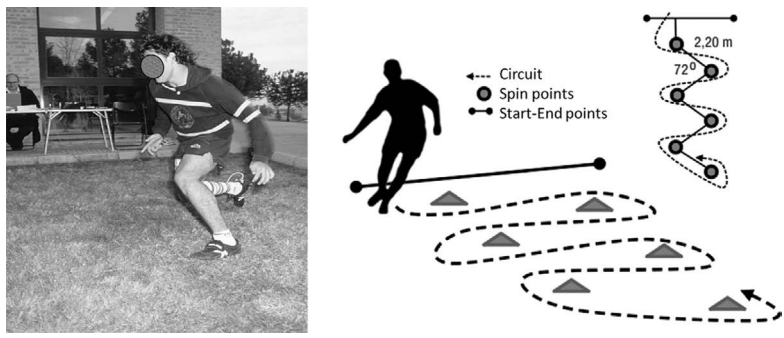


FIGURE 7.11 Example of outdoor trials to simulate real conditions: surface (natural grass) and movement.

TABLE 7.2
Footwear Questionnaire

1. DESCRIPTION OF THE SHOE:

SHOE CODE: _____ SHOE SIZE: _____

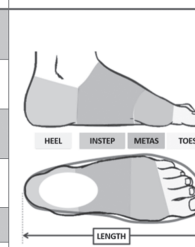
2. FIT ASSESSMENT

- The subject, sits in a chair, removes his shoes, puts on the socks provided by the technician, and puts on the shoes to test.
- The technician checks that the shoe is fastened properly: not too loose, not too tight.
- The technician checks that the shoe is the proper size.

The subject wears the shoes 5–10 minutes performing different movements. After that, the following questions will be answered first by the subject and later by the technician considering the fit areas of the picture.

3. EXPERT ASSESSMENT

The subjects' fit criteria are variable and can blur the results. It is recommended to include a more consistent fit assessment by the technician. In this case, the technician cannot feel the fit therefore, the assessment includes several objective checks as shown in Table 7.3.

	Fit perception (subject answer)					How it is preferred (subject answer)			Fit areas
Length	Very short	Short	Ok	Large	Very large	Shorter	As is	Larger	
Heel	Very tight	Tight	Ok	Loose	Very Loose	Narrower	As is	Wider	
Instep	Very tight	Tight	Ok	Loose	Very Loose	Narrower	As is	Wider	
Ball	Very tight	Tight	Ok	Loose	Very Loose	Narrower	As is	Wider	
Toes	Very tight	Tight	Ok	Loose	Very Loose	Narrower	As is	Wider	
General	Very tight	Tight	Ok	Loose	Very Loose	Narrower	As is	Wider	

2. How do you perceive the flexibility of the upper material?

Very Rigid	Rigid	Ok	Flexible	Very Flexible

3. Comments of the subject

_____

_____

FIGURE 7.12 Fit perception.

TABLE 7.3
Expert Assessment

Fit Area	Expert Assessment (Expert Answer)				
Length	Very short	Short	Ok	Large	Very large
Heel	Very tight	Tight	Ok	Loose	Very loose
Instep	Very tight	Tight	Ok	Loose	Very loose
Ball	Very tight	Tight	Ok	Loose	Very loose
Toes	Very tight	Tight	Ok	Loose	Very loose

FIGURE 7.13 Check the gap in length using gauges with different diameters.

Check the following aspects to support the expert assessment:

- **Length fit:** Check the extra room (gap) of the shoe pressing the toes if the upper material is soft or measuring the gap at the rear heel by moving the foot forward (see Figure 7.13).

  Length GAP (mm): _____

- **Heel width:** Check the fit at the internal and external sides of the heel by estimating the gaps. Hold the shoe in the heel area and ask the subject to rise the heel simulating a step. Check if the foot slides in the heel area.

  Lateral heel GAP (mm): _____
  Medial heel GAP (mm): _____

- **Upper height at the heel** (under the lateral malleolus): Check the height of the upper material under the lateral malleolus in standing posture (see Figure 7.14).

  GAP (mm): _____

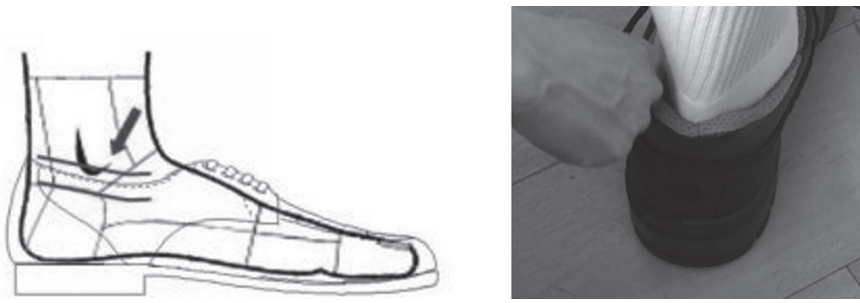
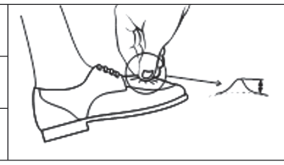


FIGURE 7.14 Check the gap in the lateral side of the shoe.

Instep fit: Check the fitting at the instep estimating the maximum gap at the summit of the instep.

Ball and toe fit: Pinch the material. How much material can be caught? (The subject must be in standing position, and equally weight bearing on both feet).

Gap internal side: _____mm	
Gap external side: _____mm	
Gap upper part: _____mm	

Comments of the expert

FIGURE 7.15 Assessment of the instep, ball, and toe fit.

Check the instep and ball and toe fit (see Figure 7.15).

Pain Points

The subject walks again 1–2 minutes. Afterward, he or she indicates any pain point at the foot surface. An image of the foot divided into zones could be used to help the user identify the zones of discomfort and pain in the questionnaire.

THE SUSTAINABLE FIT STANDARD

Once a product has been developed and a fit audit for a TP has been completed, it is possible to create a fit standard to apply to future products or to all products in one line of footwear. This helps to sustain the fit across products, so the customers know what to purchase, and it helps reduce the production of products or sizes that won't be needed. For footwear, a sustainable fit standard includes:

- Shoe last:
  - Physical or digital (last model in a CAD software) copy of the last(s)
  - Last measurements
  - Reference values of the dimensions for different footwear styles made on the last
  - Foot anthropometry for a range of people accommodated by the last
  - Tolerance of the reference values
- Rigidity of the upper material:
  - Different levels of upper material rigidity can be established
  - Control employing tensile testing
  - The tolerance of the reference values
- COF document and test protocol
- Any other component or total shoe measurable factor

TARGET USER PROFILE

Foot dimensions, shapes, anatomy, and function characteristics such as walking patterns vary with gender, age, and demographics. Adult male and female feet are proportioned differently. For example, women typically have a narrower foot for a given length than men. This can cause women to have fit issues with footwear designed for men and scaled down for women. This is a typical situation in safety footwear.

Aging affects anatomy and limits the functions of the body. Requirements related to footwear for the elderly should consider:

- As we age, we experience more deformations and pathologies such as hallux valgus and hammer toes.
- There is a reduction of soft tissue under the foot plant that provides a natural cushioning, bursitis, plantar fasciitis, dry skin, or problems related to diabetic foot that can be very serious.
- The toenails are also often affected by age. They can become thicker, raised, and more brittle, requiring more toe room.
- Aging is also associated with several changes in joint physiology that may contribute to the reduced range of motion in lower extremity joints. Due to the important role played by the foot in adapting to uneven terrain, a reduced range of motion in the joints of the foot and ankle is strongly associated with impaired balance and functional ability in older people. Movement limitations of elderly people are also related to an increase in the risk of falls which can be serious for the elderly.

For children, the foot grows very fast and their feet still have an immature structure that is going through the developmental growth phase. Hence, a poorly fitted child's shoe may cause severe foot problems. The main aspects to consider in the design of children's footwear are:

- *Feet anthropometry:* The foot is growing but the manner of increasing the size is not regular. They are also changing the proportions and shape. Thus, the children's feet are not a scaled version of adult feet.
- *Anatomy:* The anatomy is also different. During the first stages of growth, some bones are cartilages under formation. The longitudinal arch is not formed so they usually have flat feet. Children's feet are also fleshy and very flexible. Proprioception, which is the way joints and muscles send messages to the brain to help coordinate movement is an important feedback for children during the development of gait.
- *Gait pattern development:* Between 0 and 14 years old children develop their gait patterns in different stages. Their footwear should be developed according to the specific requirements of each stage and consider not only fit but also biomechanics of the gait, proprioception, and protection for thermal comfort and stability. This knowledge is important to design an

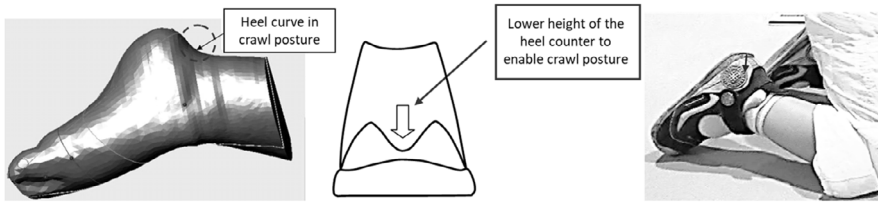


FIGURE 7.16 Posture and shape of children's foot in a crawl posture. Product developed by Garvalin (https://www.biomecanics.com) in collaboration with the IBV (https://www.ibv.org).

ergonomic structure of the shoe for children but also to consider the fit of the shoe in different situations and postures of the children according to their age and gait pattern (see Figure 7.16).

- Between 6 and 12 months are mostly crawling or may have a pre-walking gait such as that shown in Figure 7.16.
- Between 9 and 24 months they are both crawling and taking early steps.
- Between 1.5 and 4 years their gait is maturing.
- Between 5 and 14 years they are continuously improving with running and additional physical activities.

There are three main implications for fit requirements related to the geographic market:

- *Sizing charts:* Big brands are selling in an international market. It is important to label the size according to all the current existing systems: Mondopoint, European, UK, and USA (see ISO 19407:2023, Footwear – Sizing – Conversion of Sizing Systems, 2023). However, the correspondence between sizes is not direct. Full size length increment is 6.67 mm in the European sizing system and 5 mm for Mondopoint. UK and US sizing systems use half sizes for most of the shoes with a half-size length increment of 4.23 mm. This is a problem for size conversion and also to achieve a consistent fit. Local brands are developed based on a sizing system and a grading method.
- *Foot anthropometry variability:* Foot anthropometry differs along different geographic locations and in particular for main target markets, North America, South America, Europe, and Asia. This variability should be considered in the design and development of the shoe last.
- *Fit preferences:* Fit preferences are individual and influence both, the choice and perception of the fit and comfort of footwear. However, fit preferences may vary also among regions due to cultural aspects.

Foot Anthropometry and Anatomy

To understand the functions of the foot and how to maintain and support it with adequate footwear it is important to know the basis of the anatomy and functions

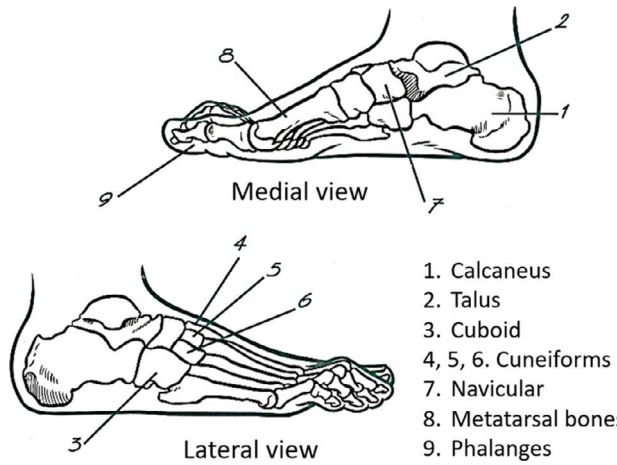


FIGURE 7.17 Foot anatomy. Main bones.

of the foot. There are many books where this knowledge is explained in detail. One recommendation is The Science of Footwear Chapters 1 and 2 (Goonetilleke, 2012). In this chapter, we have included a short overview necessary to introduce foot anthropometry.

The human foot is a complex structure formed by 26 bones (a quarter of the total skeleton), 33 joints, and more than 120 muscles, ligaments, tendons, and nerves (see Figure 7.17). These anatomical structures support and maintain the balance of the body mass during the execution of human activities. They work together to support the weight of the body, act as shock absorbers, keep you balanced, and push the body forward with each stride.

Foot as a Static Structure

The weight-bearing foot is quite different from the static or at-rest foot. It is a different foot in shape, size, and proportions. However, the shoe must fit both the static and weight-bearing foot. It can help to know what changes occur in the weight-bearing foot:

Kapandji (1970) proposed that, as a weight-bearing structure, the foot works as a dome supported by three arches, two of them longitudinal (medial and lateral) arches and one anterior-transverse arch. The three arches are in contact with the floor at three points: (A) the joint of the first toe and the first metatarsal, (B) the joint of the fifth toe and the fifth metatarsal, and (C) the posterior calcaneal tuberosity (see Figure 7.18).

The **internal longitudinal arch of the foot** (A–C), called the planar arch, is characterized by its extraordinary mobility. Its main functions are maintaining balance and adaptation to different terrains by means of its deformity and absorbing the inward and outward rotations of the leg. It is also responsible for modulating the

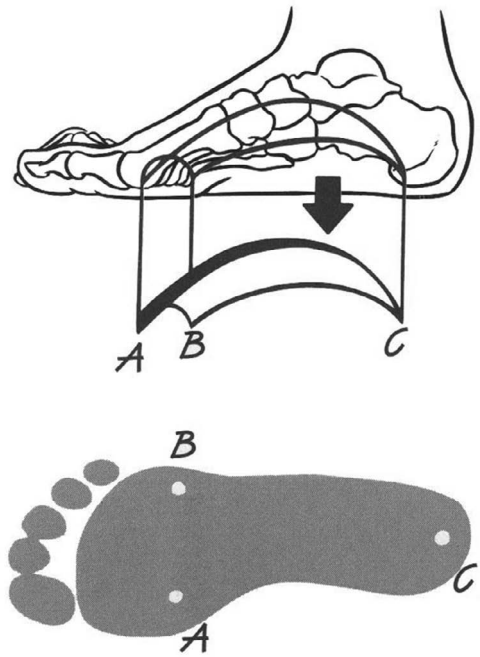


FIGURE 7.18 Foot arches (Ramiro et al., 1995).

stiffness of the foot to allow efficient propulsion during walking and running gait. The bone structure of the internal arch is supported by both ligamentous and muscular structures that span the length of this arch. The passive structures (e.g., the plantar fascia, long and short plantar ligaments and the calcaneonavicular ligament) have a particularly important mechanical role in modulating two known mechanisms that are believed to enhance human locomotion: the arch-spring and the windlass (Welte et al., 2021) (see Figure 7.19).

