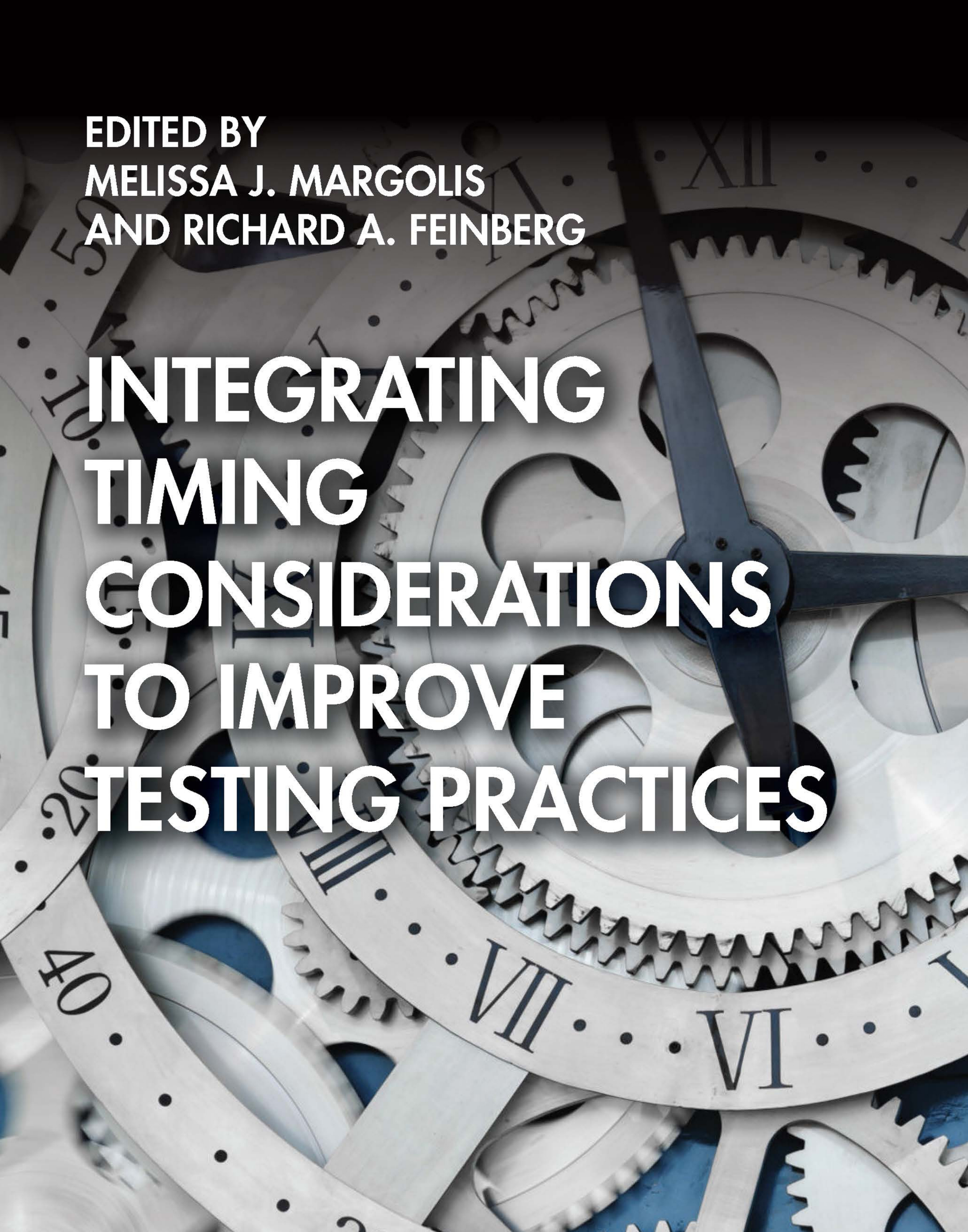
EDITED BY
MELISSA J. MARGOLIS
AND RICHARD A. FEINBERG



INTEGRATING TIMING CONSIDERATIONS TO IMPROVE TESTING PRACTICES

APPLICATIONS OF EDUCATIONAL MEASUREMENT AND ASSESSMENT

NCME national council on measurement in education

ROUTLEDGE

Integrating Timing Considerations to Improve Testing Practices

Integrating Timing Considerations to Improve Testing Practices synthesizes a wealth of theory and research on assessment-related timing issues into actionable advice for test development, administration, and scoring. One of the major advantages of computer-based testing is the capability to passively record test-taking metadata—including how examinees use time and how time affects testing outcomes. This has opened many questions for testing administrators. Is there a trade-off between speed and accuracy in test taking? What considerations should influence equitable decisions about extended-time accommodations? How can test administrators use timing data to balance the costs and resulting validity of tests administered at commercial testing centers?

In this comprehensive volume, experts in the field discuss the impact of timing considerations, constraints, and policies on valid score interpretations; administrative accommodations, test construction, and examinees' experiences and behaviors; and how to implement the findings in practice. These 12 chapters provide invaluable resources for testing professionals to better understand the inextricable links between effective time allocation and the purposes of high-stakes testing.

Melissa J. Margolis is Senior Measurement Scientist at the National Board of Medical Examiners, USA.

Richard A. Feinberg is Senior Psychometrician at the National Board of Medical Examiners, USA.

The NCME Applications of Educational Measurement and Assessment Book Series

Technology and Testing: Improving Educational and Psychological Measurement
Edited by Fritz Drasgow

Meeting the Challenges to Measurement in an Era of Accountability
Edited by Henry Braun

Fairness in Educational Assessment and Measurement
Edited by Neil J. Dorans and Linda L. Cook

Testing in the Professions: Credentialing Policies and Practice
Edited by Susan Davis-Becker and Chad W. Buckendahl

Validation of Score Meaning for the Next Generation of Assessments: The Use of Response Processes
Edited by Kadriye Ercikan and James W. Pellegrino

Preparing Students for College and Careers: Theory, Measurement, and Educational Practice
Edited by Katie Larsen McClarty, Krista D. Mattern, and Matthew N. Gaertner

Score Reporting Research and Applications
Edited by Diego Zapata-Rivera

Classroom Assessment and Educational Measurement
Edited by Susan M. Brookhart and James H. McMillan

Integrating Timing Considerations to Improve Testing Practices
Edited by Melissa J. Margolis and Richard A. Feinberg

For more information about this series, please visit: https://www.routledge.com/NCME-APPLICATIONS-OF-EDUCATIONAL-MEASUREMENT-AND-ASSESSMENT/book-series/NCME

Integrating Timing Considerations to Improve Testing Practices

Edited by Melissa J. Margolis and Richard A. Feinberg

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2020
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
A catalog record for this title has been requested

ISBN: 978-1-138-47975-3 (hbk)
ISBN: 978-1-138-47976-0 (pbk)
ISBN: 978-1-351-06478-1 (ebk)

Typeset in Minion Pro
by Cenveo® Publisher Services

Contents

Foreword

Most standardized tests require that examinees complete the required task(s) within a specified amount of time. Indeed, time limits can be regarded as a fundamental part of standardization by helping to ensure that all examinees complete the examination under comparable time constraints. In some instances, time limits may be quite stringent and the expectation is that most examinees will not complete the test. This is the case when speed of response is part of the construct being measured. In other instances, test scores are interpreted as an indication of the extent to which examinees have mastered some domain of knowledge; for these tests, speed of response is not part of the construct. This is true for most tests used in both educational testing and credentialing, where insufficient time is considered a source of construct-irrelevant variance. Nonetheless, for administrative and logistic reasons it is still necessary to impose time limits. The challenge is to establish limits that encourage examinees to work at a reasonable pace without negatively impacting their test performance.

In order to help determine timing requirements, early timing research focused on evaluating simple metrics such as the proportion of examinees who did not complete the test. More recently, however, the complexity, importance, and prominence of timing studies have increased. With the advent of computerized test delivery, extensive and precise data are available to shed light on the time examinees allocate to individual test items. In addition, the field of assessment has entered an era where change is the norm. The availability of technology coupled with the desire for more authentic assessments has led to the development of complex item formats that are intended to assess higher-order thinking skills. The drag-and-drop format, multi-media items, and case-based scenarios are just a few examples of these novel item formats that now populate high-stakes examinations in K-12 education and professional credentialing. These complex items were developed to address a more complex assessment need; not surprisingly, this complexity results in needing more time to read and answer these items than is typically required for traditional MCQs. Thus, out of necessity, monitoring response time has become a routine activity, with some testing agencies subjecting test items and test forms to rigorous analyses as part of pretesting.

Other assessment-related developments and considerations have further challenged conventional thinking about timing and its impact on score interpretation. Examples include:

- A recognition that the profession lacks evidence-based guidelines for administering tests to examinees who are granted additional testing time to accommodate a disability and the impact of additional time on construct interpretation.
- The introduction of technologies that alter the interface between examinees and the assessment tasks. Though some of these would be expected to improve examinee

performance (e.g., hover text for definitions; pacing aids), others could hinder performance (e.g., item text and graphic not fitting within a single screen).

- The use of commercial test centers for test delivery, which creates a direct relationship between testing time and cost to examinees.
- A rise in the use of computer-adaptive testing (CAT). In many instances, examinees not only see different items but also see varying numbers of items during a specified testing time.
- Evolving public expectations about test fairness, with scholarly articles and the popular press posing serious challenges to the relevance of time constraints on high-stakes admissions tests.
- A growing trend for state-wide assessments in K-12 settings to allow students as much time as needed to complete a test (within practical constraints).

It was in this context that the National Board of Medical Examiners (NBME) conducted a randomized experiment to evaluate the impact of time constraints on test performance on an examination for physician licensure. Results indicated that examinees did indeed earn higher scores under relaxed time constraints and that the benefit was more pronounced for low scoring examinees. The findings led to numerous meetings and seminars among NBME staff to determine whether test speededness was a problem and, if so, what should be done to address it. As one might expect, the obvious solutions also had obvious limitations. For example, additional time was not feasible given that the test in question already required a full day, and reducing exam length was undesirable due to its impact on reliability and content representation.

Inquiry into solutions raised numerous questions such as: *To what extent is response speed part of the construct of interest? How much does the construct change as a consequence of more generous time limits? Do examinees use their time effectively? Should two examinees who obtained the same score on a test be considered to have the same level of proficiency if one required substantially more time to complete the test? Does examinee response time provide information that can be used for predictive, diagnostic, or remedial purposes? What expectation do stakeholders (e.g., educators, consumers) have regarding time constraints on high-stakes tests and how should such expectations be factored into policy decisions regarding testing time?*

In pursuit of answers, NBME initiated a second experiment and completed evaluations of numerous alternative test design and administration models. The organization also sought input from colleagues in other testing organizations and reached out to experts in academic settings who had studied response speed and its influence on cognitive task performance. Through these interactions it became evident that the expertise on this specialized but important topic was impressive and that, collectively, there were compelling stories to tell, valuable data to share, and constructive insights to consider.

To bring this expertise together, NBME sponsored a conference in October 2017 attended by approximately 150 measurement scientists and policy leaders. The conference was affectionately branded as *TIME—Timing Impact on Measurement in Education*. Its goal was to provide a forum for scholars in psychometrics, cognitive science, and education to share research and perspectives on timing and pacing for high-stakes tests and discuss the implications for policy and practice. This volume reflects the ideas that inspired and were discussed at that conference.

The book comprises four sections. *Section I* provides a historical context for timing in standardized testing, offers a framework for evaluating the impact of time limits on score interpretations, and discusses policy considerations, including the provision of additional time to accommodate examinees with disabilities. *Section II* documents empirical research on examinee pacing, the relationship between time constraints and testing outcomes, and timing considerations in the context of different assessment formats and testing modalities. *Section III* reviews research on the relationship between speed of processing and cognitive ability, presents a model illustrating the importance of response speed to decision making in the work setting, and provides an analysis of speed-accuracy tradeoff models and their implications for construct interpretation. *Section IV* describes novel methods for using response time data to improve test construction and identify threats to validity due to examinee behavior such as lack of engagement and cheating. This book is an excitingly broad compilation of the best research on this topic area and it will be useful to testing personnel, graduate students, and faculty with an interest in almost any aspect of examination timing—from the very practical to the completely theoretical.

Those of us who were responsible for organizing the conference and editing this volume express our sincere gratitude to the authors and conference participants for making this work possible.

Mark Raymond

Acknowledgments

We would like to acknowledge the small village that provided support to us on this journey. Tremendous thanks go to the chapter authors for their dedication, perseverance, and high standards; without their efforts, there would be no content to share. Also deserving of acknowledgment are the numerous volunteers who gave of their time and expertise to provide reviews of early versions of the chapters. Specifically, we would like to thank Benjamin Lovett, Brent Bridgeman, Brian Clauser, Catherine Farmer, Chunyan Liu, Daniel Jurich, David Swanson, Jonathan Rubright, Kimberly Swygert, Mark Raymond, Matthias von Davier, Michael Jodoin, and Peter Baldwin for their thoughtful feedback that greatly improved the quality of the resulting volume by helping to ensure that the material was as relevant and accessible as possible.

We also want to thank Mark Raymond for his vision in putting together the *Timing Impact on Measurement and Education (TIME) Conference* that led to production of this volume and our employer, National Board of Medical Examiners, both for sponsoring the TIME conference and for supporting our work to translate the conference content into a comprehensive written volume documenting this important research topic.

Additional appreciation goes both to the members of the NCME Editorial Board—Brian Clauser, Fritz Drasgow, Henry Braun, Michael Kolen, Neil Dorans, and Rochelle Michel—for recognizing the importance of timing in the context of educational testing and agreeing to sponsor this topic as part of the NCME book series and to Daniel Schwartz, our editor at Routledge, for his patience and support for this volume.

Finally, we want to express our sincere gratitude to Brian Clauser for his guidance throughout the process of preparing this volume. His contributions as Editorial Board member, reviewer, mentor, and friend were invaluable, and the quality of the volume benefitted greatly from his vision and expertise.

Contributors

Paul De Boeck, The Ohio State University, deboeck.2@osu.edu

Sandra M. Botha, University of Massachusetts Amherst, sbotha@umass.edu

Brent Bridgeman, Educational Testing Service, bbridgeman@ets.org

Wayne J. Camara, ACT, Wayne.Camara@act.org

Brian E. Clauser, National Board of Medical Examiners, bclauser@nbme.org

Matthias von Davier, Boston College, Lynch School of Education, Matthias.vonDavier@bc.edu

Richard A. Feinberg, National Board of Medical Examiners, rfeinberg@nbme.org

Polina Harik, National Board of Medical Examiners, pharik@nbme.org

Deborah J. Harris, University of Iowa, deborah-harris@uiowa.edu

Daniel P. Jurich, National Board of Medical Examiners, djurich@nbme.org

Michael Kane, Educational Testing Service, mkane@ets.org

Megan R. Kuhfeld, NWEA, megan.kuhfeld@nwea.org

Patrick Kyllonen, Educational Testing Service, pkyllonen@ets.org

Seo Young Lee, Prometric, seoyoung.lee@prometric.com

Benjamin J. Lovett, Teachers College, Columbia University, BL2799@tc.columbia.edu

Melissa J. Margolis, National Board of Medical Examiners, mmargolis@nbme.org

Mark Raymond, National Conference of Bar Examiners, mraymond@ncbex.org

Frank Rijmen, Cambium Assessment, frank.rijmen@cambiumassessment.com

Stephen G. Sireci, Center for Educational Assessment, University of Massachusetts Amherst, sireci@acad.umass.edu

Rick Thomas, Georgia Institute of Technology, rthomas82@gatech.edu

Steven L. Wise, NWEA, steve.wise@nwea.org

James A. Wollack, University of Wisconsin-Madison, jwollack@wisc.edu

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

1

A History of Test Speededness

Tracing the Evolution of Theory and Practice

Daniel P. Jurich

There are many practical reasons for administering tests with time limits, most of which relate to the logistics and efficiency of test administration (Bandalos, 2018, p. 59; Morrison, 1960; Rindler, 1979). For example, time limits help to control costs for test developers who often must pay expenses associated with the testing space as well as staff costs for necessary personnel (e.g., test proctors). However, time limits can also serve essential measurement-related functions. Perhaps most importantly, they help to standardize the testing conditions and improve the ability to compare performance across examinees. Concrete evidence of timed standardized testing dates back at least to the Chinese Civil Service examinations administered in the 15th century. At that time, candidates were given one night and one day to complete poems and essays that were used to evaluate their style and penmanship (Martin, 1870). In the United States, the Army led early applications of timed structured cognitive and noncognitive testing through exams such as the Army Alpha and Beta. Beginning in 1917, these tests were used to evaluate World War I recruits on a variety of cognitive skills such as arithmetic reasoning and verbal aptitude (Gregory, 2004; Schnipke & Scrams, 2002). Since these beginnings, standardized examinations with time limits have become ubiquitous within modern society.

Although the implementation of time limits in standardized testing usually occurs due to reasons unrelated to measurement, time constraints can have substantial impact on the validity of scores. Accurate measurement is predicated on the assumption that test scores represent an examinee's true proficiency with respect to the intended constructs. When the speed with which an examinee completes a test is not of interest, a restrictive time limit that does not allow examinees to exhibit their true proficiency can have negative consequences by introducing construct-irrelevant variation into examinee performance. Even when purposefully measuring speed, an inadequately timed assessment can yield questionable or even invalid results if the degree to which speed affects scores is different from what is expected based on the construct definition. The potential for speed to threaten the validity of scores has been referred to in the literature as *test speededness*.

This chapter presents a historical overview of the testing literature that exemplifies the theoretical and operational evolution of test speededness. As will be shown, the definition of speededness has evolved throughout the history of measurement and to this day remains a debated topic. The current *Standards for Educational and Psychological Testing* provide a framework for conceptualizing test speededness as the "extent to which test takers' scores depend on the rate at which work is performed as well as on the correctness of the responses" (AERA, APA, NCME, p. 223). In other words, speededness occurs when the allotted testing time influences examinee performance such that both speed and the construct of interest contribute to score variation. Several comprehensive literature reviews have summarized different aspects of the relationship between timing and testing (e.g., Lu & Sireci, 2007; Morrison, 1960; Schnipke & Scrams, 2002). This chapter presents a historical overview that focuses on how the concept of speededness evolved and how this evolution in conceptualization has influenced the methods that practitioners have used, and are now using, for evaluations of speededness. By describing how the field arrived at current philosophies and exploring the issues that still remain unaddressed, this brief historical review intends to serve as a foundation for the subsequent chapters within this book.

The Early Years: Speed and Ability as Interchangeable Measures

As the scientific study of testing burgeoned after World War I, initial theories posited that speed would not influence response quality independent of the intended construct (Spearman, 1927). Though practitioners recognized that speed and proficiency were conceptually distinct, the prevailing theory presumed that the high correlation between the two traits made them indistinguishable from a measurement perspective (Davidson & Carroll, 1945). In other words, timing could not introduce construct-irrelevant variance because speed was interchangeable with the construct of interest. Some context of the testing era is helpful to understand the logic in this theory. It is axiomatic that numeric scores, such as number correct, will decrease when examinees lack sufficient time to consider all items. However, test scores in this era were predominately used to rank-order examinees. Although total scores can differ substantially under different time limits, rank order would stay comparable if speed and proficiency correlated near perfectly (see Ruch & Koerth, 1923).

There was also an empirical basis for considering the evolution of speed and proficiency as interchangeable. To elaborate on this work, we must distinguish between speed tests and power tests, concepts formalized by Gulliksen in 1950 but used colloquially prior to Gulliksen's work. A pure speed test is one that is intended to evaluate how quickly an examinee can complete a set of test items within a fixed period of time. As such, speed tests are designed to have strict time limits and to include items of such ease that examinees can respond to all items correctly. Scores on speed tests then reflect the number of items responded to within the time limit and provide an indication of the speed and accuracy with which an examinee processes information. In contrast, pure power tests have no time limits and contain items of varying difficulty to capture the range of proficiency on the construct(s) of interest; scores on these tests reflect the number of items examinees answer correctly out of all items and are used to evaluate ability apart from the speed with which questions are answered. The distinction between pure speed and power tests is primarily theoretical. Many educational examinations function as a mixture of both power and speed tests, intending to primarily measure the construct of interest (i.e., power), but also containing a speed component resulting from time limits that are imposed to address practical constraints (Lu & Sireci, 2007; Chapter 3, this volume). Although theoretical in nature, the concepts of speed and power tests served as a foundation for the methodological developments throughout the evolution of speededness.

Restating Spearman's theory in these terms, rank order should be consistent whether an examination is administered as a speed or a power test. The belief that speed served as a proxy for cognitive ability partially stemmed from research in the 1920s and 1930s that investigated the relationship between scores from tests taken under both speed and power conditions. This research generally involved having examinees take a timed examination with a pencil; when the time limit was reached, they then finished taking the test using a different colored pencil or pen so that scores under both speed and power conditions could be distinguished (e.g., Paterson & Tinker, 1930; Peak & Boring, 1926; Ruch & Koerth, 1923). The empirical evidence indicated that scores under the two conditions were highly correlated. For example, Ruch and Koerth (1923) administered the aforementioned Alpha Army examination to 122 examinees under two timed conditions and a power condition, and multicolored pencils were used to capture response markings under the different conditions. Examinees first were given the standard amount of time suggested by the testing manual to respond to questions using a black pencil (single time). After the first time limit expired, examinees were provided the same amount of time to continue or revise answers using a blue pencil (double time), and after that time limit expired they switched to a red pencil to complete or change responses under an untimed period (untimed). Results indicated that rank ordering remained consistent—single to double time total scores correlated at 0.966 and single to untimed total scores correlated at 0.945—and therefore seemed to support the comparability between speed and accuracy.

Distinctions between Speed and Power

Taken at face value, Spearman's philosophy implies that time limits could be applied capriciously without consequence to validity (Morrison, 1960). As the study of mental testing matured, and likely motivated by the implication of Spearman's theory for practice, empirical research began to contradict the interchangeability of time and proficiency (Baxter, 1941; Davidson & Carroll, 1945). Davidson and Carroll provided a strong theoretical and empirical critique of this accepted practice. The authors expressed strong beliefs that scores from tests administered under time limits—particularly restrictive limits—reflected a mixture of examinees' knowledge and rate. This led the authors to claim, "the indiscriminate use of time-limit scores is one of the more unfortunate characteristics of current psychological testing …" (p. 411). Davidson and Carroll first criticized the established method of correlating scores from timed and untimed administrations of the same examination because the untimed score reflects a combination of the timed component, responses to the unreached items, and any answer changes made by the examinee. As the timed scores represent a part of the total untimed score, this method spuriously inflates correlations. The problems with this approach were exacerbated when the timed condition allowed examinees to reach the vast majority of the items. In this situation, the timed scores would almost fully reflect the final untimed scores (and the two necessarily would be highly correlated).

The authors followed up their methodological critique with an empirical study focusing on establishing a distinction between speed and knowledge. Utilizing various sections from a revised Alpha Army and several other examinations measuring a number of different constructs, the authors captured responses from examinees under timed and untimed conditions. They also collected data on the time it took each examinee to finish the exam after the time limit expired. A factor analysis found that scores from the untimed administration and completion speed loaded on separate orthogonal factors representing power and speed, respectively. Moreover, scores from the timed administration loaded on both the power factor and the speed factor, indicating that timed scores represented a mixture of both factors. Stated more concretely, time limits introduced variation in score performance unrelated to the construct of interest.

Around this time, the importance of time in testing became a focus of more sophisticated empirical treatments. Studies reinforced the factorially complex nature of scores produced under time duress for a variety of constructs (Myers, 1952). Mollenkopf (1950a, 1950b) explicitly showed that rank order could be influenced by time limits. Other investigations into the speed/accuracy relationship for a variety of constructs concluded that faster examinees were not always the most accurate (Tate, 1948). Rather, rate of work appeared to be a consistent examinee trait across tasks, leading several researchers to posit that speed was more influenced by individual characteristics than by proficiency on the construct (Himmelweit, 1946; Kennedy, 1930; Tate, 1948). Researchers also began to use speed as a correlate, finding that completion time could explain additional variation above proficiency in external criteria (Lord, 1956). Literature focused on the relationship between time and testing outcomes represents an important body of work that continues to be investigated. An expanded treatment of this literature can be found in Chapter 5. These and other studies accumulated considerable evidence demonstrating the multifaceted effect that speed can have on test score validity. Due to the growing body of evidence, measurement researchers rejected the belief that speed and ability could be considered practically equivalent. Despite this understanding, Morrison (1960) expressed frustration with test developers who continued a nonchalant approach to setting time limits without empirical rationale and failed to acknowledge the threat to validity posed by inappropriate time limits.