The **transverse arch of the foot** (A–B) is formed by the five metatarsal heads. When the foot touches the ground and supports body weight, the transversal arch is flattened and extends toward the lateral sides (see Figure 7.20). It also expands in the propulsive phase of the gait or in the case of wearing high heels. The transverse arch is

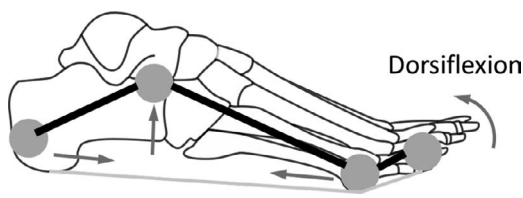


FIGURE 7.19 The windlass and arch-spring mechanisms.

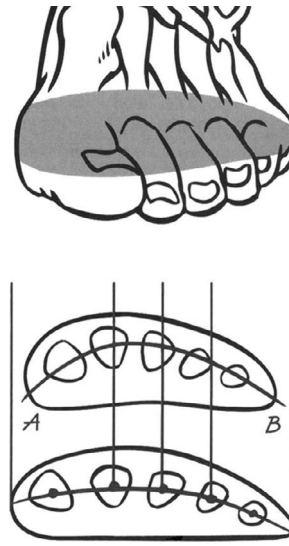


FIGURE 7.20 Transversal section of the metatarsal heads.

very relevant for footwear fit. The foot inside the shoe is totally compressed in the area of the transverse arch changing the main dimensions due to its deformation capacity.

Finally, the main characteristic of the **external arch** (B–C) is its rigidity, partly due to the force of the plantar ligaments and the bone structure. In standing posture, the deformity of this arch is minimal. The soft tissue that covers the bones of the feet in this external foot region is in permanent contact with the ground and supports the body weight. During locomotion, the rigidity of this arch transfers the movement of the rear musculature of the leg toward the forefoot.

The Dynamics of the Foot

The foot is a dynamic structure, and the shoes should be able to adapt to the deformation that occurs during the performance of different activities such as walking, running, or climbing. Knowledge about the biomechanics of the movements is essential to establish the requirements and specifications to achieve an adequate dynamic fit. There is reliable accessible literature about the gait cycle and running patterns such as (Nigg, 2010; Richards et al., 2022) a comprehensive review of the anatomy and kinematics of the lower limb, hips, and center of gravity (Ledoux & Telfer, 2022). In this chapter, the topic is introduced only to highlight the implications of footwear design.

The main joints of the foot are the ankle and the metatarsal joints:

The ankle joint: The main ankle joint is the tibiotalar (talocrural) joint (between the talus and the tibia) responsible for the movement of dorsiflexion and plantar flexion as well as some degree of ab/adduction of the foot. The combination of these motions is the result of the angle of the axis of this joint in the frontal and transversal planes (see Figure 7.21).

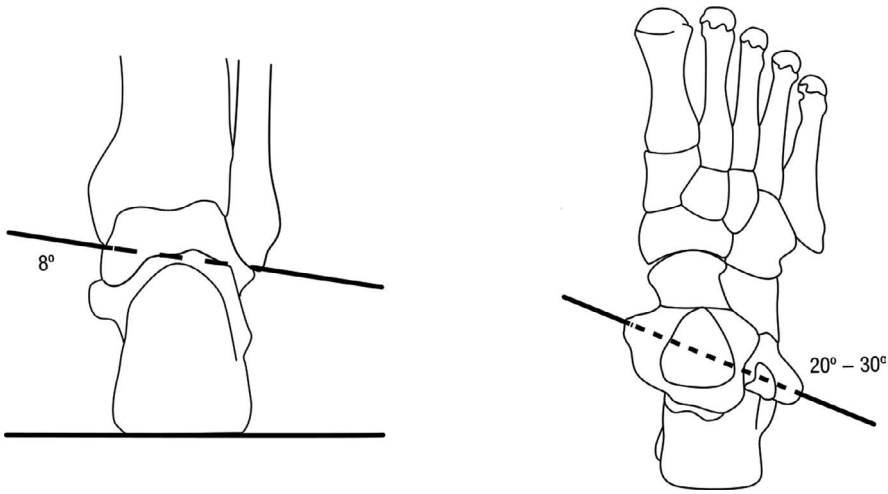


FIGURE 7.21 Axis of the tibiotalar (talocrural) joint.

Talocalcaneal (or Subtalar) Joint and the Transverse-Tarsal (or Talocalcaneonavicular) Joint

They are also responsible for the kinematics of the rearfoot that results in a combination of plantar and dorsiflexion, occurring in the sagittal plane; ab/adduction occurring in the transverse plane and inversion-eversion, occurring in the frontal plane. The subtalar joint axis (see Figure 7.22) is a line pointing from the ground surface on the posterior and lateral aspect of the foot toward the medial anterior of the foot and inclined by about 42 degrees. The rotations about the subtalar joint axis are defined as pronation and supination (Nigg et al., 2019). Pronation is the inward rotation of the rear foot about the subtalar joint axis. Supination is the outward rotation of the rear foot about the subtalar joint axis.

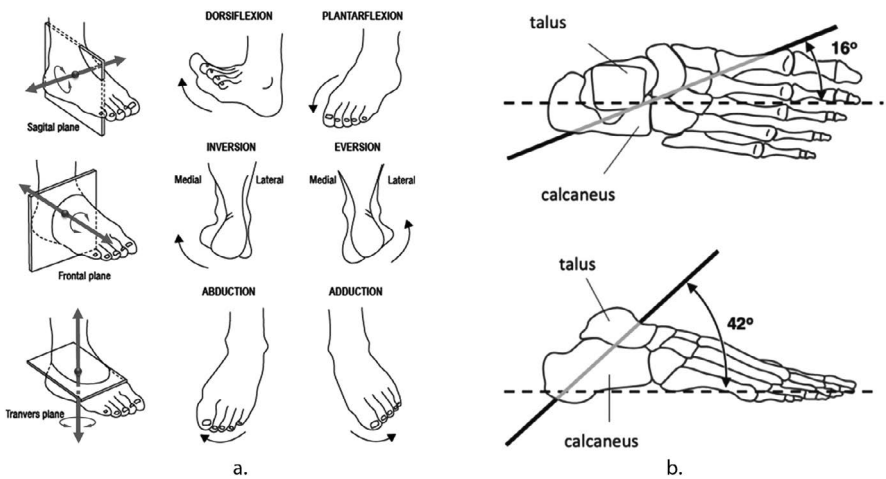


FIGURE 7.22 (a) Anatomical movements of the foot. (b) Subtalar joint axis.

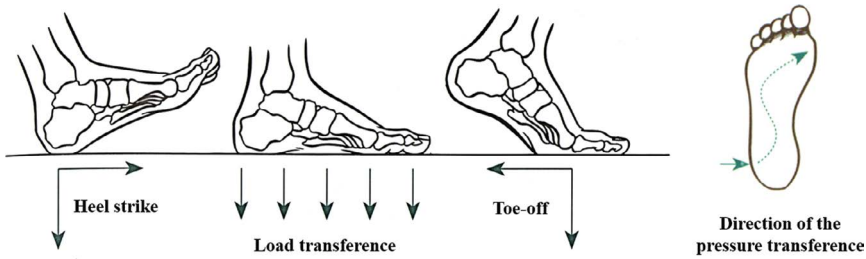


FIGURE 7.23 Stance phase of the gait cycle and pressure transference during this phase.

The Metatarsophalangeal Joints

The metatarsophalangeal joints (MTP joints) are the joints between the metatarsal bones of the foot and the proximal bones (proximal phalanges) of the toes. The main movement is flexion/extension playing a crucial role during gait but also ab/adduction in a short range.

The gait cycle consists of the stance phase and the swing phase with a total average duration of one gait cycle for men ranging from 0.98 to 1.07 s (Murray et al., 1964). The stance phase of gait begins when the foot first touches the ground and ends when the same foot leaves the ground, being approximately 62% of the total gait cycle (see Figure 7.23). The stance phase is divided into three periods: the initial contact period also known as heel strike (27%), the period of load transference (40%), and the propulsive period or toe off (33%) (Root, 1971).

Heel Strike Phase

The initial contact period begins with a heel strike. The function of the foot is to activate the natural ability of the body to absorb shocks as well as to ensure a stable position. The foot moves from heel to toe from a supinated position to a pronated position (foot rolling inward). Once the foot becomes flat, the forefoot comes in contact with the ground, and the next phase starts (see Figure 7.24). During walking the impact force produced during the heel strike is similar to the body weight. During running the impact force produced during the heel strike is almost two times the body weight. This situation requires extra protection in the design of the footwear.

Load Transference Phase

In the middle of the gait, the load transference phase, the functions of the foot are load support and overall stability. The internal foot arch undergoes the highest deformation, and the foot achieves the higher length. The foot plant is in full contact with the ground and this limb supports the entire body weight. The body weight moves forward over this fixed limb to prepare for the propulsion period. The load transference phase ends as the heel begins to rise off the ground (see Figure 7.25). Footwear design influences the dynamic pressure distribution pattern. Ideally, a plantar pressure map should avoid overpressure points that can be responsible for pain and foot problems (e.g., corns and calluses callus, metatarsalgia). The optimal design of an insole or footbed in terms of materials and anatomical shape is a good solution to enhance pressure distribution patterns.

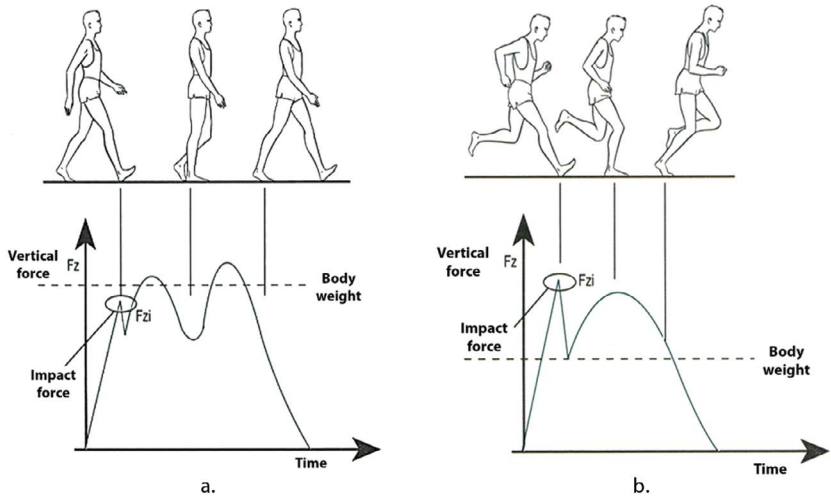


FIGURE 7.24 Vertical force during the stance phase measured using a force plate while walking (a) and running (b) (Ramiro et al., 1995).

Propulsion (or Toe Off) Phase

It is the final stage of the stance phase. It begins immediately as the heel lifts off the ground. The body is forward-moving. The heads of the metatarsals act as a support point for the rotation of the metatarsal joints. During propulsion, all the load is supported by the toes and the transversal arch achieves maximum deformation.

Foot biomechanics play an important role in the functional design of the shoe and influence the fit, particularly, the dynamic fit. That is why it is important to consider this knowledge to define fit requirements and specifications as well as design a good validation and fit test for each step of the design and sizing loop.

Sometimes, specifications related to the different functional aspects can be in conflict with each other, including the fit. Therefore, it is important to consider the

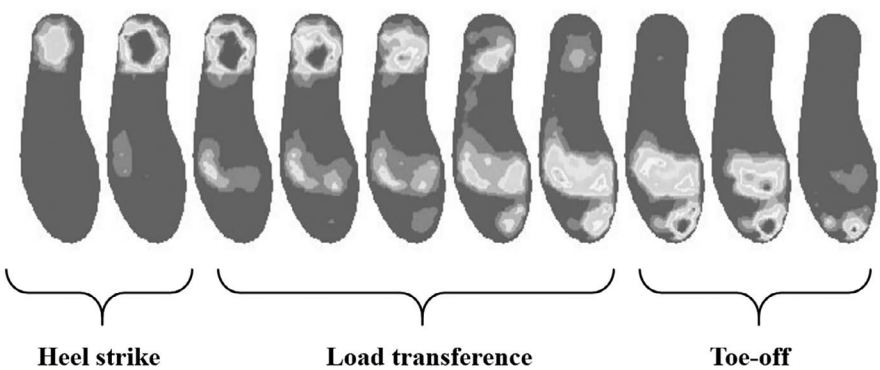


FIGURE 7.25 Normal sequence of plantar pressure distribution during the stance phase of gait.

type of footwear, the activity, and the TP to establish reasonable specifications and achieve an optimal trade-off between the various functional aspects of the footwear. For instance, a running shoe used for training should be comfortable and provide a good and healthy fit however, for a running shoe used in a competition a more aggressive fit can be considered to increase performance.

Foot Anthropometry Methods

There are very few publications and available databases of foot anthropometry. In addition, the available data usually is not comparable as different protocols and methodologies were used to measure the foot and at times have used different anthropometric definitions. As a result, the comparison of foot anthropometry from different studies is not always feasible.

Posture and Weight Bearing

Foot shape changes between different loading conditions in the static posture. These shape changes affect foot measurements (Kouchi et al., 2021) that can be larger than a shoe size grading interval. The recommended condition for the measuring protocol of foot anthropometry is half-weight bearing, the user stands erect distributing body weight equally on both feet on a flat surface. This is the most commonly used condition for footwear applications. It is recommended a separation of the feet in line with the hip joints to control the rotation and the deformation of the longitudinal arch. An alternative protocol also used especially in stores is a partial-weight bearing, sitting: the user sits on a chair resting most of his or her body weight on the chair and the foot lies on a flat surface. In this posture, it is more difficult to control the distribution of the weight that depends on the sitting position (angle of the hip and knee joints). Non-weight bearing (NWB), the foot is in the air when it is measured, not supporting any body weight. In this condition, foot pose is not standardized. It is used mainly in clinical applications when it is important to capture the shape of the arch unloaded. Since this condition is difficult for measuring the foot anthropometry and shape, the alternative method used for these applications is a foot impression taken using a foam block. This is valid to capture the foot plant in NWB condition.

Foot Measurements Should be Done Barefoot With No Socks

Socks produce pressure over the foot, modifying the foot shape and measurements in different values depending on the type of textile and fit provided by the sock. The thickness of the socks introduces an artifact in the resulting measurements. Additionally, anthropometric measurements rely on anatomical points that cannot be identified and marked with socks.

The shape and dimensions of the foot can be also affected by the time of the day and the activity done before taking the measurements. It is suggested to register this information in the measuring protocol especially when fit testing with subjects.

Manual Versus Digital Anthropometry

Foot anthropometry can be collected manually or using 3D scanning methods. Nowadays, the main footwear brands are using 3D scanning technology to capture

not only foot measurements but also foot shapes. Both methods are adequate to gather anthropometric data however, several studies show bias between digital and traditional methods therefore, it is not recommended to mix data or to compare foot anthropometry measured with different methods.

For both manual and digital methods, the foot measurements are defined using landmarks, which are key references that can be anatomical or geometric points of the foot surface. Some of the foot measurements are referred to as sections or planes that first require, the orientation of the foot and the definition of a foot axis, an imagined line indicating the longitudinal axis of the foot. However, the foot is not symmetric, this fact introduces some complexity and is derived from several definitions of the foot axis (Kouchi et al., 2021), two of which are the most commonly used (see Figure 7.26). The line connecting the projections on the ground plane of the backward point of the heel (pternion) and the tip of the second toe (ISO/TS 19408, 2015). The line connecting the projection on the ground plane of the backward point of the heel and the midpoint of the breadth of the ball cross-section. This alignment method is more stable since it is not affected by the large variability of the toe shape and some common deformities in this area (e.g., bunions).

In this chapter, the definition of the measurements has been done for digital anthropometry in order to be in accordance with the digital method to measure the

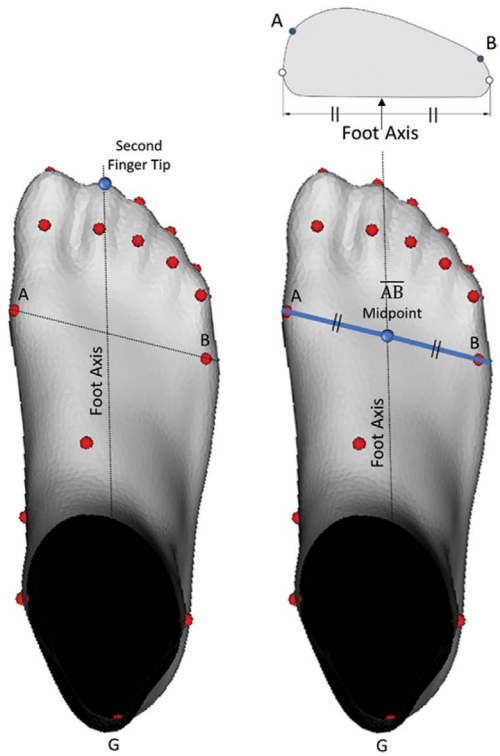
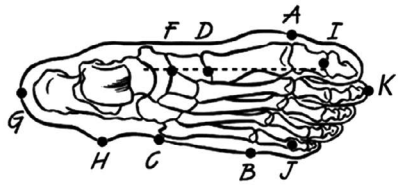
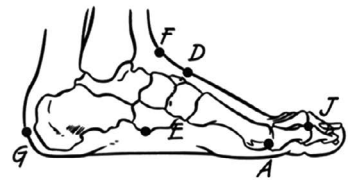


FIGURE 7.26 Two different definitions of the longitudinal foot axis (Kouchi et al., 2021).





A. 1st Metatarsal head.
B. 5th Metatarsal head.
C. Tuberosity (styloid) of the 5th metatarsal.
D. Highst point of the instep.
E. Lowst point of navicular tuberosity.
F. Junction point.
G. Pternion.
H. Lateral malleolus.
I. Highest point of the interphalangeal joint of the 1st toe.
J. Highest point of the distal interphalangeal joint of the 5th toe.
K. Tip of the longest toe.

FIGURE 7.27 Foot landmarks.

last. The definitions of the foot anthropometry described are those recommended by the authors. Note that other studies used similar measures with slight variations of the definitions. In the case of using existing anthropometric data of the foot, it is important to review the measurement definition in detail.

Foot anthropometry is based on anatomical landmarks (see Figure 7.27). For 3D foot scanning, physical markers such as stickers are used to identify the anatomical position of the bones detected by palpation. The position of these reference points and their variability across the population segmented by size is crucial to determine the location (position and angle) of the reference sections of the foot and the last.

The main foot anthropometric measurements described below are related to foot sections calculated from the 3D foot scan and they have a corresponding definition in the last or the footwear to establish the transference and contribution to design elements of the shoe (see Figure 7.28). For each section, the main anthropometric dimensions (girths, widths, and heights) can be obtained. An example

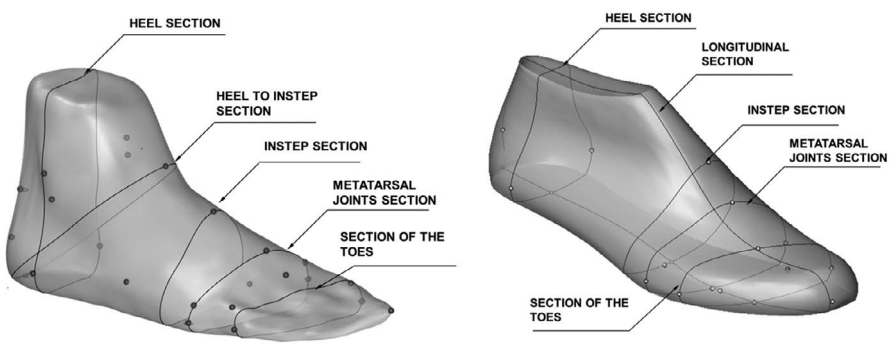


FIGURE 7.28 Anatomical sections of the foot and the equivalent sections in the last.

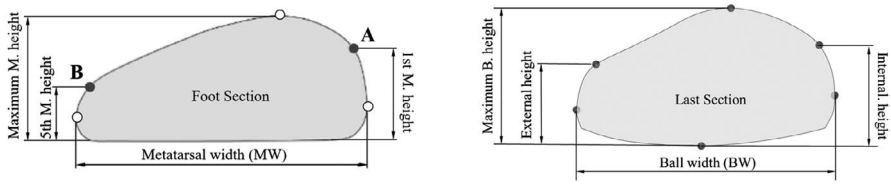


FIGURE 7.29 Example of equivalent measurements for the metatarsal joints section of the foot (name of the section of the foot based on anatomy) and the ball section of the last (equivalent name used in the footwear industry for the forefoot section in the last).

of the main dimensions for the metatarsal joint sections of the foot is illustrated in Figure 7.29.

Note that the naming used in foot anthropometry changes in some sections of the last. To properly transfer the foot anthropometry to the last, it is important to consider the following aspects (see Figure 7.30):

- Foot anthropometry is measured with the flat foot in standing position while the last includes a heel height resulting in a curved shank shape. This shape should be considered to locate properly the position of the sections.
- The section of the foot is not perpendicular to the longitudinal axis. It has an angle due to different lengths of the first and fifth metatarsal head positions.

An important issue before measuring is the last orientation in relation to a coordinate reference system and the measuring error associated with variations in the reference system. For example, the heel height measurement is influenced by the rotation of the last and how well it is defined. A small difference in the rotation can create a large difference in the heel height (see Figure 7.31). The asymmetry of the last and the lack of anatomical and reference points introduced certain difficulties in defining

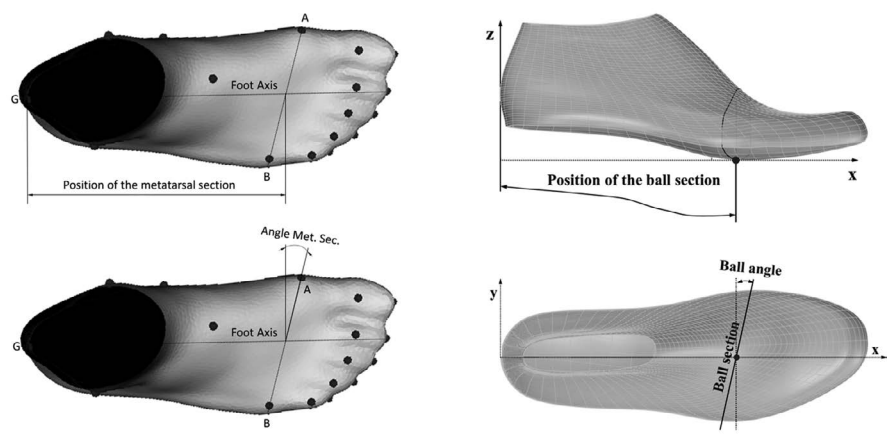


FIGURE 7.30 Equivalent position of the metatarsal foot section and the last ball section.

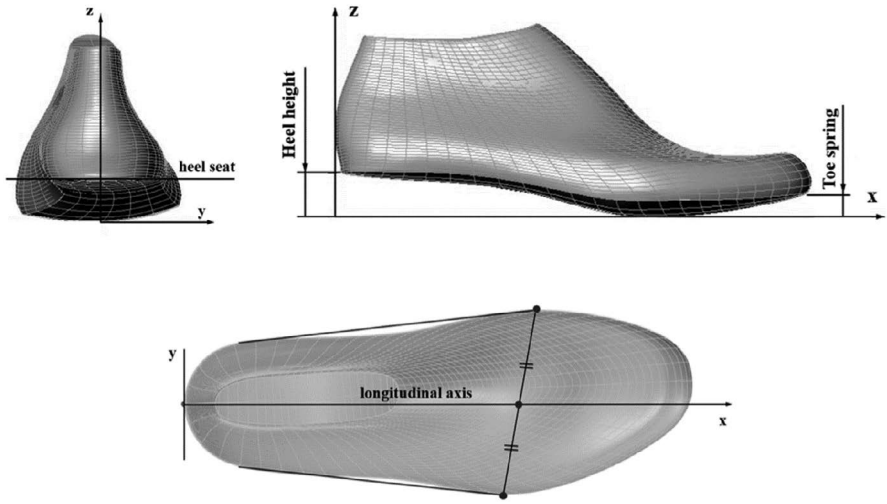


FIGURE 7.31 Orientation of the shoe last to locate the functional sections and calculate the dimensions.

the last axis and the reason why there are a variety of methods. It is important to define a consistent protocol for the computation of the last measurements.

Variability of the of the Foot Anthropometry: Sex, Age, Demographics

Foot anthropometry varies with socio-demographic factors such as gender, age, or geography. It is important to define the target market and use anthropometric information of the corresponding population. Unfortunately, the current state of anthropometric information on the body is partial, fragmented and very limited, the state of foot anthropometric databases is even worse. In this situation, most of the footwear companies, especially in the case of hi-tech footwear, create their own foot anthropometric databases.

Foot anthropometry shows relevant differences by gender. An analysis done with a foot anthropometric database of 783 subjects measured with a 3D scanner showed mean differences of 28 mm in length (see Figure 7.32), 25 mm in the metatarsal girth (ball girth), and 10 mm in the foot width (see Figure 7.33).

In addition, for the same shoe size, the width of women's feet in is smaller than in the width of men (see Figure 7.34). They have a foot that is 10–12 mm thinner, which is a full size in width. This information should be considered for the design of unisex shoes. A common mistake is to design the shoe last for men and expand the grading for both range of sizes. The result is footwear that is too wide for women. This is especially true in the case of safety footwear due to the high cost of the lasts and molds for production. Companies try to avoid creating sizes for women with different widths but similar lengths to men's footwear. A sizing loop fit audit can reveal these limitations and help to devise cost effective alternatives.

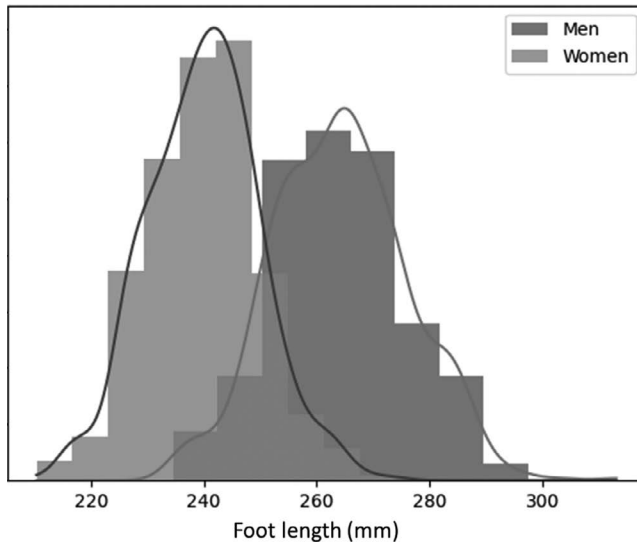


FIGURE 7.32 Foot length distribution of men and women.

Anthropometry with High Heel Shoes

Almost all shoes have a certain height in the heel support. Even in the case of men's footwear, it is commonly a 1–2 cm heel height. In sports footwear, the difference between the thickness of the sole in the heel and forefoot area is called "drop" and it is an important parameter of the footwear performance. The extreme case is the high-heeled shoes for women that change completely the foot anthropometry and dynamics.

Foot anthropometry is gathered in a flat position, weight bearing or half weight bearing. These are the conditions followed in foot anthropometric studies and research publications. However, the foot and shape dimensions change

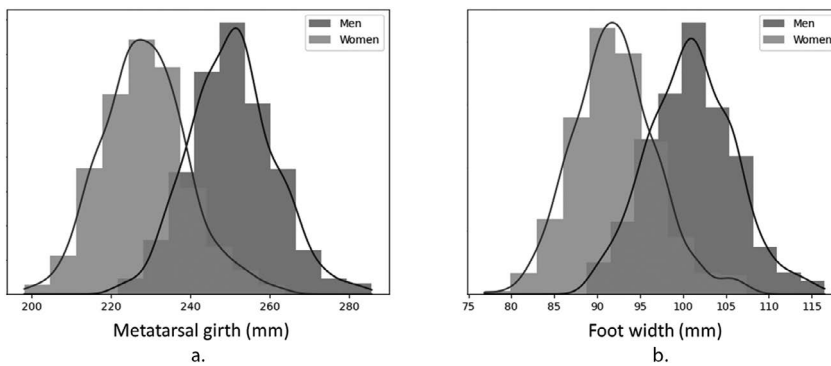


FIGURE 7.33 (a) Metatarsal girth and (b) foot width distribution of men and women.