Early Developments in Statistical Quantification of Speededness

Gullisken, Cronbach, and Beyond

With mounting evidence that speed could introduce construct-irrelevant variation into the measures, and perhaps with some frustration regarding arbitrarily set time limits, researchers began developing statistical indices to quantify speededness. Gulliksen (1950) led this advancement in his treatise, *The Theory of Mental Test Scores*. To quantify speededness, Gulliksen utilized the following characteristics from examinee response patterns:

C = Number of items responded to correctly

W = Number of items responded to incorrectly

U = Number of unattempted items at the end of the test (assumed to be not reached within the time limit)

Note, although Gulliksen explicitly assumed no examinees omitted items after consideration (1950, p. 230), later authors made a distinction between omitted items and unattempted items (e.g., Rindler, 1979). In this distinction, omitted items were assumed when there was no response to an item but there *were* responses to the surrounding items. In contrast, unattempted items were considered to be unreached by the examinee due to speed; these manifested as unmarked items at the end of the test.

It follows that an examinee's total number incorrect, referred to as error score, corresponds to $E = W + U$, where the variance of E equals:

$$\mathrm{S}_E^2 = \mathrm{S}_W^2 + \mathrm{S}_U^2 + 2r_{W,U} S_W S_{U,} \tag{1.1}$$

where S^2 represents the variance of each respective term and $2r_{W,U} S_W S_U$ represents the covariance between incorrect items and unattempted items. Gulliksen proposed that test speededness could be measured through the ratio of standard deviations of U (unattempted items) over E (error score) across examinees: S_U / S_E. The theoretical foundation for this index stems from the definition of pure speed and pure power tests, which were formulized by Gulliksen in the same chapter. In a pure power test, the unlimited time allows examinees to attempt all items, meaning that *U* will be zero for all examinees. As there is no variation in *U* for a power test, the *U* variance component (S_U^2) and the covariance component ($2r_{W,U}S_W S_U$) drop out of equation 1.1, leaving the error score variance equal to the variance of number wrong ($S_E^2 = S_W^2$). Thus, a pure power test will logically yield a test speededness ratio of $0/S_E = 0$. In a pure speed test, *W* will be zero for all examinees as any item reached will be correct, resulting in $S_E^2 = S_U^2$ and a speededness index of $S_U / S_E = 1$. Gulliksen recommended that practitioners interpret ratios below 0.10 as indicating that the test primarily measures power (1950, p. 241). In other words, an S_U / S_E ratio above 0.10 indicates that speed affected variance in scores to a nontrivial degree.

Unfortunately, the logical parsimony of this index becomes muddled in practice. As has been discussed, examinations nearly always reflect an amalgamation of speed and power. When speed and power both contribute to performance, the correlation of *U* and *W* affects the error score variance through the covariance component. A negative correlation can lead to S_U^2 being greater than S_E^2, yielding a speededness ratio greater than 1.0. Gulliksen suggested interpreting both the S_U / S_E and S_W / S_E ratios concurrently to address this complexity, where S_W / S_E can be thought of as the proportion of power contributing to total error variance. This works when one ratio is close to zero, as the other ratio will be close to a value of one and lead to the same substantive conclusion. However, this symmetry quickly diminishes. Gulliksen himself showed that a S_U / S_E ratio of 0.75—indicating considerable speededness—could lead to a S_W / S_E ratio as low as 0.25—also supporting speededness—or as high as 1.75, strongly suggesting that power influences variation in scores. Clearly, the contradictory evidence each ratio can provide, along with the capability for ratios to exceed one, makes inferences regarding speededness from these two ratios incredibly challenging (Rindler, 1979).

In a paper primarily discussing corrections for the spuriously high reliability estimates obtained from speeded tests, Cronbach and Warrington (1951) contributed a speededness coefficient (τ) based on two administrations of parallel forms (A and B) under timed (*t*) and untimed power (*p*) conditions:

$$\tau = 1 - \frac{r_{A_t B_p} * r_{A_p B_t}}{r_{A_t B_t} * r_{A_p B_p}}. \tag{1.2}$$

The value yielded by τ reflects the proportion of reliable score variance in the power condition explainable by scores in the timed condition. Because this is a correlation-based measure, it will only capture rank-order differences among examinees. This corresponds with the authors' explicit definition of speeded tests, "A test is completely unspeeded when no subject's standing would be altered if he were given additional time" (p. 184), which aligns with the early speededness conceptualizations. The index suffers some administrative complexity, as it requires the same examinees to take two versions of the same exam twice to capture the timed and power conditions. Although this can be somewhat mitigated by using split-halves of the same form, the administration required to estimate this index would seem highly irregular to the examinee and would be resource intensive for test developers. Perhaps for these reasons, the multiple administration methods for quantifying speededness would not be widely used until slightly modified versions were employed in experimental studies as described in Chapter 5.

This brief tangent describes the challenge that deriving statistical indices related to speed presented in this era of paper-and-pencil testing. This illustration is intended to provide insight into the complexity of such statistical derivations and to highlight the need for methods that were easier to apply in practice at the time. The following list paraphrases the steps for obtaining one of the lower-bound reliability estimate corrections claimed by Cronbach and Warrington as "not involved" (which might be translated to mean "straightforward" or "easy"; 1951, p. 175).

1. Determine number of items finished for each person.
2. Create a frequency distribution of examinees at each number finished and compute the variance of items finished.
3. Mark errors on the examinee answer sheet with a distinguishing color and square the average score on the last two completed items for each examinee.
4. Consider N0, all examinees who finish all items, and N1 all examinees who finish all but the last item, compute the squared average score on the last two items for N0 and multiply by N1/N0. Add the squared average to this value.
5. Enter the obtained values along with the variance of total test scores to compute a lower bound reliability estimate for single-administration timed exams.

Despite the fact that these researchers displayed remarkable ingenuity in utilizing the dearth of timing information available in paper-and-pencil-based large-scale tests, it is clear that the complexity of computation for the standard practitioner should not be understated, as it presumably resulted in these methods being met with relative indifference in practice.

In the following years, several other measures to quantify the magnitude of speededness were proposed. Helmstadter and Ortmeyer (1953) described two additional techniques that relied on performance from speed (timed) and power (untimed) administrations of the same exam. The first compared the frequency of incorrect answers for each individual item between the speed and power conditions, where considerably more correct responses within the power condition indicated that the time limit influenced scores. The second involved subtracting the mean error score (*E* from Gulliksen's formula) on the power administration from the mean error score on the speed administration. A large difference from zero would indicate that the test is predominately measuring either speed (if negative) or power (if positive). In an attempt to ease computation and interpretation of speededness indices, Stafford (1971) proposed an index called the Speededness Quotient (SQ) that was similar to Gulliksen's ratio. This measure only required calculation of the frequencies *U* and *E*, which then were summed across examinees and divided to obtain the SQ ratio:

$$SQ = \frac{\sum U}{\sum E}. \tag{1.3}$$

Thus, SQ reflects the proportion of all incorrect responses that were unattempted. Following Gulliksen's logic, an SQ value near zero indicates that the total error scores were composed primarily of incorrect responses and, therefore, that speed had little effect on scores. In contrast, an SQ value near one indicates that nearly all errors resulted from unattempted items and reflects a speeded exam. Despite the continuing work to develop measures that quantify the magnitude of speededness, a review of the literature provides little evidence that these statistical estimates were used for the purpose of evaluating the impact of time limits. As noted by Donlon (1973) 20 years after Gulliksen proposed the variance-based ratio, "No single technique for characterizing test speediness is widely established" (p. 3). Given subsequent developments, however, it can be presumed that the field was awaiting more computationally simplistic methods with clear guidance on interpretation. These methods reached operational testing when Swineford described what became the seminal guidelines on test speededness (1956, 1974).

The Reign of Swineford

In the mid-20th century, there was a general disinterest in using statistics to evaluate speededness; this disinterest was attributed to several distinct factors: (1) difficulty of interpretation, (2) complexity of estimation, and (3) resource-intensive administration procedures (Morrison, 1960; Rindler, 1979; Stafford, 1971). In the 1956 Educational Testing Service (ETS) technical manual for test users, Swineford remedied these issues by defining a rule-of-thumb-based approach for classifying a test administration as speeded or unspeeded. Under these guidelines, a test was considered unspeeded when two conditions were met:

1. All examinees reached at least 75% of the items.
2. At least 80% of the examinees reached all of the items.

Although these criteria were admittedly arbitrary (1956, 1974), the Swineford guidelines alleviated the burden on test users by offering an elegantly simple index that required only computation of examinee item completion counts and a dichotomous determination of speededness (or lack thereof) based on that computation. It therefore should come as no surprise that the Swineford guidelines, despite the author's own caution regarding the arbitrary criteria, gained acceptance as the standard for evaluating test speededness.

The rate at which the guidelines gained popularity is difficult to determine. Swineford noted that the second condition—80% of examinees completing all items—was common in 1949 (prior to the guidelines being formalized in the 1956-published technical manual; see Swineford, 1949 and Chapter 5, this volume). However, Morrison did not refer to the Swineford guidelines in his comprehensive 1960 literature review. By the 1970s, authors discussed the Swineford guidelines alongside other metrics of speededness (Rindler, 1979) and used the rules as criteria for determining speededness (Evans & Reilly, 1972). Explicit and implicit use of the Swineford guidelines can be seen in testing manuals that were published in the ensuing years, including some that were produced over 50 years after the guidelines were formalized (e.g., Data Recognition Corporation, 2018; Pearson, 2015). It can also be assumed that the guidelines influenced recommendations provided by the current *Standards for Educational and Psychological Testing* for evaluating test speededness, which state "at a minimum, test developers should examine the proportion of examinees who complete the entire test, as well as the proportion who fail to respond to (omit) individual test questions" (AERA, APA, NCME, 2014, Standard 4.14).

Flaws in Traditional Techniques

The single-administration methods used to evaluate speededness described in the previous sections, including the Swineford guidelines, rely on similar assumptions regarding examinee test-taking behavior in order to make inferences about speededness. Each method inherently assumes:

1. All marked items (*C and W*) represent a fully considered item at the pace required for the examinee to provide her or his best response.
2. Items left blank between marked items (included in *W* by some authors) must have been omitted for reasons unrelated to speed.
3. All, and only, unmarked items at the end of the examination (*U*) reflect items excluded because the time limit expired.

Thus, speededness manifests solely through unanswered items at the end of an examination. Although authors sometimes acknowledged these assumptions, they represent such an unrealistic pattern of behavior as to be highly questionable for use in practice (Rindler, 1979). Consider the following quote from Rindler's poignant critique of these methods,

> "These assumptions are somewhat questionable in theory and certainly violated in practice, but are nevertheless uniformly shared by methods conventionally employed to estimate speededness from single test administrations." (p. 265)

For these assumptions to hold, an examinee must never rush to complete items or guess randomly as time is expiring. This strategy would consequently lower the examinee's expected score in tests that do not penalize for incorrect responses, as no response guarantees an incorrect score, whereas a rushed answer or a guess has some probability of being correct. Thus, exams where scores have serious consequences for the examinees will likely yield near 100% response rates regardless of the time limit and be classified as unspeeded by conventional guidelines. In addition, the distinction between omits and unattempted items is nearly impossible to reconcile. Examinees may skip complex items expecting to return to them at a later point, a strategy endorsed by test developers (College Board, 2019). If the examinees are unable to return to these items before the exam ends, the omitted items throughout the test do reflect speededness. Essentially, these assumptions required examinees to blissfully ignore their remaining time and follow test-taking strategies that would lower their expected performance (see Lu & Sireci, 2007 and Chapter 5, this volume). Rindler concluded that the traditional single-administration measures offered no information regarding the influence of time on scores.

A Shift toward Experimental Design and Psychometric Models

Although Rindler's critiques somewhat prophetically predicted the evolution of speededness indices to consider guessing and rapid responses, the difficulties in quantifying speededness persisted throughout the 1970s and 1980s. With the increasingly apparent flaws in available methods and a lack of data for making reasonable inferences, researchers turned their focus to experimental studies that investigated different aspects of speededness. Much of this work involved randomly assigning examinees to tests of differing lengths or time limits in order to address the methodological problem associated with comparing an untimed condition to a subsumed timed condition. These studies investigated the impact of time limits on performance for various different demographic subgroups such as gender (Lawrence, 1993; Wild, Durso, & Rubin, 1982), ethnicity (Evans & Reilly, 1972, 1973; Lawrence, 1993; Wright, 1984), and students with disabilities (Munger & Loyd, 1991). The initial literature consistently indicated that differing levels of time did not alter the score differences between subgroups, but further studies revealed a more intricate relationship between speed and different subgroups (see Bridgeman, 1998; Lawrence, 1993; and Chapters 5 and 7, this volume.

Another area of focus during this time period was the development of statistical models that could theoretically incorporate response times as a parameter to explain or predict performance (e.g., Thissen, 1983) as well as early conceptualizations of models that could partition examinees into speeded and unspeeded classes (Bejar, 1985; Yamamoto, 1990). These lines of research contributed invaluably to the field's understanding of complex interactions between speed and examinee factors and served as the foundation for future statistical models addressing speed. However, in relation to conceptualizing and evaluating speededness, this era was largely a precursor to the rapid advances made with the proliferation of computer-based testing. Thus, readers are directed to Lu and Sireci (2007), Schnipke and Scrams (2002), and Chapter 5, this volume, for in-depth treatments on these studies.

The Revolutionary Influence of Computer-Based Testing on Assessment of Speededness

It is difficult to overstate the impact that advancements in computing and internet technology had on testing. With the reduced cost and increased efficiency of computing, the 1990s saw a transition from paper-and-pencil to computer-based tests (Mills, Potenza, Fremer, & Ward, 2002). Computing accessibility and power also made widespread application of sophisticated statistical models viable. The capability for large-scale computer-based testing specifically opened critical new avenues in evaluating test speededness. Perhaps most notably, computers could capture examinee response time on individual items. Computers also could track examinee actions throughout the exam. These features have made evaluating actions such as skipping items and returning to review items straightforward and have facilitated evaluations of speededness by providing a more detailed representation of examinee behavior. Returning to the deficiencies of measures based on unattempted items, response times can indicate whether examinees considered each marked item or rapidly answered items to finish the test. Response times and behavior patterns can also help determine whether unmarked items resulted from expired testing time or intentional omission.

Researchers quickly took advantage of the availability of this new information and began using it to assess speededness. In one of the earliest studies to use computer-based response times from a large-scale assessment, Schnipke (1995) used data from 7,218 examinees completing two analytical sections of a Graduate Record Examination form and analyzed response time distributions for individual items in order of the item presentation sequence. The patterns that were found showed that the assumptions required by traditional speededness indices did not reflect modern examinee testing behavior. Instead of omitting items remaining at the end of the exam, many examinees provided responses that were noticeably faster than the time required to consider an item fully. Schnipke termed these responses “rapid guessing,” and they were classified through visually identifying the item response time where the inferred distribution of rapid guesses and solution behaviors intersected. Supporting the notion that these rapid guesses did not represent examinee knowledge, the rapid guesses tended to yield average percent correct values near what would be expected by random selection of a response option. On the form examined by Schnipke, a few items elicited rapid guesses from as many as 20% of the examinees. Results also showed that the frequency of rapid guesses did not increase monotonically through the item sequence, indicating that examinees sometimes guessed rapidly on earlier items so that they could reach and consider items later in the form. In contrast, the Swineford guidelines implied that the sections were essentially unspeeded, with over 98% and 97% of examinees reaching 75% of the items and 100% and 96% being reached by 80% of examinees for the two sections, respectively. It is evident that traditional methods would have grossly underestimated the speededness of this exam by considering only unattempted items. Examinees clearly accounted for their remaining time and employed more complex response strategies to accommodate the situation, as opposed to leaving items blank. The results of these studies led Schnipke to develop a framework that would heavily influence future research and timing investigations.

In her 1995 framework, Schnipke conceptualized that examinees engage in one of two potential behaviors when responding to items: rapid guessing or solution behavior (where solution behavior reflects full consideration of an item; see also, Yamamoto, 1995). Evaluations of item response times under this conceptualization gained favor, particularly in the low-stakes assessment literature where rapid guesses typically result from low examinee motivation rather than time constraints (DeMars, 2007; Guo et al., 2016; Kong, Wise, & Bhola, 2007; Chapter 11, this volume). This perceived dichotomy of response behavior also served as the foundation for the development of statistical mixture models that classify examinees into groups representing rapid guessing and solution behavior. In an attempt to remove the influence of rapid guesses on item parameter estimates from item response theory (IRT) models, Yamamoto (1990, 1995) developed one of the earliest of these mixture models by identifying the point in the item sequence at which examinees switch from solution behavior to a rapid-guessing strategy. The proportion of examinees engaging in rapid guessing at certain item positions then can provide an indication of test speededness.

Whereas Yamamoto's model estimated behavior strategies using item responses to identify a consistent drop in examinee performance, Schnipke and Scrams (1997) directly modeled individual item response times to estimate the proportion of rapid-guessing and solution behavior for each item. Results indicated that the two-strategy model better predicted observed response times, particularly for items at the end of the examination where rapid guessing was more prevalent. Bolt, Cohen, and Wollack (2002) proposed a mixture Rasch IRT model with two classes (representing speeded and non-speeded examinees) based on item accuracy toward the end of test. The authors offered a novel take on this literature by examining the demographic characteristics of examinees in each class as an alternative for exploring potential subgroup differences. For the data in this study, speededness was unrelated to gender but was statistically associated with ethnicity, aligning with the consistently inconsistent results of studies exploring demographic relationships with speed (Chapter 5). Classification methods for estimating speededness appear to have gained some traction as researchers continue to propose new—or extend existing—models for assessing test speededness (e.g., Meyer, 2010; Shao, Li, & Cheng, 2016). Unsurprisingly, results of these studies consistently show that speededness can be severely underestimated when unattempted items are the only consideration.

Limitations with Rapid Response and Classification Models

The availability of response times and the feasibility of estimating complex models have inarguably been an aid to practitioners and researchers investigating test speededness. However, these methods have limitations when used to describe the nuances of examinee test-taking behavior. The classification of rapid guessing still requires assumptions to distinguish these responses from solution behavior. The field has no agreed-upon speed threshold—or a method to determine a threshold—for a pure guess. Numerous methods for establishing thresholds have been proposed, including visual inspection of response time distributions, static values (e.g., <5 seconds), or values conditioned on item characteristics (see Wise, 2017). More recently, Guo et al. (2016) suggested setting thresholds at the point when the cumulative probability of a correct response exceeds chance level. Of course, the selected threshold will influence any conclusions that can be drawn from the results.

More generally, recent studies suggest that the conceptualization of examinee behavior into two distinct classes oversimplifies actual examinee strategies in response to time limits (Harik et al., 2018; see also Chapter 6, this volume). As mentioned, rapid-guessing methods are particularly popular in low-stakes contexts. When the testing context offers examinees little incentive to maximize their scores, the potential lack of motivation makes complete rapid guessing a reasonable alternative. In contrast, rapid guessing should manifest only in extreme cases when scores have serious implications for examinees. In these cases, when an examinee perceives that she or he is running out of time, another rational strategy would be to alter pacing to balance the remaining time by responding quicker than the desired time for each item but still considering enough of each remaining item to inform a reasonable response. These responses would only partially represent examinee ability, but they may result in a higher expected score than rapidly guessing on a series of items near the end of an exam. Response times for examinees following this pacing strategy would likely fall above any rapid-guessing threshold, and any decrease in the probability of a correct response may not be dramatic enough to be captured by accuracy-based models (e.g., Yamamoto, 1995). In randomly assigning examinees to complete test sections with different item lengths, Harik et al. (2018) noticed patterns suggestive of this pacing change approach. Examinees with the longer test sections performed worse on the items in the same item position, even when the groups encountered the item early in the test section. Moreover, differences in examinee performance persisted when making the sections comparable in length by removing the items at the end of the longer section from calculations (e.g., removing the last 12 items from a 44-item condition to compare performance to a 32-item condition). It seems that examinees were cognizant of the time allotted throughout the exam and adjusted pacing accordingly, leading to a detrimental effect on performance across all items for those with longer sections. These results align with the speed-accuracy tradeoff discussed frequently in psychology and other fields; task accuracy tends to decrease as the task is completed more quickly (Heitz, 2014; Luce, 1986). However, the magnitude of both response time and performance decreases was largest toward the end of the examination, suggesting that the contemporary methods for evaluating speededness are likely sensitive to the most serious effects.

Effects of Speededness on Item and Test Properties

The discussion so far has focused on how speededness has been conceptualized and evaluated as a function of examinee behavior. It should be noted that speededness also negatively affects item and test characteristics. Presumably resulting from the increased accessibility of complex models, the proliferation of computer-based testing coincided with several studies indicating that IRT parameter estimates become biased under speeded conditions (Oshima, 1994; Schnipke, 1996, 1999; Wise & DeMars, 2006). To summarize the findings, speededness inflates item difficulty (b) and tends to inflate item discrimination (a) parameters, particularly for items near the end of the exam. Speededness can inflate, deflate, or have no effect on test information and reliability depending on multiple factors such as degree of speededness, item difficulty, and whether speeded responses are independent of each other (Hong & Cheng, 2018; Wise & DeMars, 2006). These findings parallel those that were found or theoretically suggested in early speededness studies (Cronbach & Warrington, 1951; Mollenkopf, 1950a) and again align with the logic of the speed-accuracy tradeoff (Luce, 1986). These biases, in addition to the contamination of scores, represent the potential outcomes of a more general issue when applying basic IRT models to speeded tests. IRT assumes that the measured trait is predominately unidimensional in nature. Under speeded conditions, responses no longer predominately reflect the construct of interest but instead reflect a mixture of speed and knowledge. Thus, the validity of both estimated scores and parameters is always in question with speeded examinations (Hambleton & Swaminathan, 1985).

Although these results may seem obvious, one potential corollary that perhaps has not received necessary attention occurs when item parameter estimates affected by speed are treated as known for other test development purposes. In his 2017 paper, van der Linden discussed two situations when this could occur. First, developers may potentially pretest items under different speededness conditions than their operational use. For example, if the pretest item's position is static and toward the end of the examination, the true item difficulty may be overestimated. Second, the differential bias in parameter estimates may hinder equating procedures if the item position has changed between the equated forms. The measureable impact of these issues has not been thoroughly addressed in the literature.

Incorporating Examinee Speed into Test Design

Given the effects that speededness can have on various aspects of examinee and test properties, recent literature has pushed for more direct evaluation and estimation of examinee-level speed components to inform critical test development tasks such as form building and time limit setting (e.g., van der Linden, 2006). Wim van der Linden has predominately led this work via a lognormal response time model (van der Linden, 2006, 2011a, 2011b, 2017). Comparable to a 2PL-IRT model, van der Linden's lognormal response time model predicts item response time distributions using an examinee speed parameter and two-item parameters. The examinee parameter characterizes an individual's level of pacing on the test. Thus, as suggested in earlier literature (Kennedy, 1930), rate of work is treated as an individual trait that can be estimated given appropriate information available from observed response times. The item parameters include time intensity, which describes the item's time demand, and a discrimination parameter that indicates the ability to differentiate between examinees with different pacing. In a series of empirical analyses (van der Linden, 2006), the model replicated observed response times well; this led the author to suggest several applications. Rudimentary uses would involve applying the model to examine response time distributions and parameters post hoc to evaluate the speededness of an administered test. Although appropriate, the model is likely overly complex to be applied for the purpose of post-administration speededness evaluations alone. Examinations of observed item response distributions would be expected to produce similar conclusions.