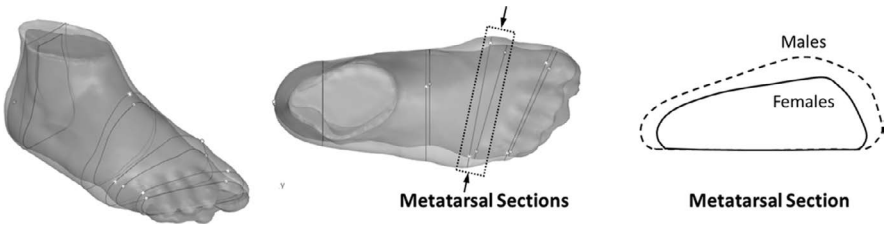


FIGURE 7.34 Differences in foot proportion between women and men.

substantially when a heel height is introduced. Figure 7.35 shows the metatarsal section of the same foot scanned in four conditions: flat, 3 cm, 5 cm, and 7 cm of heel height. It can be appreciated the difference between the section in flat position and the three sections with a heel height which became narrower (~5 mm) and higher (~8 mm).

The metatarsal girth decreases by almost 4 mm with an increase in heel height. Regarding the shape variation between flat and high heel conditions, it is interesting to point out that sections obtained in high heel conditions are more like the section of the last (see Figure 7.35). Only due to the fact that the anthropometric dimensions of the foot are not measured in the same posture adopted by the foot inside the shoe, it is necessary to consider and to introduce corrections if we try to transfer the foot anthropometry to the shoe last dimensions. However, there is a lack of publications reporting anthropometric measurements for different heel heights, sizes, and foot types. In addition, shoes can be designed in any heel height so it might be necessary to interpolate values to obtain information applicable to the last design.

The Effect of the Upper Material and the Shoe Fit

The upper material plays an important role in the shoe fit. On the one hand, the rigidity of the upper material prevents or enables the deformation of the shoe thereby

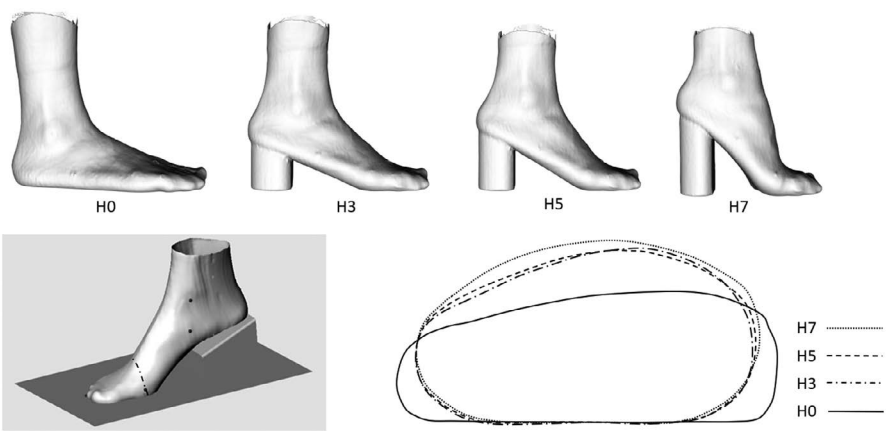


FIGURE 7.35 Metatarsal section of the same foot scanned in 4 conditions: flat, 3 cm, 5 cm, and 7 cm heel heights.

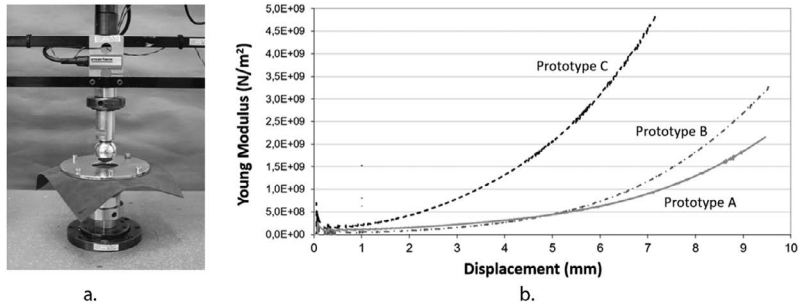


FIGURE 7.36 (a) Mechanical test to determine the rigidity of the upper material. (b) Force – deformation curves of three different upper materials made of leather.

influencing the fit. Some upper materials can be deformed, temporal, definitive or a combination of both, to enable the accommodation of different foot shapes and dimensions. On the other hand, the structure and design of the upper material are the cause of most of the critical fit problems. It is often not possible to wear a shoe due to high pressures of the upper edges or blisters caused by friction and a slide effect between the foot and the shoe.

The rigidity of the upper material can be characterized with a mechanical test using a universal strength machine. There is no standard test to "simulate" the loads involved in the foot-upper interaction. The main area of pressure and deformation of the upper that requires accommodation is the metatarsal section. A mechanical test performed with a universal testing machine can be used to simulate the pressure of the metatarsal bone against the upper to record the force-displacement curve and to obtain the upper material rigidity (see Figure 7.36).

The upper material rigidity is related to the pressure produced by the shoe over the foot dorsum. A pressure sensor matrix was developed by the IBV and presented by Olaso et al. (2007). It was used in an experiment to determine whether the rigidity of the upper material affected the pressure distribution over the foot dorsum as shown in Figure 7.37.

Two shoe prototypes were manufactured with leather upper materials of different rigidities to assess the device as shown in Figure 7.38. The shoes had a minimal design just to avoid the influence of other components on the foot pressures. A single layer of the upper tested covers the foot dorsum. A strip enables the support in the

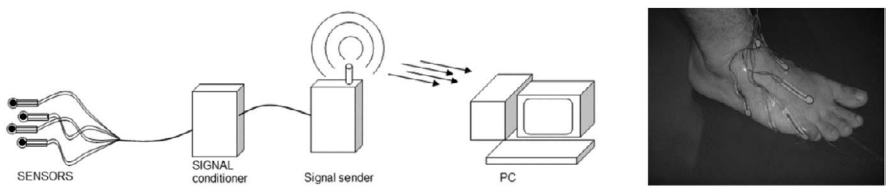


FIGURE 7.37 Electronic device for measuring pressure curves within the shoe during walking.

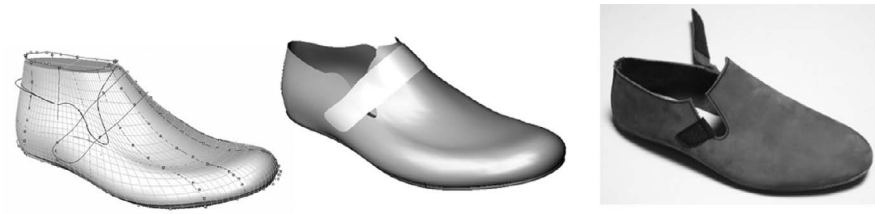


FIGURE 7.38 Shoe prototypes for measuring mechanical properties of the upper material.

high instep area. The sole was very thin and flexible to enable a normal walking pattern.

Some of the results observed with this experiment included the pressure pattern during walking and pressure patterns for different upper material rigidity.

Regarding the pressure pattern during walking (see Figure 7.39):

- The walking pattern of the pressure sensors shows the highest pressure at the insertion of the Achilles tendon in the heel contact.
- The most prominent point of the outer heel shows a point of pressure also in the heel contact.
- The most lateral point of the first toe shows the point of pressure at the late stance phase.

Results of the pressure pattern have been obtained for a specific shoe upper design: wide fit to the last, upper covering all the foot dorsum. It is expected to obtain different pressure patterns for other upper designs and shoe styles: pointed

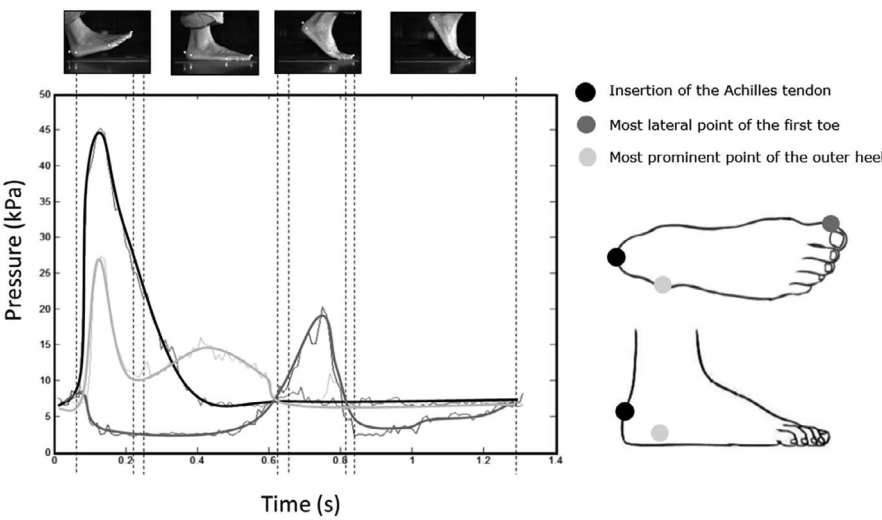


FIGURE 7.39 Pressure pattern of three points of the foot obtained during walking.

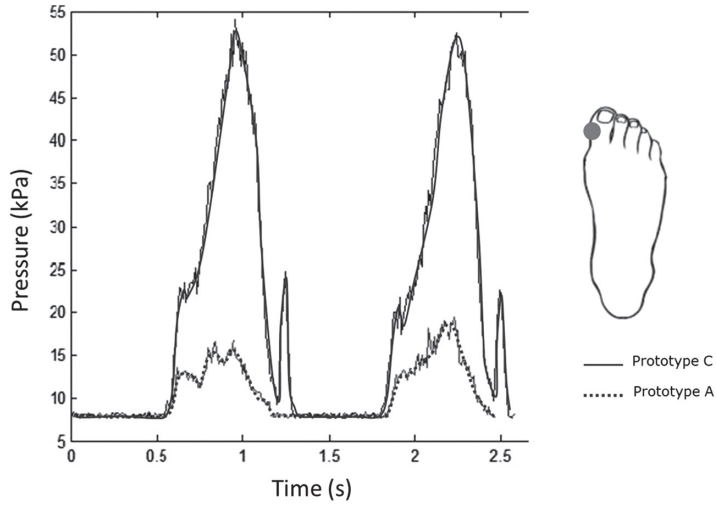


FIGURE 7.40 Pressure of the most lateral point of the first toe obtained with the prototypes of upper material rigidity A and C.

shoes, upper with opened instep such as ballerinas or pumps, high heel shoes, and boots.

Regarding the differences between the upper material rigidity, Figure 7.40 shows a significant difference in pressures between the two prototypes with the two extreme upper material rigidity. The lower pressures correspond to the upper material with lower rigidity.

The use of pressure sensors as a methodology to measure fit by an objective test however, required more knowledge to determine which are the pressure thresholds to identify fit problems in different foot zones. These thresholds can also differ among population groups due to more or less sensitivity to overpressure or caused by cultural aspects. The personal preference and experience of wearing shoes is very relevant for the perception of fit and ultimate shoe selection.

RESOURCES, PLANNING, AND PREPARATION

The preparation of the protocol for the fit test requires the definition of the following aspects:

Location

Fit testing implies the design of the protocol to test and assess the fit of a specific footwear. The fit test can be done in a laboratory however, for some specific footwear types of footwear such as sportswear, it can be interesting to do the fit test outdoors. For laboratory-based tests, it is important to have enough space for the subject, the technician, the storage of the samples, the measuring devices (e.g., foot scanner) and a free space to move, walk or perform different types of movements. Some examples are shown in Figure 7.41.

FIGURE 7.41 Preparation of fit test in laboratory conditions.

Subject Recruitment

The number of subjects will depend on the type and objectives of the test (e.g., Mini-test, testing the base size, testing the full range of sizes) and if the target users are only males, only females or both genders. Regarding the subject recruitment, it is mandatory to select users who are experts in using the type of footwear. For instance, in the case of high heel shoes, the women participating in the fit tests should be regular users of high heel shoes in order to perform the walking pattern according to this type of shoe and to give feedback in relation to the experience using this type of footwear.

Subject pathologies should also be considered. Subjects may have a variety of common foot disorders, such as bunions, calluses, blisters, plantar fasciitis, heel spur, hammer toes, or metatarsalgia. There should be a plan for recording and dealing with these pathologies. Even if the subjects with certain pathologies are to be rejected it can be important to record that the subject was randomly selected but rejected for a particular anomaly.

It is also important to check the subject's gait. Subjects with an asymmetry in the back or hips may have an abnormal walking or running pattern, that will need to be identified. The fit perception can be influenced by these pathologies.

Time of the Day or Daily Schedule Range

The foot shape varies during the day and after doing an intensive activity. It is expected that some fit variation can be in relation to the time of the day, weather conditions as well as the activity performed. The effect of these factors is usually negligible however, it is recommended to record these data in the questionnaire.

Treatment Sampling

If we are testing several prototypes, shoe models or sizes (the treatments) in the same session of a footwear fit test it is recommended to randomly assign the treatments. In other words, we need to define which treatments to test on each subject and which order to present them.

If the aim of the fit test is to establish the anthropometric ranges of accommodation of a footwear model then, it is important to establish a criterion to determine if the testing shoe size is the correct size for the subject or if he or she better fits in a larger or smaller size. For this purpose, it is useful to use samples of the shoe for a size up and a size down. As a result, it will be possible to establish the anthropometric thresholds with good and bad fit within a specific size and the overlaps with the neighbor shoe sizes.

Testing Conditions

For footwear testing, the socks to be worn should be specified as well as any environmental conditions that must be controlled. To normalize the effect of the sock thickness and material it is recommended to provide new socks to each subject in a proper size. An alternative is to have the subjects arrive in their own socks that they would normally wear with that type of shoe and record the type of sock and its relative thickness.

In some specific cases, temperature and humidity should be controlled. Such as when the tests are performed in seasons with a high difference in environmental conditions (e.g., sandals being tested in winter) or in the case of footwear used in specific weather conditions (e.g., high mountain boots).

Testing Protocol

The testing protocol includes the flow of the test.

- Subject reception. Welcome and guidance to the fit laboratory. Explanation of the test and signature of the subject's consent.
- Test:
  - Foot exploration: application of rejection criteria
  - Foot scanning or anthropometry measurements
  - Put on the shoes and check the fastening done by the subject
  - Fit test of all the shoe samples. Fulfillment of the questionnaire
- Subject farewell: Reward of the subject. Walk back to the hall. Goodbye

In addition to fit testing facilities, footwear innovation during the development process requires 3D design and prototyping. Prototyping footwear or single components requires a certain infrastructure of manufacturing. Using 3D printing can be a feasible option to manufacture some of the components such as the sole or insole but not for the upper material. In particular, if the aim of the tests is to validate fit, it is important to prototype the footwear with the real upper material. Consequently, having access to footwear manufacturing facilities becomes imperative, even for companies

outsourcing production to third-party countries. Footwear prototyping demands collaboration with various suppliers across the footwear manufacturing chain. Executing this process for only a limited number of samples can be costly but is becoming less so with advances in CAD and CAM software tools. Most of the CAD software used is specific for footwear (e.g., Shoemaster, Crispin from Delcam). Nevertheless, the use of generic CAD software can be also an alternative. In particular, Rhinoceros® is an option that is growing in popularity. The book "Rhinoceros Advanced Training Series Shoe Design and Visualization" (McNeel, 2005) is an introduction to CAD design of footwear with Rhinoceros. It includes tutorials to design different components of the shoe using sketches as a starting point, 3D design of the whole shoe, finishing and rendering.