The primary benefits of van der Linden's model emerge in applications for test development operations of future test forms. When practitioners have estimated the response time model item parameters for a pool of items and can specify the expected mean and variance for an examinee population's speed parameter (which can be informed by previous calibrations), the model offers a wide variety of operational applications. Estimates can be used to inform the selection of a time limit by evaluating the distribution of predicted total test time relative to various time limits (van der Linden, 2011a). Theoretically, the parameters in the model can also be used to construct equally speeded test forms for a population to reduce potential bias (van der Linden, 2011b). van der Linden (2017) summarizes several other applications of the model to modern test design challenges. However, the benefits of this model-based approach to speededness do incur some drawbacks. This method is mathematically complex, particularly when compared to other methods for assessing speededness. This complexity may pose challenges for operational application and interpretability of the results. Furthermore, the strong statements discussed above require certain assumptions and constraints that may not hold in practice. It remains to be seen whether practitioners will widely utilize this model, but the statistical developments and applications proposed in these studies present opportunities for future exploration.

Current State and Future Directions

Theories regarding test speededness have evolved rapidly since the early 20th century; Table 1.1 provides a selected overview of some of the more central theories along with their associated evaluation methods and limitations. Once considered ignorable, speed now is recognized as a complex factor that can influence various aspects of test validity in numerous nuanced ways. Methods to determine the occurrence and degree of speededness also have continued to advance. Estimates of spurious correlations between timed and untimed administrations of the same test were replaced by statistical indices utilizing information about unattempted items, and experimental manipulations to explore the effect of speededness grew in popularity as the limitations of traditional methods became known. The advent of computer-based testing

Table 1.1 Selected historical approaches to conceptualize and evaluate timing

Year—Author	Theory	Evaluation Methods	Limitations
1927—Spearman	With generally adequate time limits, speed and proficiency reflect interchangeable measures of general intelligence due to high correlations among these two traits.	• No evaluations, per se. • Concept established through correlations between scores from a timed version of a form and scores from a form where examinees have unlimited time to finish the form and change answers.	• Initial studies used flawed methods that inflated correlations by correlating a part score (timed test) with the whole score (timed responses + untimed period for corrections). • Later studies found that speed can have various effects on examinee behavior.
1950—Gulliksen	Distinction between speed and power tests. A speed test contains items so easy examinees should never give the wrong answer and is scored by the number of items reached. A power test allows examinees to fully consider all items and is scored by number correct. Thus, comparing unattempted items to incorrect items provides an indication of how speed influences performance.	• Examined variation in unattempted items and incorrect items. • Most prominent method involved taking the ratio of standard deviation of unattempted items (S_u) and total error (S_e): S_u / S_e. • See also, Helmstadter & Ortmeyer, 1953, Stafford, 1971.	• Untenable assumptions regarding examinee behavior when under speeded conditions, such as no guessing on items.
1951—Cronbach and Warrington	Speededness as the extent that true standard scores in the group would change if more time was allowed.	• Proportion of variance in scores due to speed as measured by independently administered halves of two equivalent forms; one half under a timed condition, and the other half under an untimed condition.	• Test administration can be contrived, resulting in idiosyncratic examinee behavior. • Multiple administrations can also be resource intensive.
1956—Swineford	Following Gulliken's focus on unattempted items, examined proportion of examinees who finish exam and percent completing individual items.	• Rule of thumb considering a test unspeeded if either: 1. All examinees reached at least 75% of the items. 2. At least 80% of the examinees reached all of the items.	• Arbitrary cutoffs. • Assumes speededness only manifests at end of exam. • Similar to Gulliksen, makes unrealistic assumptions about how examinees behave when under time pressure.

(*Continued*)

Table 1.1 Selected historical approaches to conceptualize and evaluate timing (*Continued*)

Year—Author	Theory	Evaluation Methods	Limitations
1995—Schnipke	Speededness is the extent to which some examinees are disadvantaged by the time limit on a test relative to other examinees. (p. 4) Examinees will guess on items when running out of time to maximize their potential score.	• Proposed evaluating rapid guesses that reflect quick uninformed responses under time pressure to quantify proportion of examinees exhibiting solution behavior and rapid-guessing behavior per item. • See also Wise, 2017.	• No agreed-upon method to determine a rapid-response threshold. • Dichotomous classification between rapid guessing and solution behavior may simplify examinee behavior under time duress. For example, examinees may work slightly faster than the time needed to fully consider an item but longer than what would be considered a rapid guess.
2011—van der Linden (p. 185)	Speededness described as the quantified probability of running out of time for an examinee working at a specified pace.	• Statistical model that estimates examinee pacing and item time intensity, which can be used to make various time-related inferences.	• Increased model estimation and parameter interpretation complexity relative to other speededness evaluation methods. • Involves assumptions or constraints that may not be realistic in certain operational settings.

opened up a whole new world of exploration by providing practitioners with access to direct measures of examinee speed, and increased computing power has created access to complex mathematical models that incorporate these metrics. As technological capabilities expand, we likely will see that the field is only scratching the surface of methodological potential to assess speededness. Unfortunately, the fact remains that despite advances both in methods for evaluating speededness and in understanding the complex influence of speed on test validity, the attention paid to this topic in operational settings greatly underrepresents its importance.

One of the few consistent messages echoed throughout the history of timed testing is that time limits are predominately set based on convenience and resource constraints rather than supported by evidence related to test validity (Morrison, 1960; Rindler, 1979). It would be naïve to think that this practice will change significantly, as testing organizations always must consider the costs of seat time and item development. However, the critical influence that speed can have on test validity necessitates empirical evidence defending the selected time limits through evaluations of speededness. Computations of standard test evaluation indices are not sufficient (Cronbach & Warrington, 1951), nor are outdated rules of thumb regarding unattempted items. Yet the rationale behind time limit decisions is often obscured. Test publishers rarely appear to disseminate this information in technical manuals or other sources, a frustration noted by Mollenkopf (1960) over a half century ago.

Omission of information related to timing or speed may reflect the lack of treatment speededness has received in the *Standards for Educational and Psychological Testing* throughout time. The first edition of the *Standards*, termed Technical Recommendations for Psychological Test and Diagnostic Techniques (1954), contained two standards referencing timing. The primary standard, C 18.3, noted that practitioners should provide evidence regarding time limit effects on test scores and on correlations with external variables. The manual mentions no additional details on what data could be used to investigate these effects and considered this standard as "Very Desirable," the middle category on a three-point importance scale (between Desirable and Essential). Standard D 6.1 guided practitioners away from using split-half or analysis of variance reliability coefficients with time limit tests (likely stemming from Cronbach & Warrington's 1951 article). Despite the accumulating literature on speededness, the *Standards* (2014) continue to devote only minor coverage to the appropriateness of time limits. The one standard devoted to evaluating speed, 4.14, states:

> For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure. (p. 90)

Speed is indirectly addressed through standards discussing construct definition and test specifications (1.1 and 4.2), but it is surprising that a more targeted treatment is not provided given the pervasiveness of timed tests meant to approximate power tests (Lu & Sireci, 2007). It is hoped that the thoughtful contributions in this book such as guidance on considering speed throughout the test design process (see Chapter 3) and discussion of speed in relation to validity (see Chapter 2) spur more explicit and detailed treatments of speed in technical recommendations.

There are several specific areas where additional attention would benefit the field. To help measure the construct of interest appropriately, test takers with disabilities are commonly offered accommodations in the form of additional testing time. The amount of additional allotted time appears inconsistently and arbitrarily set across testing organizations, however, and the decisions that are made lack documented support from empirical evidence (see Chapter 4). Practitioners require guidance to answer vital questions such as how to determine the appropriate amount of time necessary to neither hinder nor advantage examinees receiving accommodations. The validity of time accommodations is further complicated when speed is an intended component of the measured construct.

Innovations in testing methods will require evaluation by test developers in order to fully understand the associated timing implications. The use of features such as interactive multimedia in test items will continue to increase as technology develops. These items typically require more time than traditional items (Jodoin, 2003; Qian, Woo, & Kim, 2017). These items also represent an equity concern if subgroups complete technology-enhanced items at different rates. For example, examinees with increased accessibility to similar technology as that used in the items may respond more quickly than those requiring time to become accustomed to the new features. Other methods gaining in popularity such as simulations and game-based testing each present similar challenges.

Though this review primarily focused on the situation where test developers consider speed as construct irrelevant, it should be noted that many of the discussed validity concerns are exacerbated when speed is a meaningful part of the construct. In these situations, Standard 4.14 requires test developers to defend the degree to which speed contributes to assessment of the construct of interest. As a corollary, the developers then must empirically support that the selected time limits yielded responses with the intended level of variation explained by speed and the measured construct across the population of examinees. With increasing calls for authentic assessments of behavior, it seems likely that speed will see an increased role in construct definitions.

In conclusion, this historical perspective has described the immense advances the measurement field has made in understanding the complexity of how speededness affects examinee behavior. In addition to more refined theories of speededness, the field continues to develop increasingly sophisticated methods that incorporate more meaningful data to evaluate the impact of time limits on scores. That being said, test speededness will continue to deserve attention from both researchers and practitioners. The issues described in this chapter likely represent only a small portion of the underexplored topics related to speededness. Recent literature suggests that assumptions made by predominant methods to evaluate speededness simplify actual examinee behavior (see Harik et al., 2018; Chapter 6, this volume). More generally, the resource constraints that necessitate time limits will inevitably persist. Thus, speed will continue to present a potential source of construct-irrelevant variance that threatens the validity of inferences made from test scores; practitioners therefore must determine time limits judiciously, based on construct definitions and empirical evidence. If there has been a consistent implication throughout the history of this topic, it is that measurement professionals must improve their guidance to practitioners regarding the importance of both defining speed in relation to the measured construct and evaluating the influence of speed on scores. Given the challenges to educational measurement posed by test speededness, the evolution in understanding and practice that has developed over the past century may pale in comparison to what the future will bring.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. New York, NY: Guilford Publications.

Baxter, B. (1941). An experimental analysis of the contributions of speed and level in an intelligence test. *Journal of Educational Psychology*, *32*, 285.

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Service.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348.

Bridgeman, B. (1998). Fairness in computer-based testing: What we know and what we need to know. In Shilpi Niogi (Ed.), *New direction in assessment for higher education: Fairness, access, multiculturalism, & equity* (The GRE, FAME Report Series, Vol. 2, pp. 4–11). Princeton, NJ: Educational Testing Service.

Bridgeman, B. (2020). Relationship between testing time and testing outcomes. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 59–72). Abingdon: Routledge.

College Board. (2019). Doing your best on the SAT. Retrieved from https://collegereadiness.collegeboard.org/pdf/official-sat-study-guide-read-keys-doing-your-best.pdf

Cronbach, L. J., & Warrington, W. G. (1951). Time limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, *14*, 167–188.

Data Recognition Corporation. (2018). *Wisconsin Forward Exam Technical Report*. Retrieved from https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/Forward_Exam_Tech_Report_2018.pdf

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis. *Educational and Psychological Measurement*, *5*, 411–427.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*, 23–45.

Donlon, T. F. (1973, November). Establishing appropriate time limits for tests. Paper presented at the Northeast Educational Research Association, Ellenville, New York.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*, 123–131.

Evans, F. R., & Reilly, R. R. (1973). A study of test speededness as a potential source of bias in the quantitative score of the admission test for graduate study in business. *Research in Higher Education*, *1*, 173–183.

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications*. Needham Heights, MA: Allyn & Bacon.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: John Wiley and Sons.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*, 173–183.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Academic.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., … Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement*, *55*, 308–327.

Harik, P., Feinberg, R. A., & Clauser, B. E. (2020). How examinees use time: Examples from a medical licensing examination. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 73–89). Abingdon: Routledge.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, Article 150, 1–19.

Helmstadter, G. C., & Ortmeyer, D. H. (1953). Some techniques for determining the relative magnitude of speed and power components of a test. *Educational and Psychological Measurement*, *13*, 280–287.

Himmelweit, H. T. (1946). Speed and accuracy of work as related to temperament. *British Journal of Psychology*, *36*, 132–144.

Hong, M. R., & Cheng, Y. (2018). Clarifying the effect of test speededness. *Applied Psychological Measurement, 43*, 611–623.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, *40*, 1–15.

Kane, M. (2020). The impact of time limits and timing information on validity. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 19–31). Abingdon: Routledge.

Kennedy, M. (1930). Speed as a personality trait. *The Journal of Social Psychology*, *1*, 286–299.

Kong, X., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapidguessing behavior. *Educational and Psychological Measurement*, *67*, 606–619.

Lawrence, I. M. (1993). *The effect of test speededness on subgroup performance* (ETS-RR-93-49). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika*, *21*, 31–50.

Lovett, B. J. (2020). Extended time testing accommodations for students with disabilities: Impact on score meaning and construct representation. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 47–58). Abingdon: Routledge.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*, 29–37.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Margolis, M. J., von Davier, M., & Clauser, B. E. (2020). Timing considerations in performance assessments. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 90–103). Abingdon: Routledge.

Martin, W. A. (1870). Competitive examinations in China. *The North American Review*, *111*, 62–77.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521–538.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.). (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mollenkopf, W. G. (1950a). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, *15*, 291–315.

Mollenkopf, W. G. (1950b). Slow, but how sure? *College Board Review*, *11*, 147–51.

Mollenkopf, W. G. (1960). Time limits and the behavior of test takers. *Educational and Psychological Measurement*, *20*, 223–230.

Morrison, E. J. (1960). On test variance and the dimensions of the measurement situation. *Educational and Psychological Measurement*, *20*, 231–250.

Munger, G. F., & Loyd, B. H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *The Journal of Educational Research*, *85*, 53–57.

Myers, C. (1952). The factorial composition and validity of differently speeded tests. *Psychometrika*, *17*, 347–352.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219.

Paterson, D. G., and Tinker, M. A. (1930). Time-limit vs. work-limit methods. *American Journal of Psychology*, *17*, 101–104.

Peak, H., & Boring, E. G. (1926). The factor of speed in intelligence. *Journal of Experimental Psychology*, *9*, 71–94.

Pearson. (2015). *New York State Testing Program 2015: English Language Arts and Mathematics Grades 3–8 Technical Report*. Retrieved from http://www.p12.nysed.gov/assessment/reports/ei/tr38-15w.pdf

Qian, H., Woo, A., & Kim, D. (2017). Exploring the psychometric properties of innovative items in computerized adaptive testing. In J. Hong & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 97–118). Charlotte, NC: Information Age Publishing.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*, 261–270.

Ruch, G. M., & Koerth, W. (1923). "Power" vs." Speed" in Army Alpha. *Journal of Educational Psychology*, *14*, 193–208.

Schnipke, D. L. (1995, April). Assessing speededness in computer-based tests using item response times. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schnipke, D. L. (1996, April). How contaminated by guessing are item-parameter estimates and what can be done about it? Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Schnipke, D. L. (1999). *The influence of speededness on item-parameter estimation* (Computerized Testing Report No. 96-07). Princeton, NJ: Law School Admission Council.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*, 1118–1141.

Sireci, S. G., & Botha, S. M. (2020). Timing considerations in test development and administration. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 32–46). Abingdon: Routledge.

Spearman, C. (1927). *The abilities of man*. Oxford, England: Macmillan.

Stafford, R. E. (1971). The speed quotient: A new descriptive statistic for tests. *Journal of Educational Measurement*, *8*, 275–278.

Swineford, F. (1949). *Law school admissions Test-WLS* (ETS-RB-49-12). Princeton, NJ: Educational Testing Service.

Swineford, F. (1956). *Technical manual for users of test analyses* (SR-56-42). Princeton, NJ: Educational Testing Service.

Swineford, F. (1974). *The test analysis manual* (SR-74-06). Princeton, NJ: Educational Testing Service.

Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educational and Psychological Measurement*, *8*, 353–374.

Thissen, D. (1983). Timed testing: An approach using item response testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179–203). New York, NY: Academic.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2011a). Setting time limits on tests. *Applied Psychological Measurement*, *35*, 183–199.

van der Linden, W. J. (2011b). Test design and speededness. *Journal of Educational Measurement*, *48*, 44–60.

van der Linden, W. J. (Ed.). (2017). Test speededness and time limits. In *Handbook of item response theory, volume three: Applications*. New York, NY: Chapman and Hall/CRC.

Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement*, *19*, 19–28.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretations, and implications. *Educational Measurement: Issues and Practice*, *36*, 52–61.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150–164). Abingdon: Routledge.

Wright, T. (1984). *The effects of increased time-limits on a college-level achievement test* (Research Report No. 84-12). Miami, FL: Miami-Dade Community College.

Yamamoto, K. (1990). *HYBIL: A computer program to estimate HYBRID model parameters*. Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL Tech. Rep. No. TR-10). Princeton, NJ: Educational Testing Service.

2

The Impact of Time Limits and Timing Information on Validity

Michael Kane

Time limits play a significant role in the standardization of assessments. Standardization is useful in promoting procedural fairness and controlling random errors, but any aspect of standardization, including time limits, can have an impact on validity by supporting or undermining the proposed interpretation and use of the scores.

For performance assessments, time constraints that are similar to those associated with the kind of performance being assessed can make the performance tasks more realistic and thereby tend to support the validity of score interpretations in terms of proficiency in that kind of performance. In addition, for speed tests designed to assess how quickly test takers can perform certain kinds of tasks, time constraints are an essential element in test design.

The most complicated and contentious issues in designing and evaluating time limits occur in the context of testing in which the tasks to be performed are highly stylized (e.g., objective items, essays) and are not intended to be especially speeded. To the extent that the time limits are more or less arbitrary and tight enough that some students cannot comfortably complete the test in the time allowed, they can add both systematic and random errors to the scores and thereby limit validity. The systematic errors—which can occur at the population level, at a group level, or at an individual level—can be particularly troublesome. Time limits that have differential construct-irrelevant effects across groups can be considered a source of bias.

The errors generated by time limits can often be controlled to some extent by loosening the time constraints or decreasing the number or complexity of the tasks on a test. In some cases, it also may be possible to statistically adjust scores in ways that eliminate or reduce the more general systematic errors introduced by time limits (see Chapter 8), but generally it will not be possible to address differential effects specific to particular test takers. Alternatively, where cut scores are used to categorize test takers, it may be possible to control the general time-limit effect (for the population as a whole) by adjusting the cut scores. Test preparation that makes prospective test takers aware of the time limits and their implications also may help to limit the impact of the time limits.

The impact of time limits will depend on how tight they are, on the nature of the tasks, on whether the score scale is criterion or norm referenced, and on the extent to which time is a significant element in the construct being assessed. In addition to the direct impact of time limits on a test taker's performance, timing data (e.g., the time test takers spend on each item) also can provide information about issues such as a test taker's motivation and possible cheating.

Power versus Speed

A distinction can be drawn between attributes defined mainly in terms of speed and those defined mainly in terms of level of performance (see Chapters 1 and 7). For *speed tests*, higher scores are intended to reflect speed in performing some kind of task (e.g., typing). For *power tests*, higher scores are intended to reflect ability to perform more demanding tasks. This distinction is, of course, more of a continuum than a sharp categorization. The interpretations of scores on most educational tests are basically in terms of what the test takers can do, the kinds of tasks that they can successfully complete. That is, the results are mainly interpreted in terms of power, but they also have an element of speed in that the test taker is expected to complete the tasks in some reasonable length of time. If a student takes 2 hours to solve a quadratic equation, we might suspect that he or she is not using algebra. Although the time taken to complete a task is typically of some concern, especially if a test taker's performance is exceptionally slow, most of the attributes assessed in education are defined in terms of power, and any impact of time limits therefore typically would be considered a construct-irrelevant effect.

Although the distinction between score interpretations that involve an explicit speed component and those for which the speed component is implicit and not a major concern is a bit fuzzy, the conceptual distinction is useful in talking about the impact of time limits on the validity of the proposed interpretation of the test scores. Tests of clerical speed and accuracy are the classic examples of speed tests. The kinds of tasks under consideration are fairly limited and specific and it is assumed that test takers can perform the tasks (e.g., assembling a piece of equipment); the question of interest is *how quickly* can the test takers perform the tasks.

At the other extreme are tests that include a range of tasks that vary in conceptual difficulty from relatively simple to more complex (e.g., from arithmetic to advanced algebra or trigonometry). Similarly, a reading test might involve a range of passages of increasing complexity or questions about the passages of increasing difficulty. Traditionally, to the extent that it is feasible, the tasks in such power tests have been arranged, at least roughly, in terms of difficulty from the easiest to the most challenging so that students with relatively low levels of competence in the domain being assessed get a chance to indicate what they know without running out of time. For a power test, the time limits are designed to make it likely that most test takers have enough time to get to the end of the test. Currently, there is considerable interest in power tests based on learning progressions (Shepard, 2018).

We can also briefly consider a third category of tests in which speed can play a substantial and legitimate role. A performance test could be designed to include a strong speed component if the performance domain being evaluated requires timely actions, even if speed is not explicitly included as a separate component in descriptions of the proposed interpretation. An airline pilot landing a plane, a surgeon performing an operation, and a lawyer participating in a trial all have to make decisions and take actions in a timely way to be successful. In practice, most activities have to be completed within some time constraints, but many important activities require quick responses to evolving situations (Fitzpatrick & Morrison, 1971). In these contexts, the performances typically are intrinsically time sensitive in that they have to be fast enough for the purpose at hand, but beyond that threshold faster is not necessarily better. The performance is evaluated in terms of how effective it is in achieving some goal, and delays that interfere with the effectiveness of the performance count against the quality of the performance.

To be considered "authentic" or highfidelity, performance tests (or high-fidelity simulations) would need to reflect the time demands of the relevant real-world settings in some way and to some extent (Kane, Crooks, & Cohen, 1999). I will refer to such tests as *time-sensitive performance tests* to emphasize that the judged quality of the performance depends in part on its timeliness.

In some performance assessments, timing issues can be considered an integral part of the construct being assessed, and in these cases the core issue is whether the time constraints are appropriate given the construct being assessed (see Chapters 3, 9, and 10). For many kinds of problem solving, a faster accurate performance is considered better than a slower accurate performance.

For speed tests, the construct of interest is defined in terms of speed, and therefore time limits are not a source of error. For time-sensitive performance tests, time constraints that correspond, at least roughly, to those that occur in the real-world contexts in which the time-sensitive performances typically occur can contribute to the fidelity of the simulation. If the time limits for a time-sensitive performance test are quite different from those in real-world practice (i.e., they are either more stringent or more relaxed), the fidelity of the performance can be questioned. The time constraints inherent in a real-world situation (landing an airplane, treating a stroke patient, deciding whether to object in court) may be dictated within fairly narrow limits by the situation.

For standardized tests (e.g., objective tests, constructed-response tests, low-fidelity simulations), time limits are generally an artificial aspect of standardization and potentially are sources of error, especially if the time limits are such that many test takers do not have a chance to indicate their level of achievement by completing all of the tasks that they could perform if there were no time limit. The time limits for a real-world performance may be dictated by the nature of the performance and its context, but the time required to answer a question about the activity may be quite different. So the implications of time limits for the validity of a proposed interpretation and use are of particular concern for standardized power tests. As suggested by the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014):

> Standard 4.14: For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure. (p. 90)

To the extent that the scores on a power test have a substantial speed component, the validity of the proposed interpretation can be questioned, at least for those test takers whose scores seem to be affected by the speed component.

As noted above, there are serious timing issues for speed tests and for time-sensitive performance tests (see Chapters 1, 7, and 9), but these timing issues are closely tied to the definition of the construct being measured and will not be discussed further in this chapter. I will focus on the implications of time limits for standardized power tests.

Validity

The validation of a proposed interpretation or use of test scores requires an evaluation of the plausibility of the interpretation and the reasonableness of the proposed use of the scores. It is the interpretations (and uses) of test scores that are validated, and they are validated by developing evidence that supports the claims being made and evidence that refutes challenges to these claims. If the evidence supports the interpretation and use well enough, the interpretation and use can be accepted; on the other hand, if, given the evidence, some alternative interpretation makes more sense, the interpretation and use will not be accepted.

An argument-based approach to validation specifies the proposed interpretation and use in terms of a chain, or network, of inferences leading from the observed test performances to the claims included in the interpretation and to the uses of the test scores (Kane, 2013). An *interpretation/use argument* (IUA) includes the inferences and assumptions inherent in the interpretation (and use) of scores, and a separate *validity argument* provides an evaluation of the IUA in terms of its completeness, coherence, and plausibility.