CASE STUDIES

Case Study 1: The Design Loop of Casual and Fashion Footwear

This case study aims to demonstrate the applicability and benefits of the SPEED process for companies that develop large collections of new shoe models each season. In this context, the process use of the SPEED process must be agile and easy to implement. This requires the preparation and set up of protocols that enable the optimal performance of fit tests and the deployment of the results along the value chain.

This is an example of using the SPEED process for a casual footwear product (i.e., a sneaker). It describes the use of the COF for a footwear product, a sample from the TP to establish the base size and sizing system, and the use of fit testing to establish specifications, last proportioning, and create a footwear fit standard with a standard fit validation process.

The TP for this product was both men and women between the ages of 18 and 60 years from the United States and Europe. The fit requirements for this product were: (1) 80% of the population can slightly move their toes inside the shoe, (2) 80% prefer the fit "as is" in all areas with no pain points, and (3) the clearance in length is 5–15 mm. These requirements were assessed using an agreed upon COF and metrics as shown in Table 7.4.

The dataset used to characterize the TP was from the raw data from the CAESAR survey (Robinette et al., 2002). This data included more than 4000 adult men and women from North America and Europe. It contained two key measurements of the foot: total foot length (Foot Length) and metatarsal width (Foot Breadth). Foot Length measurements were divided into hypothetical shoe size categories using the European shoe size category definitions which was in accordance with the ISO 19407:2023 standard. The hypothetical size 38 for women had a foot length of 243.3 ± 5 mm, and the size 42 for men had a foot length of 280 ± 5 mm. The foot length range in this case was 10 mm more than the 6.67 mm which is the theoretical size grading for EU sizing. We took this decision due to the overlaps in foot lengths between sizes related to personal preferences or the impact of their foot width. A user with narrow feet for instance, might select a smaller size in length to have a better shoe fit in the width and conversely, a user with wider feet might select a bigger size (Valero et al., 2023).

TABLE 7.4
COF for the Sneaker

Fit Area	COF	Metric
Length	Optimal fit, neither tight nor loose.	Subject questions: *How would you rate the length of the shoe?* *–1 slightly tight, 0 ok, +1 slightly loose* *What would be your preference for the length?* *–1 shorter, 0 as is, +1 longer* *Are there any pain points? Yes or no.* Pass = they prefer "as is" and there are no pain points
Toes	**General fit:** Optimal fit, neither tight nor loose. **Width:** The toes are slightly compressed inside the shoe compared to barefoot. Then, the last width at the toes will be slightly smaller than toes width anthropometry. **Height:** It is important to avoid compression in the top of the toes to avoid blisters. The height of the last at the toes' section should be higher that the foot anthropometry.	Expert assessment: *Is there clearance in the toes? Yes or no.* Subject questions: *Can you move your toes? Yes or no.* *How would you rate the toe area width?* *–1 slightly tight, 0 ok, +1 slightly loose* *How would you rate the toe area height?* *–1 slightly tight, 0 ok, +1 slightly loose* *What would be your preference for the toe width?* *–1 tighter, 0 as is, +1 looser* *What would be your preference for the toe height?* *–1 tighter, 0 as is, +1 looser* *Are there any pain points? Yes or no.* Pass = they prefer "as is" and there are no pain points
Metatarsal Joints (ball of the last)	**General fit:** Optimal fit, neither tight nor loose. **Width:** Optimal fit, neither tight nor loose.	Subject questions: *How would you rate the ball of the foot area?* *–1 slightly tight, 0 ok, +1 slightly loose* *What would be your preference for the ball of foot area?* *–1 tighter, 0 as is, +1 looser* *Are there any pain points? Yes or no.* Pass = they prefer "as is" and there are no pain points
Instep	**General fit:** Optimal fit, neither tight nor loose. A slightly wide fit can be also allowed because the fastener is large and can cover until the high instep (close to the leg). **Height:** Optimal fit, neither tight nor loose.	Expert assessment: *Is there clearance in the instep area? Yes or no.* Subject questions: *How would you rate the instep area?* *–1 slightly tight, 0 ok, +1 slightly loose* *What would be your preference for the instep?* *–1 tighter, 0 as is, +1 looser* *Are there any pain points? Yes or no.* Pass = they prefer "as is" and there are no pain points

(Continued)

TABLE 7.4 ***(Continued)***
COF for the Sneaker

Fit Area	COF	Metric
Heel	**Width:** Optimal fit, neither tight nor loose. A slightly wide fit can be also allowed. **Rear heel:** Optimal fit, neither tight nor loose. A slightly wide fit can be also allowed because heel to instep fit is assure with the design of the upper material.	Subject questions: *How would you rate the heel area?* *–1 slightly tight, 0 ok, +1 slightly loose* *What would be your preference for the heel?* *–1 tighter, 0 as is, +1 looser* *Are there any pain points? Yes or no.* Pass = they prefer "as is" and there are no pain points

Figure 7.42 shows the percentage of the TP for each foot length size for men and women. Here, we see that the mode for women was size 38, and the mode for men was size 42. These then were our base sizes. If we included only sizes that have at least 5% of the population in the size, we had a size range of 35–41 for women and 39–45 for men. These size ranges accommodated 90.6% of the female TP and 93.6% of the male TP.

In this example, there were no prior reference values for last measurements versus foot measurements, and no prior validated footwear fit standard. In other words, a last labeled 38 might not fit the size 38 woman's foot. Therefore, we needed to create (or select) a suitable starting last. To evaluate the last proportioning, we examined additional foot anthropometry. Having determined the last for the base size,

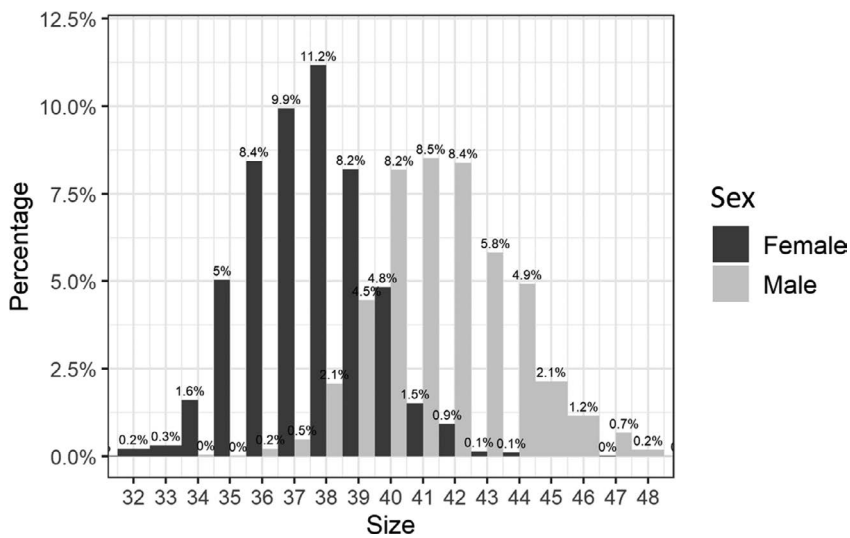


FIGURE 7.42 Distribution of population across shoe sizes based on theoretical intervals (ISO19407:2023).

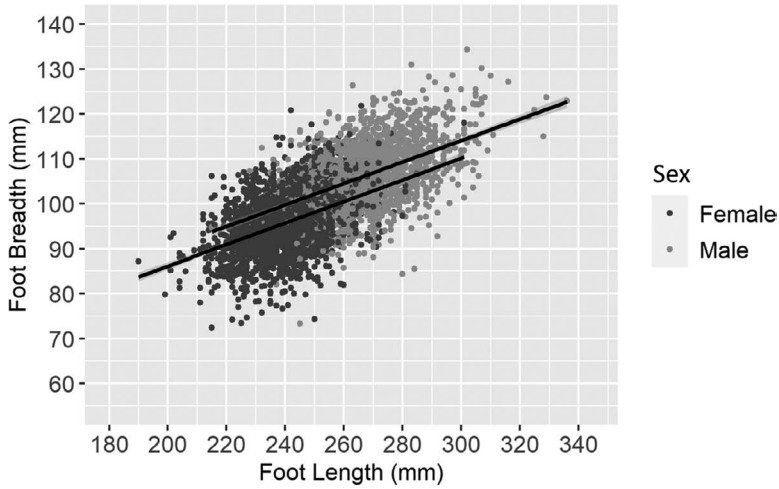


FIGURE 7.43 Bivariant plot foot length – foot width (mm) of the CAESAR database.

we created sneaker prototypes and refined the last with outcomes from the fit tests. Once we had a validated last, the anthropometry and fit scores for shoe fit with this last, we drafted a fit standard to use for both this and other styles. So, while we had several iterations to achieve our required fit this time, we would need only a few trade studies for future styles or shoe concepts.

Figure 7.43 shows the bivariate scatterplot of Foot Length by Foot Breadth with the best fit regression lines for each gender shown. These lines represent the most likely Foot Breadth values for each Foot Length value, that is the correlation between both variables. The line for females was approximately 5 mm smaller for Foot Breadth than the line for males. That means the female shoe size range should be proportioned to be narrower for each Foot Length than the male shoe size range. In other words, female feet are not simply scaled down male feet. It was important to examine all the important foot dimensions in this manner to get a good shoe and last proportioning.

Previous studies have indicated there are several other important measurements of the foot for this type of footwear including ball girth, instep girth, instep height, toe height, toe girth, and heel width. This knowledge was not available for the entire TP when we began the study, so data collection was done on small samples ($n = 100$) of men and women who were within the base size for Foot Length, and the full set of foot measurements was included in subsequent prototype fit tests once the base size prototypes were available. This allowed us to establish footwear specifications in all these areas as well as to begin building a combined foot and fit database for our products. The fit database was both structured and traceable enabling the generation of design and assessment standards to arrive at effective specifications and meet requirements faster for new products.

The additional anthropometric data obtained before fit testing gave us estimates of the range of variability for the anthropometric measurements within the base sizes. This was important to understand the accommodation range that must fit within each

TABLE 7.5
Statistical Summary of the Anthropometric Dimensions for Women's Size 38

Women's Size 38 (Foot Length 243.3 ± 5 mm)	P05	P25	P50	P75	P95	Mean	SD
Ball Girth (mm)	217.8	223.4	230.4	236.2	245.2	230.5	8.7
Foot Breadth (mm)	90.3	93.3	95.9	98.7	101.6	96.2	4.0
Instep Girth (mm)	217.5	223.8	229.4	233.8	242.7	229.3	7.8
Instep Height (mm)	57.8	60.8	63.4	66.2	68.8	63.4	3.5
Toes Height (mm)	20.3	21.9	23.6	25.2	28.8	23.7	2.5
Toes Girth (mm)	197.4	206.1	213.5	218.5	228.2	212.7	9.2
Toes Width (mm)	87.8	91.5	95.5	98.2	102.3	95.1	4.6
Heel Width (mm)	58.7	61.3	64.0	66.0	68.9	63.7	3.3

size. Table 7.5 shows the size variability for women who had foot lengths in the size 38 range (243.3 ± 5 mm).

This table was used as a reference but not transposed directly in the design or the last. As we explained in Chapter 3, percentiles are calculated independently for each measurement and combining them is not recommended. Designing for the mean or 50th percentile for all measurements might not fit anyone. Also, in some areas, we may want to use the largest measurement, in others something in the middle might be preferable.

That said, many insights can be obtained from the anthropometric table. First, the ball girth variation from 5th to 95th percentile is 27.4 mm. According to the industrial grading standard of EU sizing, the usual ball girth step between sizes is 5 mm. That means that the foot variation that must be accommodated within a size is more than 5 times the size grade.

A shoe last with a wide ball girth and deformable upper material, and a soft and deformable insole can achieve a higher accommodation rate. Such as in the case of a sneaker. A shoe last with a tight ball girth to provide a slim style and is made of rigid upper material will have a lower accommodation rate and will require several width sizes. We could only determine how much is accommodated with the different materials using trade study type fit tests. Then we could determine which materials worked best and if we needed or wanted to produce multiple width sizes in addition to the length sizes.

In the same way, the variation in Foot breadth from 5th to 95th percentile was around 6 mm while the usual industrial grading standard of EU sizing is 2 mm. According to the requirements of the COF for this example, the ball width, it was important to consider that the fit of the shoe always compresses the metatarsal arch of the foot to hold the foot in place to avoid slippage between the foot and the shoe. That means that the ball width of the last is going to be smaller than the equivalent metatarsal width of the foot. This situation is not only applicable to this sneaker case study example. Ball width of the last almost always will be smaller than the metatarsal width of the foot. How much? It depends on the type of footwear, materials, etc. This is the reason why it is necessary to start from an initial value of ball width of the

last and validate it with fit trials. Once a fit database is generated with a collection of historical fit data it can be used in the future as a valid reference for different types of shoes to create better lasts faster. A valid reference value will reduce the design loop by leading to a successful fit test on the first attempt.

The toes can be slightly compressed within certain limits and still be comfortable because they are mobile anatomical structures. Therefore, the sneaker's requirement for the last dimensions is to design a last with a smaller toe width than the foot anthropometry. Again, this is also applicable for most footwear styles. The specific values of the difference between the toe width of the last and the toe width of the foot will depend on many factors and it is necessary to start with an initial reference value and test the fit with live subjects. In the sneaker case, the design of the toes shape of the last is usually wide and the reference values will not differ much from the anthropometric values. Other types of shoes will require narrower toe areas, such as fashion shoe styles with slim and pointed toes. In this extreme, when the toes are highly compressed, the height of the toes may increase to compensate for this change, and the criteria of toe height might require a correction.

For toe height, pressure on the top of the toes should be always avoided. In fact, in our sneaker COF case study there was just enough room to be able to move the toes. For our sneakers, a starting reference point for a lower limit of the toes height might be the 95th percentile (28.8 mm) for the size 38. Is this enough? We could confirm this with trade studies of the materials and prototype fit tests of the prototype base size.

In the heel area, the fit in this area should be neutral (neither too wide nor too narrow). We used the 50th percentile (64 mm) as our starting reference point for our base size prototype.