The kinds of evidence needed to evaluate the interpretation and use are contingent on the claims inherent in the interpretation and use. A relatively simple and direct interpretation would not need much evidence to be considered valid. For example, the results of a carefully developed performance assessment in which test takers are asked to perform some kind of task, and for which the only interpretation of the scores is in terms of how well individuals can perform the task under consideration, could be validated fairly simply. For test takers who performed well on most of the tasks presented to them and therefore have high scores, a claim that they can perform this kind of task well is certainly plausible. We might have some concerns about whether the tasks might have been too easy or that a test taker had prior exposure to the specific tasks included in the assessment, but in general the validity of this kind of interpretation of high scores on a performance test is relatively easy to justify.

For test takers with low scores on a performance test, a claim that they cannot perform this kind of task well is harder to justify because there may be plausible alternative explanations for the poor performance. In addition to concerns about whether the specific tasks included in the assessment might be too difficult, we would need to rule out the influence of construct-irrelevant factors that might interfere with a test taker's performance, including language difficulties, disabilities, health issues, lack of motivation, equipment failures, tight time limits, and so on. We don't routinely try to address all such potential threats to validity, but we need to do so if there is a reason to suspect that they might be interfering with a test taker's performance.

For this kind of simple interpretation, the positive case for the validity of the interpretation in terms of expected performance on tasks like those in the test is pretty straightforward, and most of the effort in validating score-based claims about performance will involve the evaluation of potential alternative interpretations (particularly, alternative explanations for poor performance). In other cases (particularly standardized, objective tests), making the positive case for the proposed interpretation and use may be more demanding, but in all cases, validation requires an evaluation of plausible counterclaims. Cronbach made this point particularly aptly:

> The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it. (Cronbach, 1980, p. 103)

To be effective, validation requires a critical evaluation of the claims being made.

For many potential challenges to the proposed interpretation and use, the question is not one of making a binary decision about whether the proposed interpretation is true or not but rather relates to the extent to which some source of construct-irrelevant variance or construct underrepresentation is likely to interfere with the proposed interpretation and use of the scores (Messick, 1989). Score interpretations are never perfectly precise, and the claims made are usually qualified in some way (Kane, 2013; Toulmin, 1958). For example, in describing a test taker's results on a performance assessment, we might use terms like "almost always" or "usually" to indicate the consistency with which the test taker is successful.

For quantitative score interpretations, the uncertainty in the scores as estimates of the attribute of interest can be quantified as estimated errors of measurement, or standard errors, and all sources of construct-irrelevant variance contribute to this error. The sources of error are essentially sources of variability that are not consistent with the intended interpretation of the scores or are not supposed to be of any significance (Kane, 2011). For example, variability in how raters apply a scoring rubric generally would be considered a source of error, because the intended interpretation focuses on the performance of the test takers and not on the performance of the raters. Similarly, any impact of time limits on test performance, and thereby on scores, would be a source of error on a power test, because the attribute of interest does not focus on speed to any significant extent.

Time Limits as a Source of Systematic and Random Errors

Tests used to make high-stakes decisions tend to be highly standardized in the kinds of tasks included, in the response modes, in the contexts in which the test is taken, and in the rubrics and procedures used for scoring. They also are administered under fixed time limits. Standardization promotes fairness and the appearance of fairness by subjecting all test takers to essentially the same challenges (Porter, 2003), and it helps to control errors of measurement by eliminating the irrelevant variability in scores that would result if different test takers had to perform under different conditions. That being said, standardization also introduces various kinds of systematic error, because even fixed testing conditions can have differential effects on test takers' performance (Kane, 1982).

If the time limits are tight enough that some test takers do not have time to complete tasks that they otherwise could complete successfully, the competence level of these test takers will be systematically underestimated. Other test takers may not suffer any disadvantage from the time limits, because they tend to work fast (or because they have practiced working quickly in preparing for the test). Chapter 5 provides a particularly interesting review of research on the effects of speededness on test scores and concludes that the impacts of time limits depend in complicated ways on the context, the severity of the time limits, and on test formats, and therefore that simple generalizations about the impact of time limits are not possible. Interestingly, he notes that it is not necessarily the lowest scoring test takers or the highest scoring test takers who are most affected by time limits.

I will refer to the differences between hypothetical unlimited-time scores and the corresponding time-limited scores as *time-limit errors* (TLEs). These TLEs do not generally have a zero mean, and to the extent that speed is an enduring characteristic, the TLEs can be correlated across test forms; they therefore are systematic errors rather than random errors (by definition, random errors have a mean of zero and are uncorrelated with each other and with other variables).

The average TLE for a *population*, the TLE_P, is a general systematic error for the time limit, the test, and the population. The average TLE for a *group* (e.g., racial or ethnic groups, students with a disability, gender) is a group-level systematic error, or TLE_G. To the extent that it is consistent across test administrations, the TLE for an *individual* test taker, the TLE_I, is a specific systematic error. As a practical matter, it is not generally possible to estimate TLEs for individual test takers. To the extent that an individual's speed in completing test tasks varies from one test administration to another, the errors associated with the speed of performance would be considered random (e.g., in the context of test-retest reliability), while more stable differences in speed would be systematic. Systematic errors tend to be more serious than random errors, because they do not cancel out over replications of the assessment.

As the comment following Standard 4.14 (AERA, APA, & NCME, 2014) suggests:

> … When speed is not a meaningful part of the target construct, time limits should be determined so that examinees will have adequate time to demonstrate the targeted knowledge and skill. (p. 90)

We want the time limits to be loose enough that they do not interfere too much with the intended interpretation and use of the scores.

The impacts of these systematic errors tend to depend on a number of factors, including how tight or loose the limits are, the content and task types in the assessment, the assessment design, the population being assessed, and test takers' levels of motivation and test preparation. Further, in Chapter 8, Camara and Harris conclude that the modes and devices in technologically supported tests can have substantially different time requirements. So the impact of time limits will generally need to be evaluated separately for each testing program (see Chapters 5 and 8).

Impact of Population-Level Time-Limit Errors (TLE$_p$s) on Validity

The TLE$_p$ is a general systematic error estimated as the average value of the TLE over the population and is taken to have the same value for all members of the population. For norm-referenced interpretations, a test taker's score is interpreted in terms of how it compares to the distribution of scores in some population, or equivalently, in terms of the differences between the test taker's score and the scores of other members of the population. The average time-limit error for the population, TLE$_p$, is irrelevant for norm-referenced interpretations, because it has the same effect on all scores. If we subtract some number of points from everyone's score, the difference between any two scores remains the same. Note that there would be a problem if the time limits were changed and scores were compared across administrations with different time limits, but as long as time limits and the time-limit effect for the population are the same, the TLE$_p$ would not interfere with norm-referenced interpretations.

For criterion-referenced interpretations, a test taker's score is interpreted in absolute terms as indicating some level of performance (against some performance criteria, defined, for example, as a learning progression or as performance benchmarks). The TLE$_p$ is a general systematic error that reflects the average decrease in score levels associated with the time limit on the test.

If the test scores are used to make pass/fail decisions by comparing scores to a fixed passing score (or cut score), the TLE$_p$ tends to cause fewer test takers to pass by depressing the average score for the population. If the magnitude of the TLE$_p$ were known, the observed scores could be adjusted (increased) or the passing score could be adjusted (decreased) to correct for the TLE$_p$. More generally, if we have multiple cut scores, the TLE$_p$ effect could be mitigated by adjusting either the scores or the cut scores if the TLE$_p$ were known with sufficient confidence and precision (see Chapter 8).

The TLE$_p$ can be estimated directly, for example, by having a sample of test takers complete the assessment under both standard and substantially extended time conditions (with counterbalancing to control for order effects) and then comparing their performances. In this single group design, the average value of the differences between these two scores would provide a reasonable estimate of the TLE$_p$ (see Chapter 5).

Alternately, the TLE$_p$ could be estimated by dividing a sample of test takers into two randomly equivalent subsamples and either (1) administering the test to one subgroup under the standard time limit while administering the test to the other subgroup under a substantially extended time limit, or (2) administering the standard test and a shortened version of the test under the standard time limit. The difference between the average scores for the two subgroups also would provide a reasonable estimate of the TLE$_p$ (see Chapter 5).

We can also get a less direct indication of whether time limits are having any substantial impact on scores by analyzing patterns of performance across the tasks/items on the test. For example, if a substantial number of test takers has a string of omitted items, apparently random responses, or partially informed rapid responses at the end of an objective test, it is reasonable to suspect that these test takers ran out of time before completing the test (see Chapter 6).

The TLE_P typically is not a major problem. First, if the interpretation is norm referenced, the TLE_P does not generally interfere with the proposed interpretation. Second, if the interpretation is criterion referenced and the TLE_P is found to be negligible (i.e., it does not interfere with the proposed interpretation), it can be ignored. If the TLE_P is found to be significant, the time limits can be increased, the test can be shortened, and/or it may be possible to adjust the scores (or the cut scores) if the magnitude of the TLE_P is known (see Chapter 8).

Impact of Group-Level Time-Limit Errors (TLE_Gs) on Validity

TLE_Gs represent the difference in the average TLE across groups. That is, a tight time limit could have a bigger impact on one group than on another. Differences in time-limit effects for groups (e.g., race, gender, disability, first language, ability levels) raise issues of fairness. For example, test takers with visual disabilities who need to take large-print or braille editions of a test generally would need more time to complete a standardized test and therefore would be at a particularly serious disadvantage if a tight time limit were imposed on their performances (see Chapter 4). If speed is largely irrelevant to the construct of interest, then a TLE_G is a group-specific systematic error and a source of bias.

TLE_Gs are harder to estimate than TLE_Ps mainly because of the difficulty in getting adequate sample sizes. As discussed above, a TLE can be estimated most directly by having a large sample of test takers complete the assessment under the standard time limit and then under a substantially extended time limit (see Chapter 5). The average value of the differences between the two scores for the members of a group would provide a direct estimate of TLE_G. This approach works well in estimating the TLE_G for any group with a fairly large sample size (e.g., men, women, some racial/ethnic groups, groups with common disabilities). For groups with small sample sizes in the population of test takers, it would be difficult to collect large samples, but it might be possible to average results over multiple test administrations. In general, getting good, direct estimates of the magnitude of a TLE_G is difficult. Less direct approaches to detecting TLE_Gs based on patterns in item-level data (e.g., latent-group analyses) are easier to implement and can be used to detect TLE_Gs, but they are not likely to provide quantitative estimates of the magnitudes of the group-level TLEs (see Chapters 5 and 6).

Any group-level bias will have an impact on both norm-referenced interpretations and criterion-referenced interpretations, and generally it is not possible to correct for these TLEs. It is difficult to justify statistical adjustments that are applied to one group of test takers and not to other groups. Adjustments to scales (as in equating) or to cut scores that apply to all test takers can be considered part of the standardization and scaling processes that define standardized testing, but any adjustments that are made to some scores and not others tend to generate serious fairness issues.

As discussed later in more detail, group-level time-limit errors are a particularly difficult problem for groups defined in terms of disabilities, because this umbrella term includes a broad array of disabilities and a wide range of severity within disabilities. This problem is not specific to disabilities; a broadly defined group (e.g., English language learners) is a collection of special cases, but the disabilities category is particularly heterogeneous. One way to address this problem is to define more homogeneous subgroups (e.g., English language learners with different levels of competence in English), but this tends to exacerbate the sample size problem.

Impact of Individual-Level Time-Limit Errors (TLE_Is) on Validity

The TLE_Is would be specific systematic errors to the extent that they were consistent from one test administration to another, but they would function as random errors if they varied randomly across administrations (i.e., in test-retest analyses). Given that most test takers take any particular test only once at any point in their lives, it is not possible to distinguish between these two cases. In order to estimate the TLE_I for any test taker and distinguish it from sources of random error, we would need to have the test taker take the test with and without the time limit (preferably several times, with several forms of the test); in almost all cases, this is not likely to be feasible.

As a practical matter, to the extent that the TLE_Is vary from one test administration to another, they would function as random errors and could be treated as random errors (for traits, which are not expected to vary over occasions). To the extent that the TLE_Is reflect enduring differences in speed of performance that would have essentially the same impact across administrations of the test, these individual systematic errors would contribute to the universe score variance of generalizability theory or the true score variance of classical test theory (i.e., the stable component of observed scores) and therefore would tend to increase the generalizability or reliability of the scores.

To the extent that the TLE_Is contribute to the universe scores (as in generalizability theory) or true scores (as in classical test theory), they are adding a speed factor to the meaning of the scores. This is not likely to be a problem as long as the magnitude of the systematic component of the TLE is small compared to the universe-score variance (in G theory) or the true-score variance (in reliability analyses). It also is not a problem if speed is considered a legitimate part of the construct of interest.

As discussed by Bridgeman in Chapter 5, running out of time on a computer-adaptive test (CAT) can have a particularly severe impact on scores because the scoring algorithm tends to weight wrong answers toward the end of the testing period particularly heavily. Test takers who are running out of time and start guessing therefore can lower their scores considerably. This effect also has an impact on TLE_Gs and TLE_Ps, but the effect is likely to be most severe for the TLE_Is.

As indicated above, the role of the TLE_Is is complicated. Any random component in the TLE_Is would add to the overall random error in the assessment, which is of course undesirable. Any part of the TLE_I that is systematic would tend to increase the estimated generalizability or reliability of the scores but would distort the meaning of the attribute being assessed by adding construct-irrelevant variance to the scores. Both of these outcomes are undesirable, and we have no way of adjusting for any of these effects. So the best strategy would seem to be to take steps to make these TLE_Is as small as is practical.

Evaluating Time-Limit Errors

We generally don't want large errors of any kind in assessments, and we particularly don't want large systematic errors, but how do we decide if a particular source of error is too large? Two general kinds of criteria can be helpful in making such determinations. First, we don't want the TLEs to substantially increase the overall error; for this to be the case, the TLEs have to be small compared to other sources of error. Second, we don't want TLEs to interfere with the intended interpretation/use of the scores; we therefore want the TLEs to be small compared to the score differences of interest. That is, we don't want the TLEs to be large enough to make a difference in most cases.

TLEs Compared to Other Sources of Error

The overall error in scores is the major concern in evaluating assessments, and it generally includes multiple, independent sources of error. We don't want any source of error to be large, but it is the total error that is the main concern.

We generally consider item and rater sampling as sources of error, and we may consider the sampling of occasions and contexts as sources of error. In addition, it is appropriate to consider the contributions of time limits as sources of error.

Independent random errors are not simply additive because it is the error variances that are additive rather than the standard deviations of the errors. For example, given two independent sources of error, with standard errors of SE_1 and SE_2, the total standard error would be:

$$SE_{tot} = [SE_1{}^2 + SE_2{}^2]^{1/2} \tag{2.1}$$

The fact that the errors combine in this way has some important implications. In particular, relatively large errors have a disproportionally large impact on the overall error, and relatively small errors have a disproportionally small impact on the overall error. For example, if one source of error has a standard error of 5 and a second source of error has a standard error of 1, their combined standard error would be:

$$SE_{tot} = [(5)^2 + (1)^2]^{1/2} = [26]^{1/2} = 5.10 \tag{2.2}$$

That is, although the second smaller error was a fifth as large as the first error, the addition of the second error adds only about 2% to the total standard error that we would have if we eliminated the smaller source of error completely. In most cases, this small change would be considered negligible. If the second error were one third of the first error, the total error would be:

$$SE_{tot} = [(3)^2 + (1)^2]^{1/2} = [10]^{1/2} = 3.16 \tag{2.3}$$

In this case, the added error is a third as large as the first error, and the addition of the second, smaller error adds about 5% to the total standard error. In most cases, this change also would be considered negligible. For an added error that is half as large as the first error, we get:

$$SE_{tot} = [(2)^2 + (1)^2]^{1/2} = [5]^{1/2} = 2.24 \tag{2.4}$$

for an increase of about 12%.

These analyses of the relative impact of errors apply to any two or more sources of random errors or to a source of random errors with a source of systematic error. It generally does not apply to the combined effect of two sources of systematic error, because systematic errors can be correlated with each other. One rule of thumb for evaluating the impact of TLEs could be that the TLEs are substantially smaller than the total SE for other sources of error. If the TLEs were a third or less as large as the total error from other sources, the impact of the TLEs would be about a 5% increase in the total error; if the TLEs were a fifth as large as the total error from other sources, the impact of the TLEs would be about a 2% increase in the total error. So from an error-analysis point of view, the TLEs could be considered negligible as long as they were not larger than a fifth to a third as large as the total error from other sources (Kane, 2011). That is, if one can tolerate a standard error of *E* score units in some context, one can probably also tolerate a standard error of (1.02)*E* score units or (1.05)*E* score units.

Like most model-based analyses, this analysis of the relative impact of different sources of error makes some assumptions. In particular, it assumes that all errors are created equal, but this generally is not true, and in this case, it clearly is not true. The TLEs are systematic errors, and systematic errors generally are more serious than random errors because they do not tend to get smaller as the scores are based on more observations.

Even if the TLE_Gs are quite small compared to other sources of error and therefore have a relatively small impact on the total error, the kind of analysis illustrated above indicates that they tend to be problematic. To the extent that the score is not intended to have a significant speed component, any TLE_G would constitute a source of bias, and bias is a concern—even if it is small—because group-level biases raise ethical, legal, and public-relations issues.

TLEs Compared to the Score Differences of Interest

The impact of TLEs on the total SE provides an indication of how small the TLEs need to be in order to avoid increasing the overall error substantially. In addition, it is important to ensure that the TLEs (and other sources of systematic error) are small enough to not have serious impacts on the reported results or on any decisions based on the scores.

Generally speaking, it would be desirable for TLEs to be small compared to the smallest reported/interpreted score differences (Kane, 1996). For example, if scores are rounded to integers and reported as integers, and a difference of one point could lead to a change in an important decision, a TLE of half a point or more could be considered a significant problem. This issue is also likely to be most important for group-level TLEs. Because a group-level TLE is a systematic effect, it could tend to decrease the scores of everyone in the group (and, by definition, the average score of the group, by the same amount).

Both of these criteria could be relaxed to the extent that speededness is relevant to the construct of interest. As noted earlier, the impact of time limits in speed tests and relevant speed requirements in performance tests and high-fidelity simulations are not sources of error but are intrinsic to the construct of interest. More generally, some cognitive constructs can include speed as a component of the construct (see Chapter 9).

ADA Time-Limit Extensions: A Pragmatic Concern

As noted earlier, group-specific errors tend to have serious implications for fairness because they tend to introduce group-level biases that decrease scores for some groups more than others. Bias is a serious concern for any group, but for several reasons the potential for time-limit errors is a particularly serious and difficult concern for test takers with disabilities. First, test takers with some disabilities may need considerably more time to complete test tasks than comparable test takers without any disability; the adjustments that might need to be made therefore may be quite large. Second, there are many kinds of disabilities, and for each disability there are many gradations. The adjustment needed to correct for the disability would probably vary from individual to individual and from test to test, and there is no way to estimate the size of the time-limit effect for any individual on any test with any precision.

Nevertheless, time-limit extensions are the most common kind of accommodation for test takers with disabilities, even though we don't have precise estimates of how long the extensions need to be in particular cases. As a result, time-limit extensions tend to be in standard increments of 25%, 50%, or 100% more time (see Chapter 4). We want to control for possible time-limit effects due to disabilities, but we don't want to give test takers with disabilities an undue advantage. To maintain a "level playing field," it is desirable that the time limits be "loose" for most test takers so that the TLEs are controlled.

So, we have a conundrum. Time-limit extensions can improve the validity of the resulting scores for test takers with disabilities who might otherwise be subjected to a substantial construct-irrelevant barrier to their test performance. However, these extensions also raise difficult issues of fairness, especially in high-stakes contexts. If we do not provide enough of an extension, the test taker still has a construct-irrelevant disadvantage, although one that is less severe than it would be if there had been no extension. If we provide too much of an extension, we may give a test taker with a disability a construct-irrelevant advantage, which may be unfair to other test takers (see Chapter 4). The severity of the conundrum can be limited to some extent by making the time limits fairly loose for all test takers; this can be accomplished by extending the time limits or shortening the test (by including fewer tasks or simplifying the tasks). Providing prospective test takers an opportunity to practice taking released forms of the test under realistic time constraints would also probably help.

The time limits for standardized tests are often dictated by practical constraints (availability and cost of testing sites, proctors, etc.), the need for breaks during the day, and concerns about the impact of fatigue. In many cases, the most practical way to decrease speediness is likely to involve shortening the test. However, shortening the test will generally lead to a decrease in the reliability of the scores, and this tradeoff between efforts to decrease TLEs—especially TLE_Gs and the need to control random errors has to be resolved in terms of the details of specific testing programs. If the test is quite long and reliable, as is often the case for high-stakes testing programs, accepting some increase in random errors in order to reduce systematic TLEs, and thereby to reduce bias, can be a sensible choice.

The Use of Timing Data in the Scoring of Test Takers' Performances

As noted earlier, timing can play a legitimate and important role in scoring a test taker's performance. In speed tests, the test takers' scores depend mainly on how fast they complete tasks, although accuracy also plays a role. More interestingly, in a performance test or a high-fidelity simulation that uses evaluations of samples of performance from a performance domain (e.g., driving a car or truck, flying an airplane, providing medical or nursing care, baking cakes) to draw inferences about a test taker's level of competence in that domain, the score is likely to depend on timing to a significant degree because competent performance in many tasks depends on the appropriate sequencing of actions and responses and on the timing of these actions. In a driving test, one is expected to start slowing down well before one gets to a stop sign; in a test of baking skill, it is important to take the cake out of the oven soon after the buzzer sounds. As noted earlier, the time taken to perform tasks or parts of tasks may be considered an integral part of the constructs of interest, and in these cases the issue is whether the impact of time constraints is appropriate given the construct of interest (see Chapters 9 and 10).

Use of Timing Data to Check on the Functioning of the Assessment Program

With the advent of computer-based testing, much more detailed records of student performance have become available, including the time between actions and, in many cases, a detailed record of test-taker keystrokes and edits. On written tests (e.g., multiple-choice, essays), the only information we might have would be the total time a test taker spent on separately timed sections of the test, if that. This new wealth of information about each test taker's actions and the time taken to produce these actions should provide rich opportunities to develop and evaluate hypotheses about the cognitive processes engaged in by test takers, and thereby, to evaluate the validity of proposed score interpretations.

In addition to the more or less direct impact of time limits as a potential source of systematic errors in using assessment scores to estimate constructs, detailed timing data (e.g., from computer-based administrations) can provide information on additional, potential sources of systematic errors. For example, patterns in the time taken by test takers to respond to items can indicate significant sources of bias due to factors like low motivation, cheating, or item defects (see Chapters 11 and 12).

Timing information can also provide indications of whether test takers are responding to the tasks/items in a test in the ways assumed in the interpretation of the resulting scores (see Chapter 7). For example, if it were found that some test takers were taking a lot more time to complete word problems on an algebra test than they were on comparably difficult problems without much text, one could hypothesize that the test takers are slow readers. If, in addition, these test takers tended to omit tasks/items at the end of the test and word problems throughout the test, one might conclude that the reading difficulties are introducing construct-irrelevant variance into the scores.

Concluding Remarks

Standardization is very useful in controlling errors in assessments. It reduces many sources of random errors, but it has the downside of introducing some systematic errors (Kane, 1982). Systematic errors are generally more problematic than random errors because they do not cancel out over repeated assessments and thereby introduce bias. The impact of time limits on different kinds of error will depend on how tight or loose the limits are, on the content and task types involved (e.g., recognition items, routine problem-solving tasks, novel problem-solving tasks), on the population taking the test, and on their motivation and test preparation. We generally don't want large errors of any kind, but we particularly don't want large systematic errors.

For performance assessments and speed tests, time constraints need to be consistent with the intended interpretation and use of the scores. This can be operationally difficult, but it is not a fundamental problem. The more difficult issues in evaluating time limits occur in the context of standardized testing, where the variable of interest does not include a substantial speed component.

To the extent that the time limits are arbitrary, they can add both systematic and random error to the scores and thereby limit validity. The systematic errors can occur at the population level, at a group level, or at an individual level. The group-level systematic errors tend to be particularly troublesome because they raise serious questions of bias, and attempts to control this bias by extending time limits for some groups can introduce new forms of group-level bias. It is generally not practical to eliminate TLEs, but a case can be made for avoiding test designs that impose stringent time limits (see Chapter 3).

Because the time available for assessment is often quite limited, reducing speededness may require that the number of tasks/items in the test be reduced or that tasks that require a long time to complete be avoided in test design; the first of these options will tend to reduce the generalizability/reliability of scores and the second may reduce validity because of construct underrepresentation. Decreasing generalizability/reliability is, of course, not desirable, but some reduction may be advisable if it is needed to control group-level TLEs. Deciding on acceptable tradeoffs among these options is a complex problem (see Chapter 6).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bridgeman, B. (2020). Relationship between testing time and testing outcomes. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 59–72). Abingdon: Routledge.

Camara, W. J., & Harris, D. J. (2020). Impact of technology, digital devices, and test timing on score comparability. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 104–121). Abingdon: Routledge.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement: Measuring Achievement over a Decade*, *5*, 99–108.

De Boeck, P., & Rijmen, F. (2020). Response times in cognitive tests: Interpretation and importance. In M. J. Margolis & R. A. Feinberg (Eds.) *Integrating timing considerations to improve testing practices* (pp. 142–150). Abingdon: Routledge.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 237–270). Washington, DC: American Council on Education.

Harik, P., Feinberg, R. A., & Clauser, B. E. (2020). How examinees use time: Examples from a medical licensing examination. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp.73–89). Abingdon: Routledge.

Jurich, D. P., (2020). A history of test speededness: Tracing the evolution of theory and practice. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 1–18). Abingdon: Routledge.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125–160.

Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education*, *9*(4), 355–379.

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement*, *48*, 12–30.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M. T., Crooks T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.

Kyllonen, P., & Thomas, R. (2020). Using response time for measuring cognitive ability illustrated with medical diagnostic reasoning tasks. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 122–141). Abingdon: Routledge.

Lee, S., & Wollack, J. A. (2020). Concurrent use of response time and response accuracy for detecting examinees with item preknowledge. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 165–175). Abingdon: Routledge.

Lovett, B. J. (2020). Extended time testing accommodations for students with disabilities: Impact on score meaning and construct representation. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 47–58). Abingdon: Routledge.

Margolis, M. J., von Davier, M., & Clauser, B. E., (2020). Timing considerations in performance assessments. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 90–103). Abingdon: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

Porter, T. (2003). Measurement, objectivity, and trust. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 241–255.

Shepard, L. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, *31*, 165–174.

Sireci, S. G., & Botha, S. M. (2020). Timing considerations in test development and administration. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 32–46). Abingdon: Routledge.

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University press.

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150–164). Abingdon: Routledge.

3

Timing Considerations in Test Development and Administration

Stephen G. Sireci and Sandra M. Botha

Timing Considerations in Test Development and Administration

Determining the appropriate amount of testing time for examinees is one of the most significant challenges in test development, and test developers' decisions in this area are likely to set the boundaries for the fairness, utility, and validity of the overall assessment process. If insufficient time is provided to examinees, the test scores may underestimate examinees' true proficiencies. If too much time is given, the test may measure irrelevant behaviors (e.g., time management) or introduce undesirable practical effects, such as increased costs to examinees (e.g., costs for "seat time"). In this chapter, we discuss timing considerations in developing tests, beginning with defining the construct to be measured by a test and continuing through test administration. We believe that properly addressing these issues early in the test development process facilitates assessments that are fair to examinees and are more valid with respect to accomplishing the purposes of the testing program. Many of the issues we discuss are rooted in concerns for psychometric integrity, validity, and fairness. Others reflect practical realities involved in administering large-scale assessments.

Considering Testing Time from the Earliest Stages of Test Development

After deciding that a test is needed for a particular purpose, a testing program must operationally define the construct to be measured by the test. The current *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*), defines a construct as "… the concept or characteristic the test is intended to measure" (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014, p. 11). For example, proficiency in U.S. History may be a construct targeted by a high school history test, and proficiency in nursing may be a construct targeted by a nursing licensure test.

In defining the construct measured by a test, test developers must consider how much time to allocate for standardized test administration. In some cases, how quickly examinees answer test items is part of the construct to be measured by a test. For example, a test for bartending skill may want to gauge how quickly a bartender can make cocktails because the speed of task performance is relevant to performance on the job (i.e., the more drinks made in a certain timeframe, the more profit made by the establishment). In other testing situations, speed of responding to items is not part of the construct being measured. Many of the statewide achievement tests in the United States, for example, do not have time limits because they want to understand what students know and can do in general, not how quickly they can do it. For this reason, the *first* standard in the AERA et al. (2014) *Standards* states, "The test developer should set forth clearly how test scores are intended to be interpreted and consequently used … and the construct or constructs that the test is intended to assess should be described clearly" (p. 23).

Thus, in stating how test scores are to be interpreted, test developers need to address the issue of whether the test scores reflect speed of responding. Historically, tests have been described as either "speed tests" or "power tests" (see Chapter 1). The question of whether a test is considered a speed or a power test depends on whether completing the assessment is challenging because of a time limit or the difficulty of the items. Gulliksen (1950) defined a pure speed test as "a test composed of items so easy that the subjects never give the wrong answer to any of them" (p. 230). On the other hand, a pure power test is a test in which "all the items are attempted so that the score on the test depends entirely upon the number that are answered, and answered correctly" (Gulliksen, 1950, p. 231). In the first case, examinees are scored based on how many items they answer (i.e., the speed with which they complete the test), and in the second case examinees are scored on the correctness of their answer (i.e., the knowledge with which they complete the test). However, pure speed and power tests are theoretical concepts. Operationally speaking, many tests involve both speed and power components and so may not neatly fit into either one of these two categories.

Establishing Time Limits for a Test

Determining whether speed is part of the measured construct is an important part of operationally defining the measured construct and of interpreting test scores. Regardless of whether speed of response is construct relevant, all testing programs must establish standard time limits for their program. Establishing appropriate time limits for a test involves several factors. Some of these factors are practical, such as how much time can be allocated to the test given competing needs such as instruction, training, work hours, school hours, room availability, computer resources, and test administration costs. In particular, when tests are administered in testing centers, or testing proctors need to be hired, the time allotted for testing essentially dictates the test administration costs. Other factors to be considered in establishing time limits are psychometric; these involve maximizing the precision and validity of the assessment. One critical psychometric issue is *construct representation*, which requires tests to adequately represent the intended construct (Crocker, 2003; Messick, 1989; Sireci, 1998). As part of its construct definition, for example, an 8th-grade mathematics test may target different content domains such as number relations, algebraic equations, geometry, and statistics. As the number of areas to be represented increases, so too will the number of items and, by extension, testing time. To ensure that test scores provide valid interpretations, construct representation is a prerequisite. Therefore, tests must adequately represent the construct as defined in the test specifications (AERA et al., 2014; Martone & Sireci, 2009; Sireci, 1998; Sireci & Faulkner-Bond, 2014).

Another key psychometric issue is *reliability*, which refers to "the extent to which a test will give the same result on successive trials" (Wainer & Thissen, 1996, p. 22). In general, examinees' scores should not fluctuate widely over different forms of a test or across different testing occasions. In general, the more items on a test, the more reliable the test scores. As the

Standards (2014) describe, "Specifications for test length must balance testing time requirements with the precision of the resulting scores, with longer tests generally leading to more precise scores" (p. 79).

Thus, when establishing time limits for a test, test developers are faced with both practical constraints of testing costs and resources as well as the psychometric goals of construct representation and reliability. There are several steps that can be taken to optimally balance these factors. The first is to identify the number of items that is needed to adequately represent the intended construct. Next, if it is possible to pilot test the items, the amount of time examinees need to respond to the items can be estimated. Based on these pilot studies, the amount of testing time needed to fully represent the construct and to produce reliable scores can be determined (e.g., using the Spearman-Brown prophecy formula, Feldt & Brennan, 1989, or an item response theory approach based on extrapolating test information from the sample of items; see Wainer & Thissen, 1993). If the estimated amount of time is prohibitive either from a practical or a cost perspective, the construct domain can be reduced or a more efficient test design such as computerized adaptive testing (CAT) can be used.

Test developers can also conduct research to determine if the suggested time limit for a test is appropriate. If the test is *not* designed to measure speed of responding, then research regarding whether examinees can complete the test in the allotted time or whether they feel rushed to do so (and hence perform suboptimally) should be conducted. As the *Standards* (2014) recommend, "For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure" (p. 90).

In a subsequent section of this chapter, we discuss how test developers can assess whether power tests are unnecessarily speeded. At this point, we want to underscore the importance of establishing time limits that support valid score interpretations. Valid score interpretations begin with a clear definition of the construct targeted by a test and time limits that allow proper measurement of examinees' proficiencies with respect to that construct. We quote the *Standards* (2014) once again, for a description of these important test development tasks,

> In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats … [and] also specify the amount of time allowed for testing. (AERA et al., 2014, p. 85)

A summary of the steps needed to develop quality tests that promote valid score interpretations and are fair to all examinees is presented in Table 3.1. The degree to which timing considerations are needed at each step is also indicated. A clear conclusion from Table 3.1 is that timing considerations are relevant to virtually all stages of test development and are paramount from the earliest stages.

Before leaving this section on timing considerations in test development, it is important to note that Step 5, *Determine the scoring rules*, has implications for the psychometric model used to estimate scores for examinees. If the construct definition specifies that speed of response is relevant and will facilitate valid score interpretations, scoring models that include speed of response should be considered. Several item–response-theory-based scoring models that incorporate response time have been proposed. These include Wang and Hanson (2005); van der Linden (2007, 2009); van der Linden, Entink, and Fox (2010); and Molenaar, Bolsinov, and Vermunt (2018).

Test Administration Models

Test developers can choose from a number of different approaches to delivering an examination ranging from the traditional paper-based designs to the newer digital designs; the selected

Table 3.1 Summary of test development steps and timing considerations

Test Development Step	Timing Consideration
1. Establish need for test and specify testing purpose(s)	Is measuring how quickly examinees respond to items relevant to the testing purpose?
2. Define the construct to be measured	Is measuring how quickly examinees respond to items part of the construct intended to be measured?
3. Determine item formats	How efficient are the item formats with respect to gathering examinee responses? Will examinees be familiar with these formats?
4. Determine test administration design	Will the test be linear or adaptive? Are there costs for "seat time"?
5. Estimate testing time and test length	What are the practical constraints on testing time? How much time is needed to fully represent the construct?
6. Determine scoring rules	Should speed of responding be included in the scoring model?
7. Develop test specifications	Ensure construct representation within the given testing time.
8. Develop test items to represent test specifications	Develop sufficient numbers of quality items to support adaptive item bank or fixed-length test.
9. Content and measurement review of test items	Are directions clear?
10. Pilot test items	Evaluate item and test response time distributions.
11. Establish test administration conditions	Set time limits and test administration timing instructions.
12. Statistical (item analyses) and qualitative (sensitivity) item reviews	
13. Assemble test forms (or panels, or adaptive item banks)	Select items with appropriate response time distributions.
14. Fieldtest (mimic operational test length, time limits, etc.)	
15. Psychometric research: item analyses, analysis of item response time, speededness research, calibration, scaling	Evaluate testing time limits and scoring algorithms.
16. Score report design	Include information on engagement and speed of completion, if relevant.
17. If necessary: Standard Setting, Norms development	Consider effects of non-engaged or rushed examinees on norms and standards.
18. Develop technical documentation	Document results of speededness, timing, and other relevant analyses.

model directly impacts the amount of time needed and allowed for navigating and completing the test. Paper-and-pencil tests are administered in a "linear" format, which means that the items are administered in a predefined sequential order that may be the same for all test takers. Though computer-based tests (CBT) can also be administered in a linear format, many additional test-administration models are possible with CBT that are not possible with paper-based tests; these include "linear-on-the fly testing" (LOFT) and "adaptive" models (see Luecht & Sireci, 2011; Wainer, 1993; Yan, von Davier, & Lewis, 2014). The choice of administration model has implications for measurement-related areas, such as test reliability and length, as well as for administrative and policy considerations, such as test security, cost, and maintenance. Ideally, choosing a test administration model should be based on a thorough evaluation of the model's usefulness and feasibility as it relates to the goals and purposes of a testing program (Luecht & Sireci, 2011).

The benefits of CBTs over paper-and-pencil designs include improved test security, more flexible test administration schedules, immediate scoring and reporting, and the ability to include multimedia in the assessment. In addition, one of the most widely cited advantages of CBT is the ability to administer a test "tailored" to the specific characteristics of a test taker. This tailoring is achieved by using a computerized selection algorithm that selects items (or sets of items) to be administered to a specific examinee based, in part, on the proficiency of the examinee. Examinees therefore receive a test tailored to their individual proficiency. Currently, there are various computerized-adaptive administrations available that can loosely be classified into two general categories: CAT and multistage adaptive testing (MST).

CAT is a computerized test that is adaptive at the item level. Although the specifics behind the working of the item selection algorithm vary across testing programs, all algorithms involve estimating an examinee's proficiency after each item is presented and then using this provisional proficiency estimate to select the next item. This item-selection process occurs after each item is presented until the testing process terminates.

As more items are administered, measurement error decreases. Based on the characteristics of item response theory (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980), a CAT can be designed to continue administering items until the measurement error reaches a target minimum. Because examinees are not given items that are far too easy or far too difficult for them, CATs can achieve an error of measurement target (or similarly, a reliability target) using about half as many items as would be used with a linear design (Luecht & Sireci, 2011; Wainer, 1993). For this reason, many testing programs can substantially trim the testing time requirements by moving from a linear design to a CAT.

MST similarly targets the difficulty level of a test to provisional estimates of an examinee's proficiency, but the selection decisions are made after a set of items is administered and responded to, rather than after each item. After scoring a set of items, which often is referred to as a *testlet* (Wainer & Kiley, 1987) or a *module*, a provisional estimate of the examinee's proficiency is calculated and the next set of items is chosen that is best matched to that proficiency level. The examinee once again starts at a moderate point of difficulty and the test is administered in stages. Each of the modules consists of several items that are assembled according to pre-established test specifications. All items in the module are chosen so that the average difficulty matches a pre-specified target difficulty. After each module, the examinee's proficiency is estimated and used to select the next module. Although the measurement precision gains relative to a linear design are not as large as that of a CAT design, they also can substantially reduce test length (and thus testing time). Research has shown that MSTs can be almost as efficient as a CAT with respect to measurement error and have the additional advantages of better item pool usage, content coverage, and flexibility in allowing examinees to review items within a module. There are complex variations of MST designs that are beyond the scope of this chapter; for more detail, interested readers are referred to Yan et al. (2014).

CATs and MSTs have significant implications for test administration and timing. As described earlier, they can achieve reliable scores for examinees using fewer items than are typically required using non-adaptive designs. This increase in efficiency is a direct result of the item selection algorithm, which avoids administering items that are too difficult or too easy for a test taker. Therefore, these tests often are significantly shorter than paper-and-pencil tests. Although reduced testing time is frequently cited as a great benefit for developers, examinees, and administrators, there are additional timing issues to be considered in the administration design and process of CAT.

Although CAT has clear benefits, it is easy to exaggerate the real cost savings that can be credited to gains in measurement efficiency. The potential reduction in testing time might be irrelevant if the test is administered at commercial testing centers. These centers usually require a guaranteed minimum testing time while charging fixed hourly rates per examinee (Luecht & Sireci, 2011). As an example of how this comes into play, consider a situation in which the test center vendor negotiates with the developer for a 3-hour test. In this case, the same fee may apply whether the actual seat time of the examinee is one, two, or three hours. The time savings may be realized for many examinees, but not for the testing agency. Relative to linear-based tests, adaptive test designs also may require larger banks of items with sufficient statistical properties.

Timing Considerations in Item Development

Regardless of the test administration model, sufficient numbers of quality items must be developed to support a testing program. Although some testing programs cling to the 20th century paper-based format, a hallmark of contemporary educational assessments is the use of technology to allow examinees to interact with and respond to items. Item types that use technology in some way are often referred to as *technology-enhanced items* (TEIs). TEIs may use technology in varying the presentation of items to examinees (e.g., incorporating videos, or allowing examinees to access resources while responding to an item) or in changing the ways examinees provide their responses (e.g., "drag-and-drop," "point-and-click," "hot spot items," etc.; see Sireci & Zenisky, 2016). These new item formats have a direct impact on testing time, because they have the potential to greatly increase the amount of time examinees spend interacting with an item and recording their answers to it. For this reason, TEIs should have psychometric advantages (e.g., increased construct representation) to justify any increase in testing time (Huff & Sireci, 2001; Jodoin, 2003; Paniagua et al., 2017).

A recent white paper by the Association of Test Publishers (ATP) and Institute for Credentialing Excellence (ICE) provides important information regarding how TEIs affect testing time (ATP & ICE, 2017). Based on a review of the literature, they recommended that testing programs using TEIs, "Evaluate the amount of time that should be provided to examinees to complete the examination when adding [TEIs] since more than one study found that [TEIs] were more time consuming than traditional [items]" (p. 24). The studies they reviewed included Dwyer, Penny, and Johnson (2015); Jodoin (2003); Krogh and Muckle (2016); and Woo, Kim, and Qian (2014).

The ATP and ICE (2017) white paper drew from McSweeney (2013) to illustrate the importance of pilot-testing TEIs to gauge the appropriate amount of testing time needed, and also the importance of having generous time limits in those pilot studies. As they described,

> when a large [IT] company moved one of its certification exams from task-based items to project-based performance items, it discovered during the field test that examinees spent much more time reviewing each data point and rechecking instructions—to the extent that many of them timed out of the test. Since the test was not designed to be completed quickly, the IT company had to revisit the timing (McSweeney, 2013, cited in ATP & ICE, 2017, p. 21).

For this reason, ATP and ICE (2017) recommended, "It is particularly important for a program that is transitioning to a new format to plan carefully and pay close attention to examinee perceptions, item performance, and item/test completion time" (p. 21).

Given that TEIs will typically require increased testing time, they must add value from a construct representation standpoint. That is, TEIs should measure skills that are not measurable using standard item formats that take less time (e.g., multiple-choice items). As ATP and ICE (2017) concluded,

> Most certification tests are not speed dependent, so test time must account for any additional time examinees need. Significant questions are whether there is any additional value for the examinee and for the program if an increase in time is required, and can better results be obtained that can be balanced against increased expenses for seat time. (p. 21)

Although the latter quote was in the context of certification testing, it clearly applies to all tests that do not include speed of response as part of its operational definition of the construct.

When testing programs incorporate TEIs, we recommend providing practice tests. This will allow examinees to become familiar with the instructions and the interfaces and make it less likely that they will need to use valuable testing time learning how to interact with the items. Clear directions that minimize verbosity are also important, as is consideration of whether the requirements of TEIs may facilitate or interfere with the performance of examinees with special needs such as students with disabilities and linguistic minorities (e.g., English learners; Crotts-Rohor & Sireci, 2017).

The International Test Commission's (2005) *Guidelines on Computer-Based and Internet Delivered Testing* also provide useful suggestions for timing considerations. For example, they state, "When the CBT/Internet test is timed, design features so that the time required to move between questions and for the system to record the answer is not part of the timed element (e.g., the test software should deduct these times from the timing of the test or the timing clock should stop during access transitions)" (p. 16). We agree with this recommendation so that examinees are not penalized for time requirements due to factors—such as item formats or test delivery features—that are outside of their control.

Statistical Analysis of Item Response Time Data

In the previous sections, we discussed the need to consider timing issues in test development and test administration design. In this section, we discuss issues related to the analysis of item response time data, which refers to data measuring how long it takes examinees to respond to items and to complete a test. As mentioned earlier, some of the analyses of item response time data can be used in test development (e.g., selecting items for a test or to estimate total testing time) and in test scoring (e.g., Molenaar et al., 2018; van der Linden, 2007, 2009; Wang & Hanson, 2005). However, there are also other important uses of response time data; we discuss those next.

One way that item response time and total testing time data can be used is to evaluate the degree to which "speededness" is present in an assessment. Evaluation of the degree to which a test is speeded is important whether speed of response is construct relevant (i.e., the test is designed to measure speed of responding) or construct irrelevant (i.e., an undesired factor that leads to inappropriate interpretations of test scores). The *Standards* (2014) define "speededness" as the "extent to which test takers' scores depend on the rate at which work is performed as well as on the correctness of the responses" (p. 233). This definition underscores the importance of assessing speededness, and estimating its effects, to ensure proper interpretation of test scores.

Methods for Evaluating Test Speededness

There are several methods for evaluating test speededness. Some are based on experiments, others are based on calculating the percentage of examinees who complete a test or portions of a test, and still others are based on evaluations of item response time or modeling of student response data (Ying & Sireci, 2007).

Experimental methods for evaluating speededness include test-retest (alternate-form) designs, where examinees are administered tests with and without time limits, and randomly equivalent group designs, where examinees are randomly assigned to timed and untimed conditions or test sections with different numbers of items to complete within a given time period (e.g., Bridgeman, Trapani, & Curley, 2004; Harik et al., 2018; Chapter 6, this volume). In the test-retest design, if additional time has no effect on the subjects' scores, the test may be regarded as unspeeded. Cronbach and Warrington (1951) proposed a correlation-based measure of speededness for this situation, *tau*, which is based on the correlation of scores from parallel tests administered under timed and untimed conditions. Administering two parallel tests under timed and untimed conditions is resource intensive, however, and so this approach is not used often. Even when repeated test administration is possible, it is hard to keep examinees equally motivated for two test administrations. For these reasons, estimates of speededness based on a single test administration are more common.

Estimates of test speededness from a single administration of a test date back at least to Gulliksen (1950), who suggested comparing the standard deviation of the number of not reached items to the standard deviation of the number of items that were not answered correctly. If the ratio is small, the test may be regarded as primarily a power test. Swineford (1974) suggested that as long as this ratio is less than 0.25, the test may be considered unspeeded. Swineford also included a very liberal index based on the percentages of examinees completing certain portions of a test. This speededness criterion considers a test unspeeded if at least 80% of the examinees reach the last item and all examinees reach at least 75% of the items. However, as Ying and Sireci (2007) point out, "Although this 'Swineford criterion' is easy to use as a standard for flagging speededness, it allows for a speeded exam for 20 percent of the examinees, many of whom may come from subgroups that perform relatively low" (p. 33).

Estimates of speededness based on the number of items at the end of the test that the examinee did not have time to answer have fallen out of favor due both to their limitations as well as to the availability of item response time data from CBT. One limitation of these older methods is the assumption that an examinee works at a constant pace throughout the test, and when time expires, there are "not reached" items. It is much more realistic that examinees keep track of time and either skip items when they are running out of time or answer items in a random fashion as time expires. Analysis of item response time data (discussed later) makes such behaviors easier to identify.

There are, however, other methods for assessing speededness that do not rely on unattempted items and are based on analysis of data from a single test administration. Specifically, factor analysis, multidimensional scaling, and item response theory (IRT) can be used to investigate whether speed is a significant factor associated with the last section of items in the test. For example, a College Board study used factor analysis to analyze students' responses to SAT items and found that factors attributable to speed typically accounted for about 5–10% of the variance in examinees' scores (College Entrance Examination Board, 1984). Bejar (1985) suggested analyzing the fit of the most difficult items on a test to the IRT model for the examinees who were most likely to guess due to running out of time. His rationale was that on the difficult items, lower-ability examinees would perform better than predicted due to random or patterned responses. Zenisky, Hambleton, and Sireci (2002) used the Q3 statistic for assessing local item dependence within an IRT model and found higher dependence for items associated with passages at the end of a test; they claimed that this was due to speededness. Yamamoto (1990, 1995) proposed an extended HYBRID IRT model that uses multiple item response models to estimate the proportion of examinees who switch from an ability-based response strategy to a random response strategy while taking the test. The probability of correct responses when an examinee is in "guessing mode" is no longer determined by the traditional IRT models but rather by a guessing class-based multinomial model. Others have used IRT mixture models (e.g., Cohen, Wollack, Bolt, & Mroch, 2002) to classify examinees into latent classes of speeded and non-speeded examinees. Thus, there are several statistical methods for evaluating test speededness available to researchers and testing agencies.