Using the anthropometric data and our prior knowledge of what constitutes an acceptable sneaker fit we developed the specifications for the prototype lasts for our two base sizes. Then we created prototype sneakers in the base sizes to be used in prototype fit tests to verify and/or adjust the lasts. The specifications for the lasts included last measurements and production tolerances for the important measurements. The format for the specifications is shown in Table 7.6. Due to confidentiality

TABLE 7.6
Format of a Reference Value Table of the Last Dimensions

Section	Shoe Last Measurement	Reference Values (mm)
Length	Total length	L (size 38) = 253.5 ± 1
Ball	Ball girth	BG ± tolerance
	Ball width	BW ± tolerance
Instep	Instep girth	IG ± tolerance
	Instep height	>IH
Toes	Toes girth	TG ± tolerance
	Toes height	>28.8 mm
Heel	Heel width	64 mm ± tolerance
Heel-to-Instep (for boots)	Heel to instep girth	HtoI ± tolerance

reasons, all the specific reference values cannot be provided. To illustrate some reference values, we included the values for Length, Toes, and Heel.

Tolerance refers to the range of variation around the reference value within which a noticeable change is not perceived by the subject. For instance, a deformable upper material can offer significant tolerance in certain areas, allowing for a wider range of aesthetic styles, whereas a rigid material has a smaller tolerance. A lacing fastener provides a broad accommodation range, allowing for increased tolerance in both instep girth and instep height. Therefore, the tolerance values can vary for different designs, materials, or shoe last dimensions. These first tolerances were educated guesses just to get us started. The final tolerances were established through trade studies of different material properties using human subjects. As we gather databases of last trade studies and tolerances, we can establish last standards for different product types that enable us to get to a good last size and shape more rapidly.

Creating a shoe last design from scratch is not usual. It is more common to use an existing last from a footwear company or last maker which was the approach used for this case study. We used the style and aesthetic of the proposed design to identify a suitable last that met the length and girth specifications of the base size. One of the reasons was because these were the only measurements they could check at that time manually with a tape. Now we have 3D scanning and CAD tools that allow us to capture additional proportions and measurements.

The fit trials told us how to refine it for our specific product. After we had a refined last with the anthropometry and fit scores associated with it, we created a fit standard to use for all future products.

As a result, until we have a good correlation between the last and the anthropometry of the feet that are accommodated, our first last will be an educated guess that will probably require modification. Otherwise, we will have a shoe that fits the last but not the desired segment of the TP. We gathered the necessary information using fit testing and established an anthropometry, fit and last database to be able to determine this correlation.

Several prototypes were produced during the development process. The first mock-up of the sneaker was used as a first evaluation of the design, the fit and the shoe last. For this step, it was not necessary to use the real sole. A sole with the same heel height and thickness was used. This was a faster and cheaper way to check the fit with the last than using the final materials. It ensured we were in the ballpark before we made more expensive prototypes. This quick fit test was done with only the right foot. It was the beginning of the development process and there were many aspects to check and refine so prototyping both feet would be unproductive. After this first fit test, there were several modifications to carry out in the last and the substantial modifications were again prototyped and tested. Depending on how far the design is from the optimal performance, the process of achieving a successful fit test might require several loops of the redesign-prototyping-fit test. Doing these with faster, with cheaper materials saved us time and money.

The final prototype that was fit tested was a definitive version of the shoe manufactured with the final components. At this point in the development process, there is often a lot of pressure to launch the manufacturing of the complete range of sizes, so this validation is often not done. However, if not done there was a risk that some

critical problem would not be detected when the final components were not used in the fit test, which could have led to significant time and money loss. All testing is done to minimize risk, so the decision to skip this step would be dependent upon the confidence we have that the product is ready. This was the case of the sneaker development but also applies to the development of fashion footwear styles.

A fit test with subjects aims to correlate fit performance with anthropometry. It is important to represent the TP well, but also to reduce the number of subjects as much as possible. Usually, the anthropometric database that is used to characterize the TP is anonymous, so it is not possible to pull subjects from that sample to use for the test and that was the situation for this case study. We used a panel of new subjects from the TP who matched the Foot Length measurements of the base size. It was necessary to perform a new search to find good representatives. For this search, we also considered personal fit preferences and the skills to communicate the fit problems of the shoe, the good points, and the elements to improve their fit/comfort. Experience of wearing shoes of this type for the type of activity it was designed was one of our subject selection criteria. It was important to find subjects with good sensitivity that could be trained to assess the shoe effectively in a fit trial.

We preliminary chose seven new subjects for the fit test from a database of 100 women who were within the size 38 Foot Length range. Figure 7.44 shows their anthropometric profiles. Three of them have measurements that are mostly between 25th and 75th percentiles (the shaded area of the chart), representing the middle. Two of them had mostly small values representing a narrow foot and another two had measurements that were mostly high percentiles of the anthropometric values representing wide feet. Note that none of them exhibited values closely clustered around the median (50th percentile). This observation confirms the nonexistence of an average foot, revealing that the concept of a mean foot is merely an artificial construct

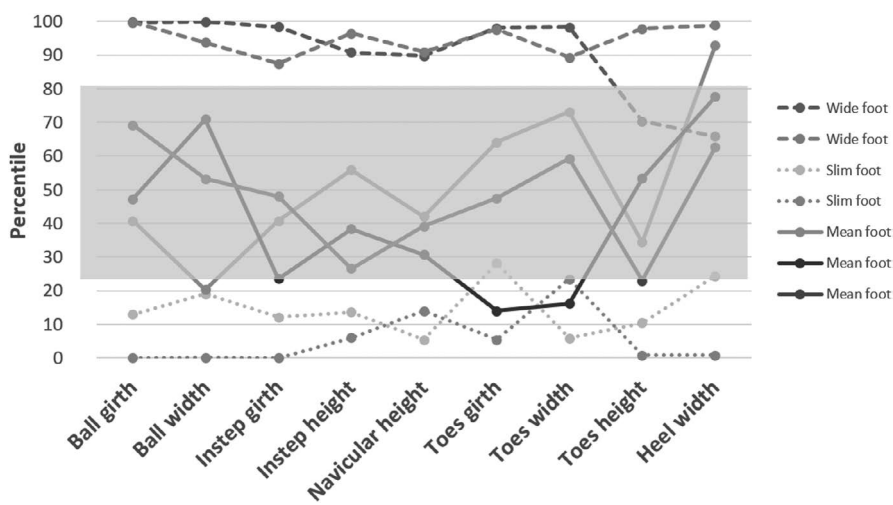


FIGURE 7.44 Anthropometric profile of a panel of subjects of a fit test.

within the statistical model. This is why we used three subjects who are somewhat in the middle to make sure the middle is well represented. More than one because one can be unusual and not representative.

The initial fit tests carried out with the first prototypes address large design or fit issues that might occur and substantial design changes that might be needed for optimal dimensions and these are detectable with as few as 3–5 subjects. In this case study, several quick trade studies were done selecting five subjects of the test panel seven subjects. The results of two of them were not consistent therefore, it was deemed optimal test to use five subjects who represented specific morphotypes of the foot anthropometry and were consistent with the subjective fit assessment of the shoes. Then prototypes with the final materials for both feet were created to validate and modify the base size last to ensure it was hitting the correct size for the TP. A summary of the fit test results for the five subjects is shown in Figure 7.45. This report is a short one-page summary of the main information gathered in the fit test. The first section is the general information:

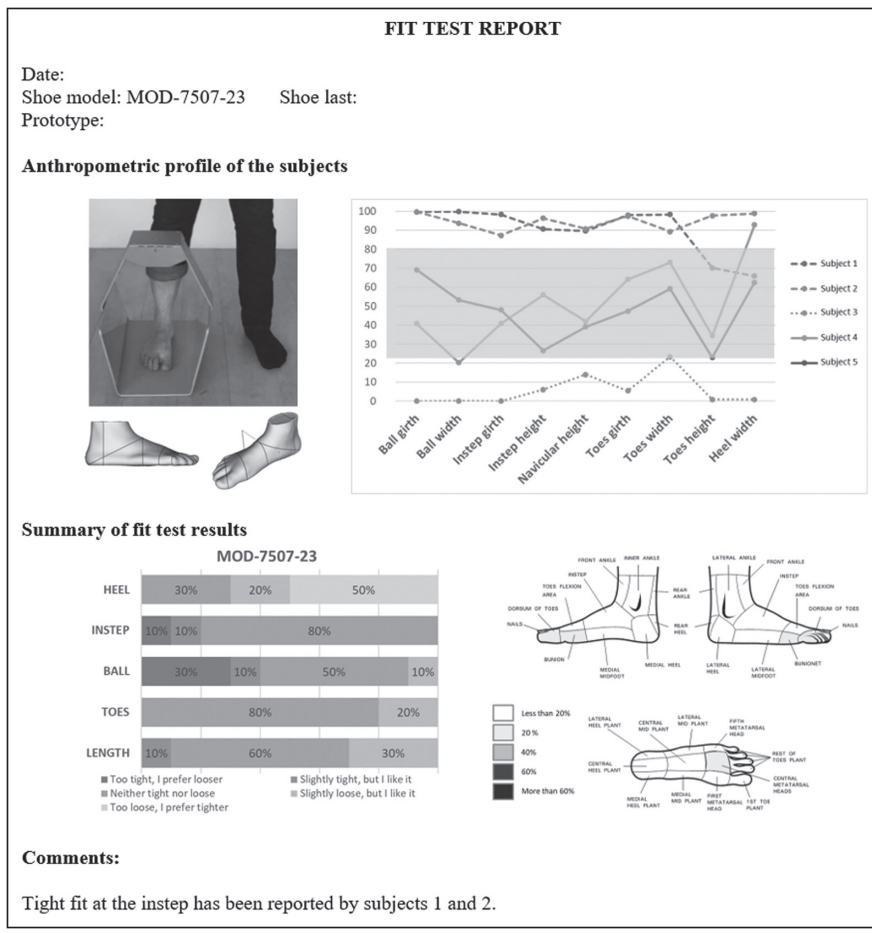


FIGURE 7.45 Fit test report.

TABLE 7.7
Recommended Modifications of the Shoe Last

Section	Shoe Last Measurement	Shoe Last Dimensions (mm) (Sneaker Example)	Modification Recommended From Fit Test Results (mm)
Length	Total length		+1
Ball	Ball girth		+4
	Ball width		(+2)
	Ball height		(+2)
Instep	Instep girth		(+2)
	Instep height		+2
Toes	Toes girth		0
	Toes height		0
Heel	Heel width		–2
Heel-to-Instep (for boots)	Heel to instep girth		(+2)

date, shoe model, version of the prototype, and shoe last reference. The second section includes the anthropometric characterization of the fit panel participants. Although they can be regular testers for the company, it is important to perform a 3D scan of both feet every session in order to track and control possible foot variations. The anthropometric characterization is shown in the plot including the percentile of the foot measurement in relation to the TP. The third section includes the results of the fitting test summarized for fit perception and foot pain zones. Both should meet the requirements of the COF established in Table 7.4 for this type of shoe. In this example, the results show fit problems at the heel, instep, and ball. This was the reason why it was recommended to modify the sections of the last in those problem areas. The resulting recommendations for the shoe last are shown in Table 7.7. These recommendations were sent directly to the last maker to make the changes.

In this table, the values in brackets were indirect modifications of the last dimensions. For instance, if the ball girth was increased by 4 mm, proportionally the width would increase by around 2 mm as well as the height.

After the modification of the last, a new prototype was manufactured and tested where it finally managed to meet the COF specifications and confirmed the fit of the shoe in the base size. The report obtained in the final test was filed as the Fit Standard Compliance Report. That means the latest version and the shoe variant of the sneaker meet the specifications of the Fit Standard designed according to the COF definition.

In this case study, the application of the SPEED process was limited to determining the base size. The design of the base size has been optimized, however the sizing was then applied according to the industrialization standards. These types of shoes are models with a short life on the market and require a fast time-to-market development process. From a cost-benefit perspective, nowadays it is very difficult to introduce the complete SPEED process.

As a result of using the SPEED process to develop and test the base size, the knowledge and information obtained was:

- Reference values of shoe last and tolerances for this type of shoe
- TP that can be used in other shoe models (in many cases, it can be applied to almost all the collection)
- COF and questionnaire
- Test protocol

This information will be an important asset for the company. Using it properly during the product development process will impact shortening the SPEED process.

Case Study 2: Application to Footwear Innovation

This case study illustrates some of the most important contributions of the SPEED process to footwear innovation. Footwear innovation began with sports shoes, but it has expanded to all types of footwear (e.g., safety, formal, children's, military). The SPEED methodology helps the footwear innovator make wise and informed decisions because the most important sources of information to build new concepts come from the users. The footwear innovation process described in this chapter was inspired by the methodologies and experience applied by the Footwear Innovation team of the IBV.

Footwear innovation projects spend more time developing requirements and COF before moving on to prototype development than is typically spent on simply creating a new aesthetic style. Good approaches to footwear innovation start with a learning phase and a creative phase before prototyping innovative new designs. These phases help determine the requirements and the design concepts, which are inputs to the SPEED process, but they use design loop trade studies and prototype tests of existing products to do so. In other words, in the development of innovative products design loop iterations are used to define the requirements, goals, and the COF for the new product. In this case study we review the learning and creative process phases and illustrate the use of trade studies and prototype testing for the development of a better athletic shoe.

The aim of the learning phase is to understand the main needs of the users, their preferences, the state of the art, the functions of the foot, key components of the shoe and the influence of the properties, etc. The learning process may include:

- User panels and questionnaires
- State-of-the-art scientific review
- Benchmarking trade studies
- Designer focus groups
- Requirements, specifications, and TP definition

Panels and questionnaires are used to understand what the expert users say are the problems and needs for the current footwear, as well as the positive elements. Panels are typically small groups of users. Questionnaires help us expand user sample size and get a better representation of different age groups, geographic areas, levels of experience, etc. Usually, user panels are done first since insights can be used to prepare the questionnaire.

A review of the state of the art is necessary at the beginning of the innovation process and includes a review of knowledge about:

- Anatomy, physiology, anthropometry, and biomechanics
- The effect of material and design of footwear components on the foot (e.g., movements, pressures, temperature, sock absorption)
- Review of the ratings of existing footwear done by users online, including ethnography which is a kind of social media research method

The information gleaned enables us to select a range of shoe models from the market that are rated the highest to establish a laboratory benchmark. A *benchmark* is a point of reference from which measurements may be made, that serves as a basis for evaluation or comparison, and against which new innovations can be judged. Benchmarking footwear includes physical tests with machines as well as with subjects. Machine testing assesses things such as durability, friction, and flexibility. The testing with subjects considers the real conditions of wear such as movements, postures, type of ground surface, and other environmental or use factors.