Assessing Speededness on Computer-Based Tests

When a test is administered digitally (e.g., CBT), the device on which the test is administered can track how long it takes examinees to respond to items as well as other test-taking behavior (e.g., clicking on different features of an item). These item response time data can be used to identify examinees who may be guessing due to running out of time or other factors. Schnipke (1995) graphed the standardized natural logarithm of response time of Graduate Record Examinations General Test (GRE) items and was able to detect examinees whose responses suddenly accelerated and became less accurate, which suggested that they were running out of time. She also examined the response time distributions for each item together with the proportion of examinees giving correct responses at each response time level. She found that items appearing later in the test had more combinations of short response times and inaccurate responses.

Although it may be easy to identify when examinees start guessing rapidly on a CBT, the issue of testing time and adaptive testing is more complex. When the stopping criterion for an adaptive test is based on score precision, such as a small conditional standard error of measurement, examinees are likely to receive different numbers of items and so would be expected to need different amounts of time to complete the test. Moreover, examinees who do well on an adaptive test will generally see more difficult items; these items may also require more time given the associated higher cognitive load (Swygert, 2003). Thus, adaptive testing may produce a test that is differentially speeded, and hence unfair, for the most proficient examinees. Testing programs may be able to mitigate this problem by using item response time statistics in the item selection algorithm to create tests of similar time requirements for all examinees (van der Linden, Scrams, & Schnipke, 1999). If not, "it may be inappropriate to have a common time limit for examinees with very different proficiencies" (Ying & Sireci, 2007, p. 35).

Response Time Engagement and Speededness

In addition to test speededness, test-taking motivation is another potential construct-irrelevant factor that affects testing time yet may be overlooked in testing research. If examinees are unmotivated, they may carelessly rush through a test. Therefore, methods for assessing examinee motivation while taking a test (e.g., Wise & DeMars, 2005; Chapter 11, this volume) are also relevant for assessing speededness.

Engagement in the assessment process is typically related to the examinee's motivation and perception of the stakes associated with the consequences of the assessment process. The most common way to conceptualize the stakes is by sorting them into two categories: low-stakes and high-stakes tests (Wigfield & Eccles, 2000; Wise & DeMars, 2005). Thus, a student might perceive a test as low stakes if there were no direct consequences related to their test score (e.g., the National Assessment for Educational Progress). In contrast, a student might perceive a test as high stakes when test scores have direct consequences, as in the case of a college admissions test. Research has shown that students are less motivated during a perceived low-stakes testing event and, consequently, might exhibit low effort during these tests (Barry, Horst, Brown, Finney, & Kopp, 2010). Thus, non-effortful behavior is more prevalent in low-stakes testing. If such behavior is not accounted for in assessing speededness, a test may appear to be speeded when it is not.

Measuring examinee effort on a test is difficult because the amount of effort given by the examinee is not usually consistent throughout the test. An examinee might show good effort on some sections of the test and no effort on other sections. Past research generally has employed either (a) self-report measures, (b) person-fit statistics, or (c) behavior-based measures to assess test-taking effort. Self-report scales are used commonly but are limited in that they provide information only at the overall test level and also are vulnerable to several biasing factors (Pintrich & Schunk, 2002; Wise & Kong, 2005). For example, examinees may understate their effort if they feel that they performed poorly on a test or overstate their effort if they fear retribution from test administrators for not responding effortfully. Person-fit statistics are based on observed responses and, though not vulnerable to the same biasing factors as self-report measures, they are sensitive to other sources of misfit that are not attributable to lack of effort (Meijer & Sijtsma, 2001; Wise, 2015).

Measures based primarily on students' item response times can be used as a more direct indicator of examinee engagement. Item response time data are collected unobtrusively and so are not subject to potential biases from examinees as is the case with self-report data. Also, item response time data are collected for each examinee for each item, which means that effort can be assessed at both the item level and the total-test level. Another advantage is that the data are collected and stored automatically, requiring no additional effort from examinees, administrators, test users, or test developers.

Measures of effort based on item response times are grounded in motivation research and based on concepts termed *rapid-guessing behavior* and *solution behavior* (Schnipke & Scrams, 1997). Rapid-guessing behavior is just what it sounds like: an examinee provides a rapid and random response to an item. Solution behavior occurs when an examinee answers an item in an effortful manner. Thus, an examinee is disengaged when she/he exhibits rapid-guessing behavior, and such behavior can be interpreted as non-effortful; the reverse then is true of examinees who exhibit solution behavior. Rapid-guessing behavior can also occur on high-stakes tests when examinees are motivated but are running out of time (see Chapter 6). Studies have shown that rapid-guessing behavior has a detrimental effect on score validity (Wise, Bhola, & Yang, 2006) and also spuriously increases the internal consistency of test score data (Wise, 2006; Wise & DeMars, 2006).

Methods for identifying rapid-guessing behavior include (a) response time thresholds, (b) the solution behavior index, (c) the response time effort index, and (d) the response time fidelity index. These methods are mostly applicable in low-stakes assessments, but they can also be used in high-stakes assessments where unmotivated examinees may be present or rapid-guessing behavior occurs because examinees run out of time. Examples of these methods can be found in Harik et al. (2018); Wise, Kingsbury, Thomason, and Kong (2004); Wise and Kong (2005); Wise (2006); Kong, Wise, and Bhola (2007); and Wise and Ma (2012).

Evaluating Differential Speededness and Speededness of Constructed-Response Items

Many of the methods for detecting speededness are appropriate for different item formats, but constructed-response items deserve a bit more discussion because the response provided—or not provided—by examinees can yield additional information about the appropriateness of the amount of time given for completion of these more extended tasks. With insufficient testing time, examinees may leave constructed-response items blank or provide answers that are much shorter than they would be if there were sufficient time. Sireci, Wells, and Hu (2014), for example, found that English language learners were much more likely to leave constructed-response items blank compared to other groups of students. Whether this problem is due to insufficient time or due to group proficiency differences deserves further study. The analysis of omit rates is one type of validity evidence based on response processes mentioned by the *Standards* (2014).

Using Item Response Data in Test Assembly

In the previous sections, we discussed timing considerations pertaining to construct definition, construct representation, establishing reasonable and fair time limits for tests, using adaptive designs for reducing testing time, and assessing speededness. In those sections, we briefly touched on how item response time could be used in test development. In this section, we focus on such use of item response time data and summarize those points.

Field-testing items and analyzing item response time data is one of the best ways a testing agency can establish test lengths and time limits. In many cases, simple descriptive statistics can provide valuable information. As mentioned earlier, using the average item response times to select items in a computerized-testing environment has also been suggested to make total testing time more consistent across examinees (e.g., van der Linden et al., 1999). Pilot data can be useful for calculating item response time statistics. However, the response time characteristics of items should be monitored over time, as items become operational.

Another way item response time data can be used in test development is for evaluating the quality of items. For example, when test items are administered across different grade levels, it may be expected that students at higher grade levels would respond more quickly. Such hypotheses could be tested by evaluating item response times across grades. As another example of evaluating item quality, Wang and Sireci (2013) used item response time to investigate whether items measuring higher-order cognitive skills had larger item response times than items measuring lower-order cognitive skills. They found significantly longer response times for items measuring higher-order cognitive skills, which they claimed provided important validity evidence based on response processes. Zenisky and Baldwin (2006) went even further in evaluating factors affecting item performance by investigating the relationship between median response time and item difficulty, item complexity, and cognitive area. Extrapolating from these studies, it is clear that response time data can be used in test development to select items whose response time characteristics are consistent with the construct theory describing what the items are designed to measure.

This section of the chapter has provided evidence that item response time data can be valuable in both test construction and test score validation. With respect to test construction, such data can be used to establish time limits for a test, to select items for a test form or adaptive algorithm, or to select quality items that function as intended. With respect to test validation, these data can be used to evaluate differences across subgroups of examinees in the amount of time taken to respond to items as well as to evaluate the cognitive processes used by examinees to respond to items.

Conclusions and Future Directions

In this chapter, we addressed several issues related to testing time. These issues included determining whether speed of responding to items is part of the construct intended to be measured on a test, factors to consider in establishing time limits on a test, different test administration designs that impact testing time, and analysis of item response time data for test development and test evaluation. Although there has been some impressive research conducted in this area, we believe that with item response time data becoming more available in computer-based testing programs, this domain of research is still in its infancy. Thus, we look forward to more research that investigates item response time with respect to test design, test administration, test scoring, and evaluation of potential biases in testing.

One area in which we think more immediate research is needed is in studying the different amounts of testing time used by students of different levels of proficiency. In some testing programs, we have observed that students who are doing very well on an adaptive test require much more time to complete a test than students who are doing less well. Such differences in total testing time make sense, because students who are doing well on an adaptive test are continually challenged by more difficult items. If the end result is that some students need much more time to complete a test than others, the testing program can be accused of systematic bias (i.e., variable [nonstandard] administration conditions). If those examinees who receive the most difficult items—and hence need more testing time—are sufficiently rewarded for their performance with high scores, perhaps there can be no criticism of bias. These possibilities underscore the need for further study of the costs and benefits of testing time and different adaptive testing designs.

A related issue is whether test scores should be flagged if examinees require and receive different amounts of testing time. We believe in the principles of universal test design (Thompson & Thurlow, 2002), which suggest that test administration conditions be sufficiently flexible so that accommodations such as extended time are not necessary. Given the 21st-century testing technology, it seems sensible that different examinees can be granted different time limits based upon their needs. However, if speed of response is an intended part of the construct, different time limits may affect score interpretation. Thus, as we have consistently emphasized in this chapter, definition of the construct must include consideration of the degree to which speed of responding to items is construct relevant. Prior research in this area on qualifying score interpretations (e.g., Sireci, 2005) and providing test accommodations (e.g., Crotts-Rohor & Sireci, 2017; Sireci, Banda, & Wells, 2018; Chapter 4, this volume) should be helpful in this regard.

In summary, consideration of timing issues in test development, administration, and evaluation necessitates that testing agencies first decide if speed of response is relevant or irrelevant and then design the test and its administration accordingly. As Messick (1989) pointed out, "Tests are imperfect measures of constructs because they either leave out something that should be included … or else include something that should be left out, or both" (p. 34). If speed of response is relevant to the construct a testing agency intends to measure, then it should be factored into scoring and score reporting. If it is not, speed should not impact scores. Once again, we turn to the *Standards* (2014), which emphasize continual focus on the construct measured to support valid interpretations of test scores. Specifically, the *Standards* state, "All steps in the testing process, including test design, validation, development, administration, and scoring procedures should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population" (p. 63). We concur with this advice, and we recommend that it be used by testing programs to develop their test development and validation research agendas.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Association of Test Publishers (ATP) and Institute for Credentialing Excellence (ICE). (2017). *Innovative item types: A white paper and portfolio.* ATP and ICE.

Barry, C. L., Horst, S. J., Brown, A. R., Finney, S. J., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*, 342–363.

Bejar, I. I. (1985). Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language. *ETS Research Report Series*, *1985*(1), 1–57.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41*, 291–310.

Cohen, A. S., Wollack, J. A., Bolt, D. M., & Mroch, A. A. (2002, March). *A mixture Rasch model analysis of test speededness*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

College Entrance Examination Board. (1984). *The College Board technical handbook for the scholastic aptitude test and achievement tests*. New York, NY: Author.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, *22*(3), 5–11.

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika*, *16*(2), 167–188.

Crotts-Rohor, K., & Sireci, S. G. (2017). Evaluating computer-based test accommodations for English learners. *Educational Assessment*, *22*(1), 35–53. Online version published February 8, 2017, doi: 10.1080/10627197.2016.1271704.

Dwyer, A. C., Penny, J. A., & Johnson, R. L. (2015, February). Scoring alternative item types: There's many a slip between the cup and the lip. Paper presented at the Association of Test Publishers Innovations in Testing Conference, Palm Springs, CA.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: American Council on Education and Macmillan.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., … Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement*, *55*, 308–327. doi:10.1111/jedm.12177

Harik, P., Feinberg, R. A., & Clauser, B. E. (2020). How examinees use time: Examples from a medical licensing examination. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 73–89). Abingdon: Routledge.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, *20*(3), 16–25.

International Test Commission. (2005). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, *6*(2), 143–171.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, *40*, 1–15.

Jurich, D. P. (2020). A history of test speededness: Tracing the evolution of theory and practice. In M. J. Margolis, & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 1–18). Abingdon: Routledge.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606–619.

Krogh, M. A., & Muckle, T. J. (2016). Assessing the psychometric properties of alternative items for certification. *Journal of Applied Measurement*, *17*(4), 489–501.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lovett, B. J. (2020). Extended time testing accommodations for students with disabilities: Impact on score meaning and construct representation. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 47–58). Abingdon: Routledge.

Luecht, R. L., & Sireci, S. G. (2011). A review of models for computer-based testing. *Research report 2011-2012*. New York, NY: The College Board.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction, *Review of Educational Research*, *4*, 1332–1361.

Meijer, R. R., & Sijtsma, J. (2001). Methodology review: Evaluating person fit. *Applied Measurement in Education*, *25*, 107–135.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: American Council on Education and Macmillan.

Molenaar, D., Bolsinov, M., & Vermunt, J. K. (2018). A semi-parametric, within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 205–228.

Paniagua, M., Swygert, K. A., Billings, M., DeRuchie, K., Haist, S. A., Hussie, K., … Merrell, J. (2017). *Constructing written test questions for the basic and clinical sciences* (4th ed.). Philadelphia, PA: National Board of Medical Examiners.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Upper Saddle, NJ: Merril Prentice-Hall.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83–117.

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher*, *34*(1), 3–12.

Sireci, S. G., Banda, E., & Wells, C. S. (2018). Promoting valid assessment of students with disabilities and English learners. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices: Issues, innovations, and application* (pp. 231–246). Springer International Publishing: Switzerland.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*, 100–107. doi: 10.7334/psicothema2013.256.

Sireci, S. G., & Zenisky, A. L. (2016). Computerized innovative item formats: Achievement and credentialing. In S. Lane, T. Haladyna, & M. Raymond (Eds.). *Handbook of test development* (pp. 313–334). Washington, DC: National Council on Measurement in Education.

Sireci, S. G., Wells, C., & Hu, H. (2014, April). *Using internal structure validity evidence to evaluate test accommodations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Swineford, F. (1974). *The test analysis manual* (ETS SR, 74-06). Princeton, NJ: Educational Testing Service.

Swygert, K. A. (2003). *The relationship of item-level response times with examinee and item variables in an operational CAT environment* (LSAC Computerized Testing Report 98 10). Newtown, PA: Law School Admission Council.

Thompson, S., & Thurlow, M. (2002, June). Universally designed assessments: Better tests for everyone! *Policy directions, Number 14*. Minneapolis, MN: National Center on Educational Outcomes.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W.J. (2009). Conceptual issues in response time modeling. *Journal of Educational Measurement*, *46*, 247–272.

van der Linden, W. J., Entink, R. H. K., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*(1), 15–20.

Wainer, H., & Kiley, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103–118.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*(1), 22–29.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wang, X., & Sireci, S. G. (2013, April). *Investigating the relationship between item response time and cognitive level.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wise, S. L. (2006).. An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education*, *19*, 95–114.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*, 237–252.

Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150–164). Abingdon: Routledge.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: the normative threshold method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Bhola, D., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice*, *25*(2), 21–30.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Woo, A., Kim, D., & Qian, H. (2014, November). Exploring the psychometric properties of innovative items in CAT. Paper presented at the 14th annual Maryland Assessment Conference, College Park, MD.

Yamamoto, K. (1990). *HYBIL: A computer program to estimate HYBRID model parameters*. Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model (*TOEFL Tech. Rep. No. TR-10*). Princeton, NJ: Educational Testing Service.

Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing theory and applications*. Boca Raton, FL: CRC Press Taylor & Francis Group.

Ying, L., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29–37.

Zenisky, A. L., & Baldwin, P. (2006). Using item response time data in test development and validation: Research with beginning computer users. *Center for educational assessment report No*, *593*.

Zenisky, A.L., Hambleton, R. K., & Sireci, S. G. (2002) Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*(4), 291–309.

4

Extended Time Testing Accommodations for Students with Disabilities

Impact on Score Meaning and Construct Representation

Benjamin J. Lovett

By their very nature, standardized tests are designed to be administered to all examinees under similar conditions. Indeed, variability in administration across examinees is a clear threat to test fairness (Wollack & Case, 2016). That being said, *testing accommodations* must be provided when a disability condition keeps an examinee from accessing the test under standard administration conditions. Two examples of such accommodations—which change the administration conditions of the test in some way without changing the actual test content (Lovett & Lewandowski, 2015)—are reading the test items aloud to a visually impaired examinee and providing an examinee in a wheelchair with a wheelchair-accessible desk for the testing session. Accommodations are required under disability discrimination and special education laws when needed for examinees to access tests, and the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) recognizes the responsibility of test developers and users to develop and provide appropriate accommodations.

Testing accommodations are common; during the 2015–16 test cycle, the College Board (which administers the SAT, among other tests) received 160,000 requests for accommodations—double the number received just 5 years prior to that (Yellin, 2017). Accommodations often modify the examinee's response mode (e.g., allowing dictation of answers), the presentation format of the test (e.g., an audio recording of test items), or the scheduling of the test (e.g., in the afternoon, due to health problems that flare in the morning). The most common accommodation on virtually every test is extended time, and it is exactly what it sounds like: giving an examinee additional time to work on the test (see, e.g., U.S. Government Accountability Office, 2011). The present chapter focuses on this accommodation, with particular attention to its impact on the meaning of resulting test scores and more broadly on the ability of the (accommodated) test to measure all aspects of the intended construct.[1]

The Popularity of Extended Time Accommodations

Many students receiving extended time do not have any sensory or physical handicaps; instead, they have been diagnosed with learning, cognitive, or psychiatric disabilities—conditions that together are often called "hidden disabilities." On admission, certification, and licensing tests, learning disabilities and attention-deficit/hyperactivity disorder (ADHD) are the most commonly accommodated conditions (U.S. Government Accountability Office, 2011). Similarly, in school settings, the highest-incidence special education classification is learning disabilities (Heward, 2013). Many testing agencies also have reported an increase in the number of students requesting extended time for psychiatric conditions, particularly anxiety. The high incidence of hidden disabilities is one reason why extended time is such a common accommodation; another reason is that students who use auxiliary aids or human assistants (e.g., a live test reader) often need additional time to utilize these other accommodations.

Of course, time pressure during tests is an experience that is common to examinees both with and without disabilities. Time limits are a known contributor to test anxiety, and so it is not surprising that time extensions would be viewed as desirable by many examinees. In their review of the literature on students' perceptions of testing accommodations, Lovett and Leja (2013) concluded that accommodations reduce anxiety and discomfort for students with and without disabilities. Extended time specifically has been found to do this; for instance, Elliott and Marquart (2004) found that giving fourth-grade students 40 minutes instead of 20 minutes to complete a math test (i.e., 100% extended time) made the majority of students with and without disabilities feel more relaxed (78% and 75%, respectively). The reason for this relaxation may be an expectation of score improvement; in one large-scale survey, Lewandowski, Lambert, Lovett, Panahon, and Sytsma (2014) found that the vast majority of college students with and without disabilities (approximately 87% of each group) felt that 50% extended time would lead to at least somewhat higher scores for them on a high-stakes test.

Students' perceptions of extended time correspond to the actual effects of the accommodation; extended time consistently has been found to improve scores on standardized tests for students with and without disabilities. Although some disability advocates (e.g., Shaywitz, 2003) claim otherwise, three systematic literature reviews (Cahan, Nirel, & Alkoby, 2016; Lovett, 2010; Sireci, Scarpati, & Li, 2005) all have found this nonspecific effect of extended time. Whether students with disabilities benefit *more* from extended time than do nondisabled students is a more complex issue that appears to depend on just how speeded the test is. On moderately speeded tests, some nondisabled students are more likely to finish within the standard time limits, making the extended time of negligible value to that subgroup; on highly speeded tests (cf. Lewandowski, Lovett, Parolin, Gordon, & Codding, 2007), most or all nondisabled examinees are working throughout the full (extended) time allotment and complete more work during that allotment than do examinees with disabilities, thus benefiting more. Finally, some teacher-made classroom tests have such liberal time limits that extended time is of little value to anyone with or without a disability. On these tests, most students who receive extended time accommodations actually complete their exams within the standard time allotment (see, e.g., Spenceley & Wheeler, 2016). However, when there is at least some time pressure, students with *and* without disabilities tend to benefit from extended time.

Because extended time is so widely desired (and so beneficial), it is especially important to develop guidelines for when it should be given. Although no consensus currently exists on detailed decision-making algorithms for extended time, Phillips (1994) offered five questions that should be asked before providing any accommodation. The questions, which follow, raise issues that remain salient a quarter of a century later:

1. Will format changes or alterations in testing conditions change the skill being measured?
2. Will the scores of examinees tested under standard conditions have a different meaning than scores for examinees tested with the requested accommodation?

3. Would nondisabled examinees benefit if allowed the same accommodation?
4. Does the disabled examinee have any capability for adapting to standard test administration conditions?
5. Is the disability evidence or testing accommodations policy based on procedures with doubtful validity and reliability? (Phillips, 1994, p. 104)

The present chapter will focus on the first two of these questions, presenting recent research results and analysis (for an earlier review of research on all five questions, see Lovett, 2010). Phillips's second—and narrower—question is addressed first: do test scores obtained under extended time conditions have a different meaning than those obtained under standard time conditions? The focus here is on studies examining the predictive validity of test scores obtained with and without extended time accommodations. The chapter then turns to Phillips's first question: do time extensions change the nature of the skill(s) being measured by a test in ways that are problematic, leading to construct underrepresentation for students receiving accommodations and to construct-irrelevant variance given the varying time allotments for different examinees? Even if the meaning of the score changes, does that imply that the test is no longer measuring the construct adequately? This broader question is examined by sampling results from a variety of different types of relevant research. The chapter concludes with recommendations for making decisions about extended time accommodations based on the research reviewed.

Extended Time and Score Meaning

Investigations of the effect of accommodations on score meaning have utilized a wide variety of techniques, including factor analysis and differential item functioning. However, predictive validity is often "where the rubber meets the road"; score users need to know whether an examinee's score has the same probability distribution with regard to a predicted outcome regardless of whether or not the score was obtained with accommodations. This is all the more important now that most if not all testing agencies have stopped "flagging" or "annotating" scores to indicate that the test was administered in a nonstandard fashion (see, e.g., Sireci, 2005).

Admission tests are an obvious area where predictive validity is paramount. Searcy, Dowd, Hughes, Baldwin, and Pigg (2015) conducted an impressive longitudinal predictive validity study of students who took the Medical College Admission Test (MCAT) with either a standard time limit (n = 76,262) or extended time accommodations (n = 449). These investigators obtained data from two sets of outcomes: performance on various parts of the United States Medical Licensing Examination (USMLE) and graduation from medical school. The two groups of MCAT examinees obtained almost identical average MCAT scores (e.g., for the total MCAT scores, d = 0.05), suggesting that if the predictive functions were the same, the two groups' outcomes should also be the same. However, the examinees who obtained MCAT scores with additional time failed the USMLE at far higher rates (about three times as often) and failed to graduate from medical school at similarly higher rates. For instance, 6% of students who took the MCAT under standard time conditions failed Step 1 of the USMLE on their first attempt; 17.9% of students who received extended time on the MCAT failed Step 1 on their first attempt. Differences in graduation rates for standard-time and extended-time examinees are even more striking: 32.8% of extended-time students failed to graduate from medical school in 4 years compared to 13.9% of standard-time students. MCAT scores are clearly not the only factor considered in medical school admissions; as such, Searcy et al. also investigated a model that used both MCAT scores and undergraduate GPA as predictors and found that there was still significant overprediction of USMLE performance and medical school graduation probability among those students who had received extended time accommodations on the MCAT.