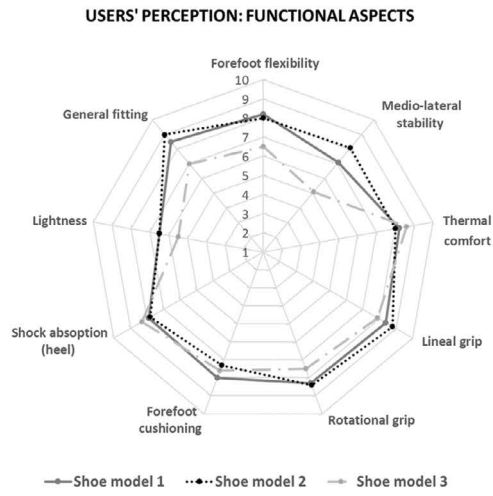
For this case study, we did a benchmark study with ten subjects who had extensive experience wearing the specific type of footwear under study. This was a trade study comparison of three models of shoes that were currently on the market. The goals of the benchmarking were to obtain: (1) a list of reference values for each test representing the performance of the best products in the market to enable us to track the strong and weak points of the new designs; (2) insights about which design elements and materials of the tested models are working well, or working poorly; and (3) innovation opportunities since it is possible to identify potential aspects for improvement.

Each of the subjects wore each of the shoe models and rated the shoes on a scale of 1–10 for each of the nine criteria (comfort test):

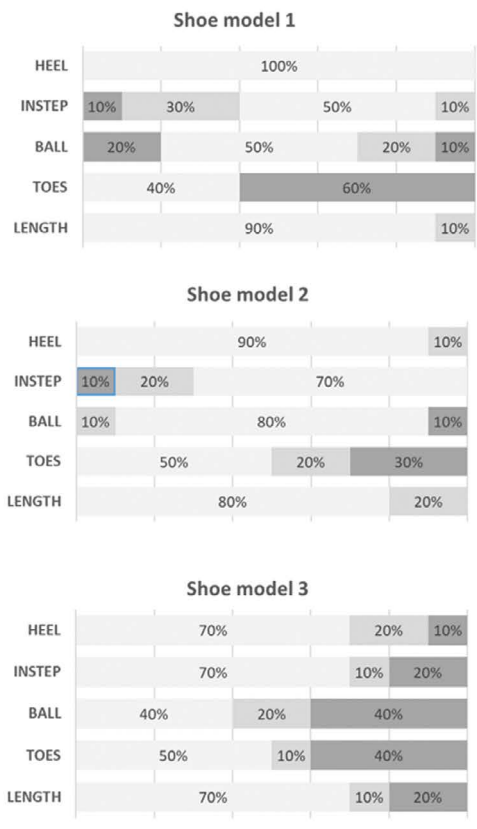
1. Forefoot flexibility
2. Medio-lateral stability
3. Thermal comfort
4. Linear grip
5. Rotational grip
6. Forefoot cushioning
7. Shock absorption
8. Lightness
9. General fitting

Fit in different areas of the foot was also assessed using the five-point scale: 1 being too tight, 2 being slightly tight but they like it, 3 being neither tight nor loose, 4 being slightly loose but they like it, and 5 being too loose such that they prefer tighter. t-Tests (described in Chapter 4) were done to determine which shoe results were significantly different.

The results are shown graphically in Figure 7.46. The radar chart summarized the results of the comfort test by illustrating the mean scores obtained from analyzing three shoes on a scale of 1 to 10 in the benchmark test. The bar charts illustrate the results of



a.



b.

FIGURE 7.46 (a) Rating of nine functional aspects of sports footwear obtained in the benchmarking test with subjects. (b) Fit assessment of the three models by foot zone.

the fit test by zones being light gray for the score 3 (neither tight nor loose), darker gray on the left corresponding to score 2 (slightly tight but they like it), the darkest gray in the left is the score 1 (too tight), darker gray in the right corresponds to score 4 (slightly loose but they like it), and the darkest gray in the right is the score 5 (too loose).

The "lightness" functional aspect of the shoe has the lowest scores for all three models; therefore, it could be a potential point for innovation.

Shoe model 3 had significantly worse forefoot flexibility, general fitting, and medio-lateral stability. This model had a material in the sole with greater rigidity and density. This material seemed to be affecting both the weight of the footwear and the reduction of flexibility in the forefoot.

The stability issue may have more to do with looseness, since model 3 had the worst stability and was looser in all areas except the toes, than the other models. Shoe model 1 has more looseness in the toes than the other two models and it has the second worst stability. Looseness may also have influenced the general fit score. Shoe model 3 had the worst rating for general fitting and shoe model 1 was the second worst.

These results provided good insights for defining and refining the COF for the new innovative shoe.

All the information generated in user panels, state-of-the-art review and benchmark testing was then reviewed by the design team and designer focus groups. We used multidisciplinary teams that included experts on anatomy, biomechanics, manufacturing methods, materials science, and user experience. They prioritized the objectives of the product innovation and set the requirements and specifications for the new design.

The learning process was finalized with a document that included the footwear requirements, KPIs, specifications and the COF. This was the input information as indicated in the SPEED process.

The specifications might include the methods of verification (machine tests, subject tests) and the KPI thresholds. The inputs will include specifications of the whole shoe and specifications for the components. For example, in the case of a removable insole, requirements might include among others, the thickness that can be variable at the heel midfoot and toes, the anatomical elements (heel cup, arch support, metatarsal support), material properties (shoe hardness), material properties and thickness after fatigue, moisture absorption, and drying speed. The COF included the test protocol, the fit questionnaire, the range of reference values for the last, and the acceptable fit score values (the KPIs for shoe fit).

Next came the creative phase. The objective of this phase was to create and develop the innovation concept. This is a critical phase of the innovation process since it is very difficult to generate something new. At the same time, it must be feasible from various perspectives such as manufacturability, cost, and integration with shoe components. The design team develops different design concepts that they hope can meet the requirements, drafts CAD versions and uses rapid prototyping of components or complete form factor shapes that can be touched and viewed to make decisions and discard options that, in a quick view, are not going to be winners. The best options are then chosen to move on to the prototyping phase.

The aim of prototyping at this point is to materialize the innovation concept into a complete shoe. Then design loop testing is used to refine the product. The process is iterative and may include mini-loops after the fit and performance tests as well as it is conceived the SPEED methodology. Depending on the complexity, it usually starts with a coarse design that will be refined in different iterations. In this process, the use of simulation methodologies and prototyping technologies can speed up the process significantly.

Since creating a lighter shoe seemed to be a good innovation opportunity, for this case study the goal was to arrive at a shoe that was equivalent to the best benchmark shoes, but much lighter. The concept to achieve this goal was to make a lighter sole. After the creative phase, the design concept decided to reduce the weight of the sole by material reduction in the internal side of the sole. The first step in the prototype evaluation was to perform several simulations in a finite element model. The input forces for the simulation were pressure patterns from biomechanical studies of the TP. The simulations enabled us to compare the deformation of the sole considering different types of materials and thicknesses of the vertical walls. This resulted in a material reduction in some areas but with a higher concentration of material at the heel and the metatarsal arch.

The design that appeared best in the simulation was then prototyped using rapid prototyping technologies and tested mechanically to validate the simulation results and check mechanical properties such as fatigue, that is the resistance under long cyclic loads. This is illustrated in Figure 7.47.

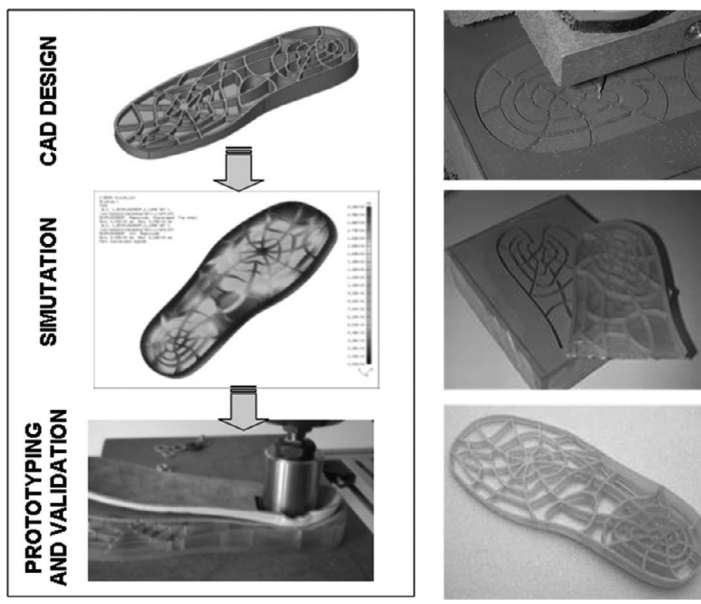


FIGURE 7.47 Simulation, prototyping and testing the innovation concept.

At this stage, it is usual to redesign some elements due to manufacturing specifications. For instance, the design of the previous sole required a redesign in the heel to avoid small-deep cavities that prevent the removal of the sole from the mold. Next, the complete shoe was prototyped at this stage using the same process described in the first case study, with tests of the base size last, refinement, and iterative prototype tests with subjects and comparison against the benchmark shoes. For these tests, which required active movement while wearing the shoes it was important to prototype the shoes for both feet. The prototypes were tested in real conditions, moving the testing set up to a sports facility, both indoor and outdoor. The type of sports surface influences footwear performance and the assessment results.

The test protocol must include the critical and most frequent sports gestures and the testing protocols, the questionnaire must be adapted for each specific project. As a result, the test sessions can be a lot longer than a regular fitting test, and if there are repeated tests with the same subjects it is often necessary to have a paid test panel for at least the first set of tests. That was the situation for this case study.

For the first tests where we assessed the base size, we used expert users of this type of sports footwear. We selected ten subjects that we could use repeatedly for testing throughout the project. This provided some consistency in their responses because by understanding the COF and participating in all the testing, the subjects became expert fitters. Since there was a risk associated with using a small sample size like this, we checked their foot sizes and relative foot proportions by examining their percentile scores for each of the seven foot measurements.

A practical representation of the anthropometric profile of the subjects is shown in Table 7.8. In this table, the values are shaded to indicate their relative size. The smallest values are white, and the largest values are the darkest.

We see that we have a good range of subjects with values ranging from the 4th percentile for subjects 4 and 7 to nearly the 100th percentile for subject 6. Subjects 4 and 7 show small anthropometric dimensions in the ball area (narrow foot) while

TABLE 7.8
Anthropometric Characterization of Subjects with the Corresponding Percentile of the TP

Subject	Ball Girth	Ball Width	Ball Height	Toes Girth	Toes Height	Toes Width	Instep Height
Sub1	64	67	49	92	33	93	62
Sub2	61	44	58	92	41	95	35
Sub3	90	97	22	80	89	73	85
Sub4	7	5	30	6	54	4	90
Sub5	28	29	27	21	18	22	48
Sub6	60	49	77	75	100	53	88
Sub7	11	13	16	18	13	28	4
Sub8	42	42	43	44	26	47	39
Sub9	48	39	46	59	35	61	21
Sub10	44	47	51	85	59	87	29

subject 3 shows a wide foot in the ball. If the results of the shoe fit test show good fitting results in the ball area for all the subjects, then the accommodation range of the shoe for that area is well designed and optimized to cover most of the anthropometric variability. Subject 6 shows an extreme value for toe height. It was interesting to check the fit assessment of the shoe at the toes in relation to the other subjects. This type of representation of the subjects ensured we had good representation in our sample and allowed us to compare their fit scores against their foot proportions relative to the TP.

Once we had the new shoe model base size validated, we tested it against the baseline shoe models using the same panel of subjects to ensure that it was both lighter and as good or better for all the other criteria. Sometimes, the testing against the baseline can indicate the need for small changes to refine the final shoe and proceed to the final full-scale fit test. For these tests, it is important to use the final materials and designs for all components, including the insole, upper material, upper design, and sole. It is also important that the number of shoe pairs manufactured should be enough to cover tests with subjects and tests with machines. In the subject test, each participant should be given a new pair of shoes to ensure consistent sample conditions and prevent potential effects of shoe breakdown.

Once the final design is approved, it is time to test the full size range in a sizing loop full fit test. This ensures that the scaling did not introduce problems for some sizes and that the correct assortment of sizes for the TP is used. This process requires time to do the grading of all the components (last, upper patterns, sole, and insole). Soles and insoles are manufactured with metal molds that include the right and left foot. Usually, the insoles can cover a couple of sizes, but the molds of the sole are one size, which is why it needs more time for manufacturing.

The final full fit test is also helpful if the brand is interested in publishing the results in a scientific journal. This is a strategy that is starting to be followed by top brands since it is a way to certify the results and marketing claims. In that case, a minimum of 50 subjects is required for the study. Descriptive statistics as well as t-test and discriminant analyses are methodologies that can be used to analyze the results comparing different alternatives.

In this case study, we used a sample of 50 subjects drawn from the TP so that we had multiple people wearing each size. We checked the sample against a larger sample drawn from the TP to verify how well they represented the TP. We illustrate the coverage in bivariate plots such as the one in Figure 7.48. This shows the Foot Length (FL) versus the Metatarsal Girth (BG) for the fit test sample versus the full sample from the TP. The large dots are the fit test sample subjects and the small dots are the subjects in full sample from the TP. Both samples cover the range of European sizes 35–44 and include men and women.

For the final fit test, it was necessary to test each subject in the size that fit them the best. In this case study, the best fitting size was first determined by having the subject try their usual shoe size and either a larger or smaller size as well to confirm, use, and record the best size. Since the improvement over the baseline shoe models was confirmed in the base size testing it is not necessary to test against other shoe models at this stage. However, doing so can provide additional marketing materials and proof of improvement.