The Law School Admission Council (LSAC) has conducted several similar studies. Sweeney, Lauth, Trierweiler, and Pashley (2017) examined the performance of 121,378 law students who had taken the Law School Admission Test (LSAT) without accommodations and 880 students who had received extended time. As with the MCAT, the two groups' LSAT performance was almost identical ($d = 0.05$), but following their first year in law school, the group who received extended time obtained an average grade that was almost half of a standard deviation ($d = 0.43$) lower than the unaccommodated group. Sweeney et al. also found that even when LSAT scores and students' undergraduate GPAs were combined into an index to predict law school performance, the index overpredicted performance of students who received extended time on the LSAT, with residuals between 3 and 4 points on a scale with a standard deviation of 10. Interestingly, Sweeney et al. also ran regression models comparing nonaccommodated students to those who received accommodations *other than extended time* and found that the latter group's law school performance was actually *underestimated* (although with a residual of only about 1 point). Although only one outcome—first-year law school performance—was examined in this study, the LSAC has replicated their results across multiple samples and cohort groups (see Amodeo, Marcus, Thornton, & Pashley, 2009; Thornton, Reese, Pashley, & Dalessandro, 2001, for earlier work with similar findings). The authors of all three studies concluded that LSAT scores obtained with extended time were simply not comparable to those obtained under standard time allotments.

Other studies have been conducted with college admissions tests. For instance, Cahalan, Mandinach, and Camara (2002) found that SAT scores of students with learning disabilities receiving extended time accommodations were substantially weaker predictors of first-year college GPA than were the scores of students who received no accommodations (and who generally had no disabilities). These investigators also found that SAT scores obtained with extended time overpredicted college GPA for male—but not female—students. Including high school GPA in regression models corrected the overprediction for male students, but it caused underprediction of female students' college performance. In this study, then, extended time accommodations appeared to further complicate the problem of differential predictive validity by gender that other research has noted (cf. Zwick, 2006).

Admittedly, these predictive validity studies generally have two limitations. First, they largely confound disability status and time allotment, making it unclear if the differences in prediction functions are due to the extended time *per se* or due to the nature of the students' disabilities. (It would be preferable, albeit impractical, to also examine the performance of nondisabled students with accommodations and students with disabilities who do not receive accommodations.) Second, the studies generally do not record whether students receive accommodations on the outcome measures (e.g., law school exams, or the USMLE). Despite these limitations, score users should know that scores from students who receive extended time accommodations often have a different meaning from other scores, even if the reasons for this are not entirely clear.

Extended Time and Construct Representation

Time limits during tests are often considered to be a source of construct-irrelevant variance (e.g., Lu & Sireci, 2007), because some of the variance in scores will reflect variability in the proportion of the test that different examinees have an opportunity to reach. In some situations, then, extended time accommodations reduce that construct-irrelevant variance by allowing additional time for examinees who need it and ensuring that all examinees have the opportunity to complete the test. Of course, an assumption lurking behind that analysis is that the speed with which examinees complete the test is entirely irrelevant to their skill level. Psychometricians and disability advocates often make this assumption after noting obvious exceptions where speed is part of the intended construct to be measured (e.g., a typing test).

But is the importance of speed really limited to such unusual tests? Or could variability in examinees' test-taking speed actually be a source of construct-*relevant* variance? To put this in a legal context, when is the extension of a time limit a "fundamental alteration" to a test that is not a reasonable accommodation under the Americans with Disabilities Act (Weber, 2010)?

Evidence relevant to answering these questions is spread out widely across different areas of research. Before reviewing that evidence, two red herrings are worth noting; they are irrelevant research findings that have led some scholars (e.g., Jolly-Ryan, 2007; Kelman & Lester, 1997) to doubt the value of speed. First, on any particular test taken by a group of examinees, there often is no substantial linear relationship between how long an examinee takes to complete the test and the accuracy of his or her answers (e.g., Lovett, Lewandowski, & Potts, 2017). However, this research finding only implies that there are roughly as many slow-and-accurate examinees as there are fast-and-accurate examinees. The finding leaves unanswered the question of whether test users would want to rank the fast-and-accurate examinee higher (in skill, competence, etc.) than the slow-but-accurate examinee. A second red herring is that a given examinee will show a greater tendency toward making errors when given a tighter time limit, a phenomenon known as the *speed-accuracy tradeoff* (e.g., Heitz, 2014). However, examinees still vary in how much time they need to achieve a given level of performance (accuracy), again raising the question of whether it is valuable to be accurate under a tight time limit versus only under a generous one. These two common findings, then, raise more questions than they answer about the relevance of speed with respect to skill level.

Direct Evidence of the Construct Relevance of Speed

Although the question of the value of speed is a complex one, at times it has been addressed by very clear and direct evidence. One example of such evidence concerns the licensing examination that is taken by osteopathic medical graduates (those with a DO degree) who want to practice medicine in the United States. Like most standardized licensure exams, the Comprehensive Osteopathic Medical Licensure Examination of the United States (COMLEX-USA) exam series is administered with strict time limits; if medical students and resident physicians with diagnosed disabilities need additional time to complete an exam, they submit an application for such an accommodation to the National Board of Osteopathic Medical Examiners (NBOME). On one hand, the NBOME has a duty to make its exams accessible to examinees with disabilities; on the other hand, the NBOME has a mission to protect the public by maintaining high standards of competence for osteopathic physicians.

NBOME (2013) surveyed 290 faculty members at osteopathic medical schools about "knowledge fluency," which was further defined as the "ability to recall and apply information accurately and *without hesitation*" (emphasis added). Over 93% of respondents agreed that "Knowledge fluency is an osteopathic professional standard of competency critical to the NBOME mission to 'protect the public'." Consistent with this valuing of knowledge fluency, over 85% of respondents *disagreed* with the proposal to remove time limits from COMLEX exams. Almost half of respondents went further and agreed with proposals to report the time that each examinee took to complete each exam, so as to provide further information about knowledge fluency. Based on responses to these and similar questions, the NBOME added the following statement to its *Bulletin of Information* for the COMLEX exams:

> Each COMLEX-USA examination is administered in a standardized, time-measured environment, as the ability to recall, interpret, process, and apply clinical knowledge and skills without hesitation and in a fluid manner ("knowledge fluency") is fundamental to a generalist osteopathic physician's competence to practice osteopathic medicine and therefore is one of the fundamental competencies and skills the COMLEX-USA examinations assess. (NBOME, 2017, p. 5).

It is perhaps worth noting that most COMLEX items involve patient scenarios in which examinees must assess a situation and determine an appropriate diagnostic test, a likely diagnosis, or a useful treatment. Therefore, the cognitive processes used in responding to COMLEX items are likely to overlap substantially with the cognitive processes used by physicians in practice, and if speed is relevant to real-world practice, it is likely to be relevant on the COMLEX as well.

One of the COMLEX exams—Level 2-PE—involves demonstrating relevant clinical skills while interacting with simulated patients (i.e., actors who are trained to simulate having different medical problems). After careful consideration, the NBOME determined that the ability to perform these skills under time pressure was a core competency; as such, the NBOME simply will not grant additional time to complete this portion of the exam sequence (NBOME, n.d.). However, examinees may request additional break time between the different patients or additional time to electronically document their findings from the patient encounters (which is a required component of the test). This decision demonstrates how accommodations policies can be tailored to specific aspects of exams to ensure construct validity, and it further shows the relevance of speed to work in the profession of osteopathic medicine.

Another piece of direct evidence comes from the field of law. Bar exams are timed, and an increasing number of bar examinees request extended time for disability conditions, raising the question of whether speed is part of the construct that the bar exam is designed to measure. Millman (1994, cited in Mehrens, Millman, & Sackett, 1994) studied this question with regard to the New York bar exam by surveying over 200 New York attorneys about “how important for competent lawyering it is” to “be able to read, think, and write under tight time constraints” (p. 43). Questions were posed for each of those three behaviors (reading, thinking, and writing) and for each of nine separate skills that the New York bar exam sought to measure (e.g., identifying and formulating legal issues, generating alternative solutions and strategies, etc.). The vast majority of respondents felt that it was at least somewhat important to be able to read and think under tight time constraints when performing any of the nine skills, and at least 60% of respondents felt the same way with regard to *writing* under tight time constraints. Many respondents (between 20% and 71%) reported that it was *very* important to be able to perform those activities under tight time constraints.

The National Council of Bar Examiners (NCBE) provided updated support for Millman’s findings in its most recent job analysis of new attorneys. As part of the full job analysis, more than 1,600 attorneys who had been licensed for between 1 and 3 years were surveyed about the skills needed by newly practicing attorneys (i.e., those who were relatively close in time to having taken the bar exam; Case, 2013; Nettles & Hellrung, 2012). Respondents were asked to use a 0–4 scale to rate the importance of a variety of skills for performing well as a newly licensed lawyer. “Working within established time constraints” was rated with an average importance of 3.44, which scales between “quite significant” (3) and “extremely significant” (4). Moreover, this average rating was identical to the average rating for “interpersonal skills” and higher than that of many other traits and skills, such as “diligence” (3.26), “advocacy” (3.24), and “interviewing” (2.92). If a bar exam—or, for that matter, a law school exam—is designed at least in part to elicit skills relevant to actual legal practice, speed would seem to be a legitimate part of the construct being measured (see also Pardy, 2016, for a law professor’s argument that his exams are designed in part to measure speed).

As these examples illustrate, direct evidence of the construct relevance of speed is available from subject matter experts and individuals who must actually demonstrate the skills that tests are designed to measure. Such evidence should be obtained anew for each particular test because it is always specific to an individual test with a particular purpose. However, obtaining such evidence does not always require large-scale surveys or other expensive, intensive data collection techniques. For instance, in many educational settings, teachers, professors, and administrators have substantial authority over the curriculum and over determining which skills are important. These professionals, especially when aided by guided reflection, can determine whether speed, fluency, automaticity, and other time-related aspects of a skill are important and expected outcomes in their particular settings. The key issue is that test designers and test score users at every level should pause to consider whether there is any desirability to processing the information in test items and responding to them within a limited span of time.

Indirect Evidence of the Construct Relevance of Speed

In addition to direct evidence from particular settings, psychological theory and research can address the issue of the construct relevance of speed in a more general way. For instance, in behavioral psychology, skill development has long been viewed as happening in a series of steps or stages where *accuracy* (performing a response correctly) comes first, and then comes *fluency* (performing the behavior accurately *and* quickly; e.g., Haring, Lovitt, Eaton, & Hansen, 1978). Research has repeatedly shown the difference between mere accuracy and deft fluency in skills; the time needed to perform a response correctly is an indicator of the depth of a learner's competence. Not only is fluency needed before learners can move on to the *generalization* stage in skill development, which allows them to infer how to solve related problems not encountered before (Martens & Witt, 2004), but fluency is also needed for *retention* of skills over time (Singer-Dudek & Greer, 2005). Obviously, generalization and retention are at the heart of validity arguments for most tests; based on an examinee's responses to a relatively small sample of test item stimuli, test users want to make inferences about how the examinees will respond to similar but distinct stimuli (generalization) at some time point after the exam is over (retention). Another bonus of fluency is *endurance*; learners who are able to perform responses on tests fluently can persist in making those responses for longer periods of time without fatigue or a decline in accuracy (Binder, Haughton, & Van Eyk, 1990). In many industrial assessment contexts, a selection test provides a brief work sample for the examinee to respond to, and so the examinee's fluency would index her/his ability to persist in performing that work for an entire day. Similar examples are found in childhood education; measuring a child's reading *fluency* is a better indicator (than mere accuracy) of that child's ability to persist in reading through a lengthy passage or an entire book chapter without tiring and giving up. In sum, then, fluency—which can only be measured with time limits—is an essential component of high skill levels (for further discussion, see Binder, 1996; Kubina & Morrison, 2000).

Theory and research in cognitive psychology are also relevant; in that field, skill development is studied as well, but it is done so from a perspective interested in the internal mental processes that change as learners' skills improve. Cognitive psychologists who study expertise consistently note that, when compared to novices, experts at a skill are both faster *and* more accurate (Anderson, 2000). Therefore, when we compare across examinees who differ in experience or expertise, or we follow the same person as she/he acquires experience toward expertise, we find that speed and accuracy both improve together, rather than one virtue being sacrificed for the other. For instance, in the domain of medical practice, a classic study compared five groups (junior and senior medical students, residents, general practitioners, and dermatologists) on their ability to identify various skin lesions; as expertise increased, the proportion of correct answers increased in a linear fashion, but the amount of time taken to reach correct answers decreased (Norman, Rosenthal, Brooks, Allen, & Muzzin, 1989). More recently, Nodine et al. (1999) found the same result in the context of judgments of breast lesions based on mammography images. As an explanation of these and similar findings, Kellman (2013) noted that, compared to novices, experts (1) focus selectively on relevant information when solving problems and (2) encode information in terms of larger and more deeply meaningful units, increasing efficiency and speed.

Time pressure, then, tends to allow examinees with deeper knowledge to shine. Woods, Howey, Brooks, and Norman (2006) further tested this claim by teaching undergraduate college students information about various diseases. (To control for any prior medical knowledge, the diseases were entirely fictitious.) One group of students was taught only the signs and symptoms for each disease, so as to be able to recognize each one algorithmically (superficial knowledge). A second group of students was taught the signs and symptoms for each disease along with causal explanations of how the diseases worked (deep knowledge). When presented with new cases of people to diagnose, the students with deep knowledge only outperformed those with superficial knowledge under time-pressured conditions, suggesting that the students with superior understanding showed greater fluency but not greater (untimed) accuracy.

At times, experts may even outperform *themselves* when placed under time pressure, directly contradicting the speed-accuracy tradeoff at a within-person level. Beilock, Bertenthal, McCoy, and Carr (2004) examined putting performance by novice and expert golfers under timed and untimed conditions. (All golfers were undergraduate students, and the "experts" needed to have 2 years of high school golf experience or a handicap below 8.) Novices showed the expected speed-accuracy tradeoff, improving their performance under untimed conditions, but experts *improved significantly under time pressure*. Although putting is obviously different from many of the skills measured by typical tests in educational settings, Beilock et al.'s provocative results should challenge casual assumptions that time pressure has a negative effect on performance for all examinees.

Space limitations preclude detailed discussion of additional work showing the relationship between time pressure, level of skill, and examinee speed and by extension the relationship between increased expertise and the efficiency of problem-solving strategies. Whether involving simple arithmetic questions (Campbell & Austin, 2002) or college physics problems (Lasry, Watkins, Mazur, & Ibrahim, 2013), and whether using behavioral measures of skill (Furlan, Agnoli, & Reyna, 2016) or neuroscience tools examining activation of parts of the brain associated with different cognitive processes (Price, Mazzocco, & Ansari, 2013), there is ample evidence that speeded problem-solving forces examinees to rely on skills associated with deeper levels of competence.

Implications for Extended Time Accommodations Policies

If time extensions have the potential to alter the meaning of scores and prevent tests from measuring what they are designed to measure, are extended time accommodations necessarily inappropriate? In a word, no. However, the research reviewed in this chapter suggests a need for caution and care when making decisions about altering the time limits for some examinees and not others. A comprehensive discussion of accommodation decision-making procedures is beyond the scope of the present chapter and must take into consideration laws and regulations, the ability of students with disabilities to adapt to standard conditions, and the complex issue of just how much additional time is needed (for coverage of all of these topics, see Lovett & Lewandowski, 2015). However, the rest of this section provides a brief overview of a psychometric framework for how decisions ideally would be made.

A first step toward appropriate decisions involves distinguishing between two sets of skills needed to succeed on a test: *target skills*, which the test is designed to measure, and *access skills*, which are assumed to be present in adequate levels in all examinees to allow for meaningful participation in the test (Ketterlin-Geller, 2008). For instance, doing well on a typical exam in a college anthropology course requires knowledge of the anthropology content on the test (a target skill) but also adequate vision (an access skill). Variance in target skills leads to construct-*relevant* variance in test scores; variance in access skills leads to construct-*irrelevant* variance in test scores. The presence of examinees with disability conditions can increase both types of variance, because a disability condition may cause an examinee to have low levels of target skills, access skills, or both, depending on what a particular test is designed to measure.

Test developers, in consultation with subject matter experts and test users, should carefully consider whether any trait that might vary with test-taking speed (e.g., fluency, automaticity, using problem-solving processes requiring deep competence) is a target skill. If after thoughtful and searching consideration there really is no place for speed-related traits in the target skill set, the time limits should be made very liberal and examinees should be able to request even more additional time (up to some logistically practical amount) without a need for extensive review of disability documentation. This is consistent with principles of "universal design" procedures for increasing accessibility of assessments for all examinees (Ketterlin-Geller, 2005). However, the research reviewed in this chapter suggests that speed-related traits often will be (or should be) among the test's target skills. What then?

Disability experts have a key role in determining whether a particular examinee has deficits in access skills that will lead to a need for accommodations, including extended time. At independent testing agencies, higher education institutions, and private K-12 schools, disability experts (full-time internal employees and/or contracted disability consultants) who have (a) detailed knowledge of the diagnostic assessment procedures used to identify disability conditions as well as (b) knowledge of the task requirements of the tests for which accommodations are being requested should review disability documentation submitted by examinees requesting additional time. A similar process exists in public K-12 schools, where clinically trained professionals conduct a special education evaluation and then review the data with a larger set of educational and administrative professionals. The professionals should determine if a legitimate disability condition (e.g., a learning disability) is present (based on whether the submitted data shows that official diagnostic criteria for the condition are met), and if so, the evidence for deficits in access skills should be carefully scrutinized. If the examinee has a deficit in one or more access skills, accommodations then may be appropriate. In particular, if the examinee has a deficit in a speed-related access skill, extended time accommodations may be appropriate.

Consider the example of Susan, a 21-year-old college student with a diagnosis of a learning disability in reading who is applying for accommodations on a test used for admission to graduate/professional schools. Susan might submit documentation that includes reports from psychological evaluations, transcripts from college and high school, score reports from college admissions tests, information about when and where she has received accommodations before, and a personal statement describing how her learning disability affects her life, including her educational and test-taking experiences. A disability professional can determine whether Susan has provided sufficient evidence of actually meeting the official diagnostic criteria for a learning disability in reading and then can go on to search for evidence of deficits in relevant access skills. Perhaps Susan has consistently obtained below-average scores on norm-referenced diagnostic tests measuring her reading fluency and timed reading comprehension skills, and she also performed at the 10th percentile on the SAT critical reading section without accommodations but at the 40th percentile on the ACT reading section with extended time; this generally would be evidence consistent with a need for additional time to access exams, although the disability professional should always take all of the documentation into account.[2]

Variations in Susan's case show some of the complexities of accommodations decision-making. If Susan were requesting accommodations on a test where reading speed or reading fluency were part of the target skill set, her deficits in these skills would *not* make extended time appropriate; a student who had orthopedic problems leading to slowed motor speed (an access skill rather than a target skill) still might require extended time on such a test. If Susan showed no deficits in timed reading skills but had evidence suggesting that she takes longer than most people to retrieve information from memory and apply that information to problem-solving, extended time might be inappropriate because a test's target skills might include being able to recall and apply information under time-limited conditions. In short, it is difficult to describe how decisions should be made without detailed information about the examinee's disability evidence and similarly detailed information about the target skills and access skills of the exam in question.

Conclusions

Although the term "standardized test" is used in multiple ways, it originates in the understanding that tests should be administered in the same way to different examinees so that resulting score variability is attributable to variability in examinee skill levels and not administration procedures. Research shows that even slight changes in administration conditions can sometimes affect scores, and altering time limits often *will* affect scores as was reviewed above (Lee, Reynolds, & Wilson, 2003). At the same time, some examinees have disability conditions that will prevent them from meaningfully and fairly participating in a test unless some aspect of the administration procedures is altered. Due to these two opposing considerations—standardization and the ensuring of access—testing entities should be hesitant to alter time limits, *unless there is evidence that a particular examinee needs such an alteration due to deficits in access skills*. "Evidence" is the key word here. Rather than assuming that a test does not intend to measure anything that is speed-related or assuming that because an examinee reports a disability condition she/he meets the criteria for the condition and requires a time extension, ideally evidence should be sought at every point where decisions are made.

Notes

[1] The author wishes to thank Lawrence Lewandowski for comments on a draft of the present chapter, as well as Cynthia Searcy and Marc Kroopnick for suggestions on the chapter's coverage and structure, and discussion of their own relevant research.

[2] Just how much extended time to provide to Susan would also depend on a complete review of her disability documentation as well as information about the test that she is about to take. In the United States, examinees are most commonly granted 50% or 100% extended time, although other countries routinely grant 33%, 25%, or even just 10 additional minutes per hour of testing (about 17%). The few empirical studies comparing different amounts of extended time suggest that we in the United States may be overaccommodating even many students with genuine disabilities, allowing these students to complete (access) more test items with accommodations than nondisabled students can complete under standard time; see Lovett and Lewandowski (2015) for discussion.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Amodeo, A., Marcus, L. A., Thornton, A. E., & Pashley, P. J. (2009). *Predictive validity of accommodated LSAT scores for the 2002-2006 entering law school classes* (LSAT Technical Report 09-01). Newtown, PA: Law School Admission Council.

Anderson, J. R. (2000). *Cognitive psychology and its implications* (5th ed.). New York, NY: Worth.

Beilock, S. L., Bertenthal, B. I., McCoy, A. M., & Carr, T. H. (2004). Haste does not always make waste: Expertise, direction of attention, and speed versus accuracy in performing sensorimotor skills. *Psychonomic Bulletin & Review*, *11*(2), 373–379.

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst*, *19*(2), 163–197.

Binder, C., Haughton, E., & Van Eyk, D. (1990). Increasing endurance by building fluency: Precision teaching attention span. *Teaching Exceptional Children*, *22*(3), 24–27.

Cahalan, C., Mandinach, E. B., & Camara, W. J. (2002). *Predictive validity of SAT I: Reasoning Test for test-takers with learning disabilities and extended time accommodations* (College Board Research Report 2002-5). New York, NY: College Board.

Cahan, S., Nirel, R., & Alkoby, M. (2016). The extra-examination time granting policy: A reconceptualization. *Journal of Psychoeducational Assessment*, *34*(5), 461–472.

Campbell, J. I., & Austin, S. (2002). Effects of response time deadlines on adults' strategy choices for simple addition. *Memory & Cognition*, *30*(6), 988–994.

Case, S. M. (2013). The NCBE job analysis: A study of the newly licensed lawyer. *Bar Examiner*, *82*(1), 52–56.

Elliott, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children*, *70*(3), 349–367.

Furlan, S., Agnoli, F., & Reyna, V. F. (2016). Intuition and analytic processes in probabilistic reasoning: The role of time pressure. *Learning and Individual Differences*, *45*, 1–10.

Haring, N. G., Lovitt, T. C., Eaton, M. D., & Hansen, C. L. (1978). *The fourth R: Research in the classroom*. Columbus, OH: Merrill.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, article 150.

Heward, W. L. (2013). *Exceptional children* (10th ed.). Boston: Pearson.

Jolly-Ryan, J. (2007). The fable of the timed and flagged LSAT: Do law school admissions committees want the tortoise or the hare? *Cumberland Law Review*, *38*, 33–70.

Kellman, P. J. (2013). Adaptive and perceptual learning technologies in medical education and training. *Military Medicine*, *178*(10S), 98–106.

Kelman, M., & Lester, G. (1997). *Jumping the queue: An inquiry into the legal treatment of students with learning disabilities*. Cambridge, MA: Harvard University Press.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *The Journal of Technology, Learning and Assessment*, *4*(2). Available from http://www.jtla.org

Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice*, *27*(3), 3–16.

Kubina, R. M., & Morrison, R. S. (2000). Fluency in education. *Behavior and Social Issues*, *10*, 83–99.

Lasry, N., Watkins, J., Mazur, E., & Ibrahim, A. (2013). Response times to conceptual questions. *American Journal of Physics*, *81*(9), 703–706.

Lee, D., Reynolds, C. R., & Willson, V. L. (2003). Standardized test administration: Why bother? *Journal of Forensic Neuropsychology*, *3*, 55–81.

Lewandowski, L., Lambert, T. L., Lovett, B. J., Panahon, C. J., & Sytsma, M. R. (2014). College students' preferences for test accommodations. *Canadian Journal of School Psychology*, *29*(2), 116–126.