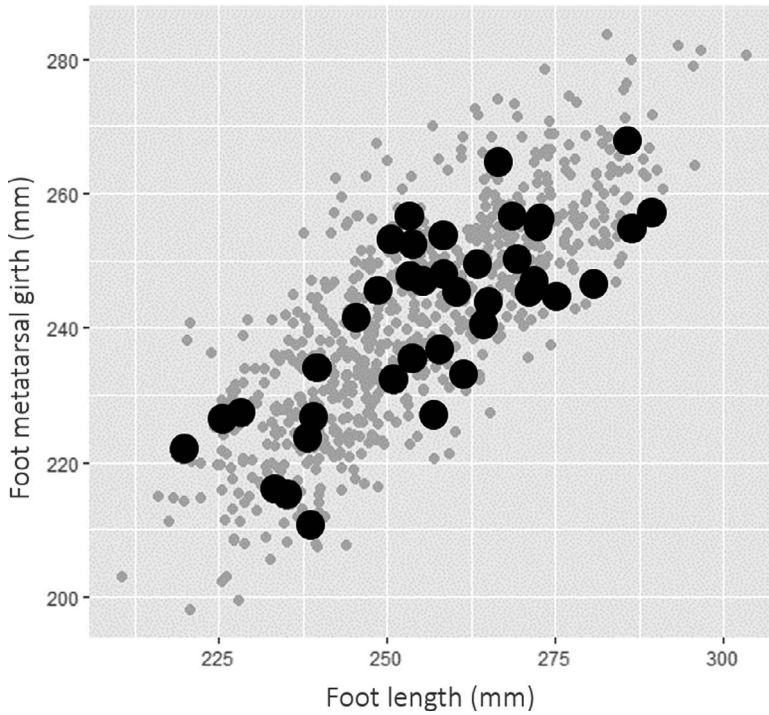


FIGURE 7.48 Bivariant plot foot length – metatarsal girth (mm) of the subjects overlapped with the TP (males and females).

As a result of using the SPEED process in the development process of an innovative footwear test on the base size, the knowledge and information obtained was:

- A functional benchmark of current shoe models of the market, some of them from competitors, and the positioning of the new development.
- Inspiration and ideas to cover a functional demand from the TP with an innovative concept.
- Guidelines of design elements of the shoe that are providing a functional value as a result of physical tests and subjective perception test with subjects.
- The repeated application of this methodology will facilitate the creation of databases, establish thresholds for physical tests associated with comfort perception, and ultimately define KPIs related to functional aspects and footwear fit.
- Test protocols to validate the concept and to generate the marketing claims supported with research experimentation.

This information will be an important asset for the company. Using it properly during the product development process will impact shortening the SPEED process for future shoe development.

REFERENCES

García, A. C., Porcar, R., Bataller, A., Page, A., Martínez, A., & González, J. C. (2021). Functionality and personalization. Importance of design aspects and methodological approach. In 2001 World Congress on Mass Customisation and Personalisation.

Goonetilleke, R. S. (2012). The Science of Footwear. CRC Press.

ISO 19407:2023 Footwear — Sizing — Conversion of Sizing Systems. (2023).

ISO 20345:2021 Personal protective equipment — Safety footwear. (2021).

ISO/TS 19408:2015—Footwear—Sizing—Vocabulary and terminology. (2015).

Kapandji, I. A. (1970). The Physiology of the Joints, Edinburgh. E & S Livingstone.

Kouchi, M., Ballester, A., McDonald, C., Jurca, A., Dessery, Y., Armitage, Z., Schwartz, L., Martirosyan, V., & Dubey, S. (2021). IEEE SA 3D Body Processing Industry Connections–Comprehensive Review of Foot Measurements Terminology in Use. IEEE SA 3D Body Processing Industry Connections–Comprehensive Review of Foot Measurements Terminology in Use, 1–63.

Ledoux, W., & Telfer, S. (2022). Foot and Ankle Biomechanics. Academic Press.

McNeel, R. (2005). Shoe Design and Visualization. Institute of Biomechanics of Valencia, Rhinoceros Advanced Training Series.

Murray, M. P., Drought, A. B., & Kory, R. C. (1964). Walking Patterns of Normal Men. JBJS, *46*(2), 335–360.

Nigg, B. M. (2010). Biomechanics of Sport Shoes. University of Calgary.

Nigg, B., Behling, A. -V., & Hamill, J. (2019). Foot Pronation. Footwear Science, 11(3), 131–134. https://doi.org/10.1080/19424280.2019.1673489

Olaso, J., González, J. C., Alemany, S., Medina, E., López, A., Martín, C., Prat, J., & Soler, C. (2007). Study of the influence of fitting and walking condition in foot dorsal pressure. Proceedings of the 8th Footwear Biomechanics Symposium–Taipei 2007, 41.

Olaso Melis, J. C., Priego Quesada, J. I., Lucas-Cuevas, A. G., González García, J. C., & Puigcerver Palau, S. (2016). Soccer players' Fitting Perception of Different Upper Boot Materials. Applied Ergonomics, *55*, 27–32. https://doi.org/10.1016/j.apergo.2016.01.005

Ramiro, J., Alcántara, E., Forner, A., Ferrandis, R., García-Belenguer, A., Durá, J. V., & Llana, S. (1995). Recommendations Guide for Footwear Design. Institute of Biomechanics of Valencia, 135–151 (In Spanish).

Richards, J., Levine, D., & Whittle, M. W. (2022). Whittle's Gait Analysis-E-Book. Elsevier Health Sciences.

Robinette, K. M., Blackwell, S., Daanen, H., Boehmer, M., & Fleming, S. (2002). Civilian American and European Surface Anthropometry Resource (caesar), final report. Volume 1. Summary. Sytronics Inc., Dayton, OH.

Root, M. L. (1971). Biomechanical Examination of the Foot. (1st ed. Volume 1). Clinical Biomechanics Corp.

Valero, J., Ballester, A., Nácher, B., Solves, C., Luca, C., Pesce, M., & Alemany, S. (2023). A Statistical Size Recommender for Safety Footwear Based on 3D Foot Data. https://doi.org/10.15221/23.40

Welte, L., Kelly, L. A., Kessler, S. E., Lieberman, D. E., D'Andrea, S. E., Lichtwark, G. A., & Rainbow, M. J. (2021). The Extensibility of the Plantar Fascia Influences the Windlass Mechanism During Human Running. Proceedings of the Royal Society B: Biological Sciences, 288(1943), Article 1943. https://doi.org/10.1098/rspb.2020.2095

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Glossary

Accuracy: The ability of the tool to capture the true measurement.

Allometric grading: Three-dimensional grading that matches the body growth in different body areas – also see **grading**.

Allometry: The study of the relative growth of parts of the body in relation to the growth of the whole.

Allowable error: Errors with repeated measurement that are deemed acceptable.

Ambinocular: The total field of vision seen by at least one or both eye/s (left or right).

Anthropometrist: Someone trained in measuring the human body.

Anthropometry: The scientific study of the measurements and proportions of the human body.

Average: In mathematics and statistics, this refers to the arithmetic mean which is the sum of the observations divided by the count.

Base: (1) The starting point for the development of a size range. (2) In mathematics, it is the number of different digits that a system of counting uses to represent numbers. For example, the most common base used today is the decimal system, which uses numbers 0 to 9 to count. Computers use a binary base, 0 and 1.

Benchmark: A standard or point of reference against which things may be compared or assessed.

Bivariate: An analysis, graph, table, or chart of two variables.

Block: (1) A two-dimensional master pattern shape that contains all the fit elements and the best or truest representation of the desired fit, to be used as the "master" reference. (2) In experimental design, a block is a group of experimental units that are like each other, for example, male and female in gender or 18–25 age block in age.

Body form: See **manikin**.

Body grade: A chart of scaled body measurements.

Boundary cases: Cases at the extreme edges (boundary) that we wish to accommodate.

Calibration: The process of determining, checking, or rectifying the settings or gradations on a measuring instrument or other piece of precision equipment.

Case: A single individual to be represented in a product design or evaluation. This representation can take three forms: (1) a list of measurements of an individual, (2) a 3D (or 4D) model of an individual, or (3) the actual individual. The actual individual is called a fit model or a live model.

Case study: A detailed example in a real-world context.

Concept-of-Fit (COF): A description of how the product should fit, what constitutes a good fit versus poor fit, and methods for assessing or measuring it.

Correlation: A measure of how two or more variables are related to each other after they have been standardized. It is denoted r or R and has a value between –1 and 1.

Covariance: A measure of how much the deviation of one variable from its mean matches the deviation of another variable from its mean in its original

measurement units. The deviation is measured using the variance, which is the squared standard deviation.

Cut ratio: The per unit ratio of items to cut per size.

Database: Collection of many samples or datasets.

Design element: The parts of the design that form a new style.

Durometer: A standardized way of measuring the hardness of materials like rubber or plastic.

Ethnicity: (1) A group that shares cultural, traditional, and familial bonds and experiences. (2) For purposes of product development, this may be defined as a group having similar physical traits since the purpose for asking is to ensure all groups are adequately represented and accommodated. (3) Sometimes called race.

Experimental design: Designing and setting up the conditions for testing.

Feature envelope: An area that contains a selected percentage of a particular landmark or body feature, such as an ellipse that encloses 95% of the right pupil landmarks from a sample.

Fit audit: A fit test of the full set of sizes that uses a sample representative of the target population (TP) for one or more of the following purposes: (1) assessing a fit standard, (2) determining if a product meets a fit standard, or (3) determining if the sizes are appropriate for a new TP.

Fit element: Design elements that impact fit.

Fit intention: The desired fit appearance and intended functionality of the product.

Fit mapping: Fit mapping is the process of comparing (or mapping) the range of fit of each size with the total variability in our sample.

Fit model: (1) A person employed by a product development team to try products on for fit, function, and aesthetic evaluation. (2) A case.

Fit standard: Part of a suite of quality control and quality assurance standards. Its function is to ensure consistency in fit and sizing between styles within a brand or company. It includes such things as body measurements, blocks/slopers, manikins, pattern measurements, human models, and grading.

Form factor: Fit intentions expressed in a 3D form either physically or digitally. When it is a wearable physical form, it is called a mock-up or a prototype.

Frequency distribution: A description of the number of times (the frequency) every possible value of a variable occurs. It is depicted using a graph or a frequency table.

Grading: The practice of creating additional sizes from a pattern by moving points on its perimeter along two-dimensional coordinate axes by predetermined values.

Histogram: A chart that depicts the count of a variable's values in a series of bars.

Homologous: Having the same relative position, value, or structure.

Informed consent: A process used for the protection of human subjects in research. It involves disclosing to potential subjects, the information needed to make an informed decision, and promoting the voluntariness of the decision about whether to participate. Consent is documented using a form signed by the subject, the investigator, and a witness.

Interface: The interaction between two things, that is, person and wearable or two types of cloth.

Inter-observer error: Measurement differences between measurers or observers.

Intra-observer error: Differences in measurements taken by one measurer or observer.

Key performance indicators (KPIs): Quantifiable measures of product performance that serve as the list of requirements for the product.

Last: (1) For footwear, it is a form shaped similarly to a foot that is used to make or repair shoes. (2) For hats, it is a form for making a hat that is shaped like both the designed hat shape and the head shape where it will be worn.

Manikin: (1) A 3D representation of the human body or its body segments, either digitally or physically. (2) A model or replica of a human being or a stylized representation of a human. (3) Workroom manikin: a figure used for displaying or fitting clothes, otherwise known as a tailor's dummy, and usually used in the pre-production process.

Mannequin: See **manikin**.

Measurement uncertainty: Measurement uncertainty is an estimate of the level of accuracy and precision with which a measurement can be taken with a given tool or process.

Median: (1) The point in the distribution of a variable that divides the observations into two halves. (2) The 50th percentile.

Mock-up: A prototype that is not fully functional and/or not made from the intended materials.

Mode: (1) The peak in the distribution with the highest number of observations. (2) The most frequent number.

Multivariate analysis: A statistical analysis with simultaneous observation and analysis of multiple outcome variables.

Non-fit design elements: The design elements that, when changed with a new style, are not constrained by the fit standard, that is, do not impact fit.

Objective measures: Measures based on observable phenomena uninfluenced by emotions or personal prejudices.

Pants-form: A manikin suitable for bifurcated garments such as trousers.

Parallax: The observed displacement of an object caused by the change in the observer's point of view.

Pattern: A model, template, or guide to make a product in either digital or physical format. It might consist of shaped pieces of metal, card, paper, or other material used as a template for processes such as cutting out, shaping, or drilling.

Percentile: A statistic indicating the percentage of observations that are smaller for a single variable. In other words, k percent of the observations fall below the kth percentile.

Pilot test: A test to guide to try out data collection procedures and the COF before finalizing them.

Point cloud: A set of 3D points from a scan.

Population: Every member of a group of people or things.

Precision: The consistency of measurement taken on static or fixed objects.

Principal component analysis (PCA): A statistical method used to understand higher dimensional relationships for the purpose of reducing the number of variables to a more manageable set.

Product development process: The cycle of developing a product from concept to the prototype approved for production, also called the pre-production process.

Product/garment grade: Scaled garment patterns who together create a range of sizes in one style.

Prototype fit test/prototype test: Prototype tests are fit tests that evaluate the fit and performance of the entire product-human system using either mock-ups or fully functional prototypes of the product.

Race (see also ethnicity): (1) For purposes of product development, this usually is defined as a group having similar physical traits. The purpose of asking is to ensure all groups are adequately represented and accommodated. (2) A group sharing a common lineage. (3) Sometimes called ethnicity.

Resolution: The closeness of the settings, gradations, or points such as how many decimal points are on the instrument. When gradations or points are close together, they are high resolution and when they are far apart, they are low resolution.

Sample: (1) A part or fraction of a whole thing, whole group, or a population. (2) In apparel – a prototype garment that serves as an example for others to copy in the mass production of that garment. (3) In statistics – a selected segment of a population is studied to gain knowledge of the whole.

Sampling bias: A tendency to favor the selection of units with particular characteristics.

Simple random sample: A sample where each item from a population of items has an exactly equal chance of being selected.

Sizing survey: A collection of anthropometric data.

Sloper: See **block**.

Stratification: The dividing of a population into groups and selecting samples from each group.

Subject: A participant in a study.

Subjective measures: Measurements that require an opinion.

Sustainable fit standard: A **fit standard** that can be followed and maintained for all products of a type and provide a consistently good fit for the target population without unnecessary sizes or size duplication.

Sweet spot: The area or location that is the most effective.

Target population (TP): The population of intended users. The group of people for which a product is made.

Tariff: A table indicating the number of products in each size category to be produced and/or purchased. This can be expressed as the per unit amount (the cut ratio), the percentage, the number per ten thousand units, or the number for some other total number of units.

Trade study: Comparisons of a small number of treatment options, such as different shapes, paddings, adjustment mechanisms, environmental conditions, or user tasks for the purpose of narrowing down design options or verifying design changes are effective and do not cause issues elsewhere.

Treatments: The product combinations and the conditions we plan to use.

Univariate: Statistical observation and analysis of just one variable.

Weighting: A technique in research where the observations are adjusted to reflect a project population more accurately. This is used when the sample has more representatives from some groups and fewer representatives from other groups than the population.

Index

Note: *Italic* page numbers refer to *figures* and **bold** page numbers refer to **tables**

D

H

I

K

L

M

N

O

P

Q

R

S

T