Lewandowski, L. J., Lovett, B. J., Parolin, R., Gordon, M., & Codding, R. S. (2007). Extended time accommodations and the mathematics performance of students with and without ADHD. *Journal of Psychoeducational Assessment*, *25*(1), 17–28.

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, *80*(4), 611–638.

Lovett, B. J., & Leja, A. M. (2013). Students' perceptions of testing accommodations: What we know, what we need to know, and why it matters. *Journal of Applied School Psychology*, *29*(1), 72–89.

Lovett, B. J., & Lewandowski, L. J. (2015). *Testing accommodations for students with disabilities: Research-based practice*. Washington, DC: American Psychological Association.

Lovett, B. J., Lewandowski, L. J., & Potts, H. E. (2017). Test-taking speed: Predictors and implications. *Journal of Psychoeducational Assessment*, *35*(4), 351–360.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29–37.

Martens, B. K., & Witt, J. C. (2004). Competence, persistence, and success: The positive psychology of behavioral skill instruction. *Psychology in the Schools*, *41*(1), 19–30.

Mehrens, W. A., Millman, J., & Sackett, P. R. (1994). Accommodations for candidates with disabilities. *Bar Examiner*, *63*(4), 33–47.

National Board of Osteopathic Medical Examiners, Inc. (2013). Knowledge fluency and time limitations. Unpublished manuscript.

National Board of Osteopathic Medical Examiners (NBOME). (2017). *COMLEX-USA Bulletin of Information 2017-2018*. Retrieved from www.nbome.org

National Board of Osteopathic Medical Examiners (NBOME). (n.d.). Request for test accommodation instructions. Retrieved from www.nbome.org

Nettles, S. S., & Hellrung, J. (2012). *A study of the newly licensed lawyer*. Applied Measurement Professionals.

Nodine, C. F., Kundel, H. L., Mello-Thoms, C., Weinstein, S. P., Orel, S. G., Sullivan, D. C., … Conant, E. F. (1999). How experience and training influence mammography expertise. *Academic Radiology*, *6*(10), 575–585.

Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of Dermatology*, *125*(8), 1063–1068.

Pardy, B. (2016). Head starts and extra time: Academic accommodation on post-secondary exams and assignments for students with cognitive and mental disabilities. *Education and Law Journal*, *25*, 191–208.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, *7*(2), 93–120.

Price, G. R., Mazzocco, M. M., & Ansari, D. (2013). Why mental arithmetic counts: Brain activation during single digit arithmetic predicts high school math scores. *Journal of Neuroscience*, *33*(1), 156–163.

Searcy, C. A., Dowd, K. W., Hughes, M. G., Baldwin, S., & Pigg, T. (2015). Association of MCAT scores obtained with standard vs extra administration time with medical school admission, medical student performance, and time to graduation. *Journal of the American Medical Association*, *313*(22), 2253–2262.

Shaywitz, S. (2003). *Overcoming dyslexia*. New York, NY: Vintage.

Singer-Dudek, J., & Greer, R. D. (2005). A long-term analysis of the relationship between fluency and the training and maintenance of complex math skills. *The Psychological Record*, *55*(3), 361–376.

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher*, *34*(1), 3–12.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457–490.

Spenceley, L. M., & Wheeler, S. (2016). The use of extended time by college students with disabilities. *Journal of Postsecondary Education and Disability*, *29*(2), 141–150.

Sweeney, A. T., Lauth, L. A., Trierweiler, T. J., & Pashley, P. J. (2017). *Predictive validity of accommodated LSAT scores for the 2007-2011 entering law school classes* (LSAT Technical Report 17-04). Newtown, PA: Law School Admission Council.

Thornton, A. E., Reese, L. M., Pashley, P. J., & Dalessandro, S. P. (2001). *Predictive validity of accommodated LSAT scores* (LSAT Technical Report 01-01). Newtown, PA: Law School Admission Council.

U.S. Government Accountability Office. (2011). Higher education and disability: Improved federal enforcement needed to better protect students' rights to testing accommodations. Report GAO-12-40.

Weber, M. C. (2010). Unreasonable accommodation and due hardship. *Florida Law Review*, *62*, 1119–1178.

Wollack, J. A., & Case, S. M. (2016). Maintaining fairness through test administration. In N. J. Dorans & L. L. Cook (Eds.), *Maintaining fairness through test administration* (pp. 33–53). New York, NY: Routledge.

Woods, N. N., Howey, E. H., Brooks, L. R., & Norman, G. R. (2006). Speed kills? Speed, accuracy, encapsulations and causal understanding. *Medical Education*, *40*(10), 973–979.

Yellin, D. (2017, January 26). Growing number of students seeking accommodations for SAT. Retrieved from www.northjersey.com

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (pp. 647–679). Westport, CT: Praeger.

5

Relationship between Testing Time and Testing Outcomes

Brent Bridgeman

In many testing situations, the primary reasons for imposing time limits are administrative convenience and reduced cost. In these cases, whether test takers can work quickly as well as accurately is not assumed to be a part of the construct that is being assessed, and the impact of the time limit on scores would be considered a source of construct-irrelevant variance. Alternatively, if the ability to work both quickly and accurately *is* part of the construct being assessed (e.g., on an educational or psychological test), it is to be expected that time limits can and should impact test scores. Descriptions of the constructs to be assessed by high-stakes admissions tests such as the SAT®, ACT®, and Graduate Record Examinations General Test (GRE®) suggest that speed is at best a minimal part of the construct being assessed. The claim for the current version of the SAT Reading Test is as follows:

> The redesigned SAT's Reading Test is intended to collect evidence in support of a broad claim about student performance: Students can demonstrate college and career readiness proficiency in reading and comprehending a broad range of high-quality, appropriately challenging literary and informational texts in the content areas of U.S. and world literature, history/social studies, and science (College Board, 2018a, pp. 41).

For the SAT Math Test, the claim is:

> The redesigned SAT's Math Test is intended to collect evidence in support of the following claim about student performance: Students have fluency with, understanding of, and the ability to apply the mathematical concepts, skills, and practices that are most strongly prerequisite and central to their ability to progress through a range of college courses, career training, and career opportunities (College Board, 2018a, pp. 132).

Although the word "fluency" in the Math claim might suggest a speed component, it is not addressed in any of the subsequent descriptions.

Similarly, the ACT *Technical Manual* says very little about the importance of rapid responding, but in the section on item tryouts it indicates that, "The time limits for the tryout units permit the majority of students to respond to all items." This suggests that speed of responding is not a part of the intended construct (ACT, 2019). A description of the version of the GRE that was introduced in 2011 notes, "… it is specified that no test section should be delivered under speeded conditions" (Robin & Steffen, 2014). Given that (1) speed is not part of the intended construct for high-stakes admissions tests but that (2) these tests have time limits such that at least some students struggle to finish within the time allowed for the test, it is critical for test publishers to establish the effects of the time limits on various item statistics and most crucially on students' scores.

This chapter focuses primarily on the effects of time limits on admissions tests and K-12 accountability assessments in which rapid responding is not part of the construct that the test is intended to measure. Different issues can be at play in licensing tests in which speed of responding can be a legitimate issue (e.g., would you want to give a pilot's license to someone who came up with the appropriate action in an emergency situation only after taking a few minutes to respond?), and such tests are not covered here. The complex issues in a medical licensing context also are not addressed here, but see Chapter 6 for a discussion of the relevant issues in this context. Similarly, speededness concerns for essay tests are discussed in Chapter 7.

The organization of this chapter is essentially chronological. It begins in the 1940s when completion statistics were the primary tool for addressing speededness, moves on to the 1970s, when concerns about group fairness with speeded tests became a major issue, and then proceeds to discuss the new speededness issues that emerged with the introduction of computerized adaptive tests (CATs). Issues with the speededness of state accountability assessments are then briefly discussed. Finally, some suggestions for future research efforts are presented.

Early Research on the Impact of Time Limits on Scores

Research from the 1940s and 1950s

Educational Testing Service (ETS) was founded in 1947, and one of the first research reports that was produced was entitled *Item-Analysis Data from an Experimental Study of the Effects on Item-Analysis Data of Changing Item Placement and Test Time Limit* (Mollenkopf, 1949). The study reached the wholly unremarkable conclusion that items in speeded tests are more difficult (i.e., have a lower proportion correct) compared to the difficulty for those same items in unspeeded test administrations. Specifically, the author concluded, "The proportion right of those attempting the item, the Delta index, and the biserial *r* were all found to have undesirable characteristics for items appearing late in a speeded test." Although today this conclusion is obvious, such a study was likely important in the early days of large-scale testing; at that time, there may have been a belief by some that parameters such as item difficulty were inherent in items in specific populations rather than dependent on where in the test that item was placed. This issue of parameter determination and item placement remains equally relevant today and is a concern because the incorrect specification of item difficulty can affect final examinee scores.

Another relatively early ETS research report (Lord, 1954) explored the impact of speed factors on test validity. Tests of vocabulary, spatial relations, and arithmetic reasoning were administered to students at the U.S. Naval Academy. Included in the battery were speeded and unspeeded—but otherwise parallel—tests of vocabulary, spatial ability, and arithmetic reasoning. The unspeeded vocabulary test included 15 items in 7 minutes, while the most highly speeded version had 75 items in 5 minutes. The percent of examinees finishing was 97% for the unspeeded version and 2% for the speeded version, but Lord pointed out that these finishers likely included rapid random guessers. (Note that with modern computer administrations, such rapid random responding can be accurately tracked; this technology was not available in 1954.) The speeded and unspeeded tests were analyzed together in a maximum-likelihood factor analysis and ten factors were extracted. Most factors did not have a speed component (e.g., unspeeded tests of verbal reasoning and mathematical reasoning), but factors containing the speeded verbal tests and speeded perceptual tests were identified. Small positive correlations were found between the speed factors and grades, suggesting that speededness could produce a small improvement in predictive validity of grades in the Naval Academy.

Swineford Guidelines

In 1949, Francis Swineford, a psychometrician at ETS, published a statistical report on characteristics of the Law School Admissions Test (LSAT; Swineford, 1949). Table 5.1 in that report presented information on the speededness of the test sections with the table description stating, "It is generally assumed that timing is satisfactory when it allows about 80 percent of the candidates to reach the last item." Although this 80% guideline later would be incorporated into what became known as the Swineford guidelines, it is worth noting that Swineford did not invent this guideline; it was already "generally assumed" in 1949. Of the ten sections on the LSAT at the time, only five met the 80% guideline, and on two of the sections fewer than half of the examinees reached the last item. This led Swineford to conclude that, "it is clear that two sections are much too highly speeded." Although these sections were clearly speeded, there could not be any assurance that sections that were completed were unspeeded; the score on each section simply was the number correct, and there was no correction for guessing. Completion would not be a meaningful indicator, because testwise examinees would randomly guess to complete a section before allowing time to expire. It therefore appears that LSAT test takers in 1949 either were not testwise or reflected cultural norms about the appropriateness of random guessing or of responding to an item if they didn't have a good idea of the answer. This assumption that test speededness could be assessed through simple inspection of completion rates permeated much of the early research on test speededness and psychometric tools that relied on estimates of test speededness.

Gulliksen provided a method for estimating the reliability of speeded tests given "some reasonable approximation of the 'number of items' *K* in the speeded part of the test …" (Gulliksen, 1950). But there is no way to determine what the "speeded part of the test" actually is. Some might naively assume that the speeded part of the test is represented just by the items near the end of the test. Indeed, this might have been a reasonable assumption in 1950 when most students were not coached on effective strategies for dealing with tests with strict time limits. Students would simply start with the first item and continue answering items in order until time ran out. Lord also made this assumption when he published a Research Bulletin entitled *A Method for Estimating from Speeded Test Data the Power Condition Scores and Item Difficulties* (Lord, 1950). In order to use this method, Lord noted, "It is assumed that all items are scored 0 or 1

Table 5.1 Final scores of examinees with identical ability estimates on item 29 (Theta = 1.0) by amount of time available to complete the section

Time to Complete Section	*n*	GRE-A *M*	GRE-A *SD*	GRE-Q *M*	GRE-Q *SD*	GRE-V *M*	GRE-V *SD*
More time	400	664	32	632	108	479	100
Less time	419	639	43	631	100	520	97

(there is no correction for guessing), that the examinee reads the items in the order in which they are administered, and that he does not attempt items that he has not read." Today, we could say that Lord's method would work only if examinees behaved irrationally.

This line of thinking continued in 1956 when Swineford made a more formal declaration of speededness guidelines in the *Technical Manual for Uses of Test Analyses* (Swineford, 1956). This manual asserted,

> A test may be considered unspeeded if:
>
> 1. virtually all candidates reach 75% of the items, and
> 2. at least 80% of the candidates respond to the last item.

Note that the Swineford guidelines make sense only if examinees behave in a way that rational examinees in the 21st century would not be expected to behave. Examinees now are given explicit advice that runs counter to the Swineford, Gulliksen, or Lord assumptions that students will seriously consider each item in the order presented, continue until time runs out, make no random guesses at the end of the test, and not revisit any previously seen items. In advice from the College Board, examinees are advised to *not* just proceed in order from the first item to the last. Specifically,

> Don't dwell on questions that stump you. Circle ones you decide to skip so that you can return to them quickly later. Remember that a question you answer easily and quickly is worth as much as a question that you struggle with or take a lot of time on. (College Board, 2018b)

The College Board guidelines go on to say,

> Remember that there's no penalty for guessing, so you should answer all questions before time is up. When you're not sure of an answer, make an educated guess. (College Board, 2018b)

Therefore, if used in the 21st century, the Swineford guidelines could make a truly speeded test look unspeeded if examinees skip early items (hoping to return to them later), answer the last item, and then time runs out before they can return to the skipped items. By conventional definition, they have reached the last item in the test, which therefore is declared to be unspeeded. The opposite problem also can exist: Use of the Swineford guidelines can make an unspeeded test that has a correction for guessing appear to be speeded because examinees may intentionally omit the last few difficult items at the end of the test. On a test with no correction for guessing (which describes almost all tests now), even when 100% of the examinees answer 100% of the questions, the test could be speeded or unspeeded (i.e., scores *might* have been higher if examinees had more time). Given the ambiguity that results when attempting to assess the speededness of a test from looking exclusively at completion statistics, random-assignment experimental studies may be needed to adequately estimate the speededness of any test. When Lord (1954) attempted to assess the contribution of speed factors to predicting grades in academic courses, he noted, "The exact degree of speededness that will result in the highest validity for the admissions examinations can only be determined by an experimental study of the admissions examinations themselves." By the 1970s and beyond, experimental studies to determine the impact of speededness on test scores became more common, although some version of the Swineford assumptions continued to be used well into the 21st century (e.g., Shao, Li, & Cheng, 2015). As experimental methods that did not rely exclusively on test completion statistics became more common, focus shifted to how these methods could be used to address the most pressing concerns related to test speededness; of particular concern was investigating how speededness might differentially impact the performance of minority group members.

Differential Impact of Speededness on Examinee Subgroups

The 1970s brought increased interest in investigating score differences across gender and ethnic groups and in identifying test characteristics—such as speededness—that might be related to subgroup score differences. In 1972, a study by Evans and Reilly—entitled *A Study of Speededness as a Source of Test Bias*—sought to determine whether time limits on the Law School Admissions Test (LSAT) were differentially impacting Black students. They reduced the number of items to produce an experimental form that was less speeded than an experimental form with the standard time limit and randomly distributed both forms to students in regular test centers and in special centers that had been established to serve colleges whose students were predominantly Black. In these special centers, the usual testing fee was waived; the authors referred to the students in these centers as "fee-free candidates." Within these fee-free centers, 230 students received the speeded form and 235 received the less-speeded form. In the regular centers, approximately 5,000 students took each form. The experimental forms were administered as part of an operational administration, but the scores on these forms were separate from the operational scores. Because students with both the speeded and unspeeded forms needed to be tested in the same room, the time limit had to be the same for all students (40 minutes). Again, speededness was manipulated by reducing the number of questions to create a less-speeded form: The standard form had four reading passages with 35 reading comprehension questions (eight or nine questions per passage), and the less-speeded form had three passages and a total of 27 questions. The dependent variable was the score on the 27 questions that were common to both forms. The study authors did not ask the participants to explicitly state their ethnicities, but nearly all of the "fee-free candidates" were Black and only a very small fraction of the students in the regular centers were from minority groups. Although the authors did not need to rely on applying the Swineford guidelines to answer their research question, they nevertheless did observe that with the standard timing fewer than 70% of the fee-free candidates answered the last question compared to 90% of the students in the regular test centers. Because the LSAT had no correction for guessing, any well-coached student should complete 100% of the test; the results therefore suggest that in 1972 most Black students were not well coached. In the experimental analysis of the test scores (not dependent on applying the Swineford guidelines), students from regular test centers scored about 22 scale score points higher on the less-speeded form (with a scaled score standard deviation of about 100) and students from the fee-free centers scored about 33 scale score points higher on the less-speeded form. Although all students gained from extended time, the gain for the fee-free candidates was not significantly larger than the gain for the regular candidates; the authors therefore concluded, "reducing speededness is *not* more beneficial (in terms of increasing the number of items answered correctly) to fee-free than to regular center candidates." The authors were accepting the null hypothesis (rather than failing to reject it), but the sample size was sufficiently large that it was reasonable to assume that whatever differential gain existed from reducing speededness was likely to be trivial. Furthermore, because students in the fee-free centers were less likely to randomly guess as they were running out of time on the speeded form, the gain from using the unspeeded form was slightly overstated for these students.

Wild, Durso, and Rubin (1982) used another approach to evaluate differential speededness for women and minorities (Black/White only) on the GRE. Rather than having forms with differing numbers of questions that could be administered within centers where the time limit was fixed, they had different timings in different test centers. Although long or short forms were randomly distributed to test centers, the authors could not assume that the students in centers where the test section had a short time limit were randomly equivalent to students in centers where the section had a longer time limit. Therefore, they used the subgroup means in a center (rather than persons) as the unit of analysis. The experimental test was administered following an operational test administration in approximately 553 domestic test centers; 250 centers used the standard 20-minute time limit and 253 centers used a 30-minute (extended) time limit. Within each center, a random half of the forms contained an experimental verbal form and the other half contained the experimental quantitative section. The verbal form consisted of 26 questions (11 passage-based reading comprehension questions and 15 discrete questions); the 20-minute time limit allowed 46 seconds per question and the 30-minute time limit allowed 69 seconds per question. The quantitative test consisted of 14 questions; this translates to about 86 seconds per question for the 20-minute timing condition and about 129 seconds per question in the 30-minute timing condition. At the time of this study, the GRE had a correction for guessing. The dependent variable was the mean formula score (across centers) for each subgroup; the operational formula score for the corresponding (Verbal or Quantitative) test section in each center was used as a covariate. In both the gender and ethnicity subgroups, scores were higher in the extended-time administrations, but the gains were approximately equal across subgroups; the authors concluded that test-taking time is not a "biasing agent" for gender or Black/White subgroups. Following the focus on subgroups based on gender and ethnicity, attention shifted to another subgroup with increasing visibility: students with disabilities. Although there were no a priori reasons to expect timing impacts on gender and ethnic groups, impacts for students with disabilities were somewhat different because they routinely could get an accommodation for extended time. This accommodation was intended to make the assessments fairer for all applicants, but there were some concerns that some students were gaming the system to get possible benefits from extended time even in the absence of a legitimate disability.

The Impact of Time Limits on Scores by Testing Context

Timing Accommodations and Effects of Extra Time on High-Stakes Admissions Tests

In 1998, a test candidate took the Graduate Management Admissions Test (GMAT) to apply to business school. The candidate had no hands and so was allowed extra time to take the test. At the time, nonstandard test administrations, including extended time, resulted in an indication on the score report that the test had been taken under nonstandard conditions. The candidate did not question the appropriateness of the accommodation, but he did sue ETS for flagging his test score. He argued that the flag was a violation of the Americans with Disabilities Act because it identified him as a person with a disability and could suggest that his score was artificially inflated. The flag was consistent with published professional standards that suggested that scores that were obtained under nonstandard conditions should be identified to test users. Thus, there was a conflict between the established professional standards and the implications of the federal law. ETS initially set out to defend flagging, but as depositions were being obtained, there was a change in leadership at ETS and the new president wondered why ETS was, from his perspective, defending the wrong side. ETS changed its position, settled the suit, and put procedures in place to drop flags for almost all accommodations including extra time. After reviewing its procedures, the College Board followed suit and decided to drop the flag for SAT scores.

Although there was little concern with granting accommodations to students with a well-documented need, there was a concern that some parents would pay to get a diagnosis that might not be fully legitimate. This would be less of a concern if the benefits of extra time were not substantial, but little was known about the effects of extra time on SAT scores. The College Board commissioned a large-scale study to explore this question.

The study was conducted in the fall of 2000 using special test sections that were part of operational SAT tests. Because some of the results were surprising, a follow-up study was conducted in the fall of 2001 to determine whether the initial results could be replicated using the same procedures but a different set of questions. Results for both studies were published together (Bridgeman, Trapani, & Curley, 2004). Because of the very large sample sizes and replication of some surprising results, these studies are discussed in considerable detail below.

In 2000, every operational form of the SAT contained two 30-minute and one 15-minute verbal sections and two 30-minute and one 15-minute math sections. In addition, each test contained a 30-minute variable section with either verbal or math questions. This section, which did not contribute to a student's score, could be used for test equating, trying out new questions, or other experimental purposes. The questions in this section looked like questions in the operational sections and had the same time limit as the operational sections. The test takers were not told which section was the variable section, so motivation on this section was comparable to motivation on the operational sections. Because the timing for this section was fixed at 30 minutes, time per item was manipulated by creating forms with differing numbers of items. Specifically, ten forms were created: four forms (two verbal and two math) had the standard number of items, two forms (one verbal and one math) had a reduced number of items equivalent to time-and-a-quarter, and two forms were equivalent to time-and-a-half. To create the short forms, items were deleted from various positions in the full-length forms; just deleting the items at the end would have produced an easier form because the items at the end tended to be more difficult. However, the last item in the full-length form was retained as the last item in each shortened form. Analyses were based on the items that were common across all timings (e.g., the 23 items comprising the shortest verbal form were the same 23 items that were scored when embedded in the longer verbal forms). The ten experimental forms were spiraled in each batch of test forms sent to a testing location such that the distribution was random in effect. As part of a large-scale national administration, sample sizes were over 8,000 examinees for each of the ten forms. In order to investigate possible differential effects of extra time for students at different ability levels, three ability groupings were created based on scores on the relevant (verbal or math) operational sections on the standard 200–800 scale. Most students were in the middle group (410–600), but in each form there were over 800 examinees in the lower group and 1,600 in the upper group.

Analyses were conducted at both the individual item level and at the form score level. One of the surprising findings for anyone who believed in the relevance of the Swineford guidelines was that effects of a less-speeded test could affect item performance throughout the test and not just on the last few items in the test. For item 9 out of the 23 common items in one of the verbal forms, the proportion correct was 0.64 in the 35-item test and 0.70 in both of the shorter forms. On common item 12 out of the 17 common items on a math form, the proportion correct was 0.51 in the long form and 0.61 in the shortest form. Similar results were found for an experimental administration of the computer-delivered version of the ACT under three speededness conditions (Li, Yi, & Harris, 2016). For Reading item 30 out of 40, the proportion correct was 0.50 with the standard timing and about 0.65 in both extended-time conditions. In both the ACT and SAT studies, differences across timing conditions were actually smaller for items near the end of the test than for some items earlier in the test. There have been recent attempts to assess speededness by examining items only near the end of the test or beyond the "change point" (e.g., Shao, Li, & Cheng, 2015); these attempts may have some value in identifying extreme cases, but they should be interpreted cautiously as they require belief in the assumption that speededness affects only items near the end of a test and this assumption has been proven to be incorrect.

In the Bridgeman et al. (2004) study, speededness effects at the total verbal form score level, when expressed on the familiar 200–800 scale, were less than 7 points for both forms; differences for the math forms were somewhat more variable. For the form that consisted primarily of quantitative comparison items, speededness effects were small, again under 7 points on the

200–800 scale. But for the form with more standard multiple-choice questions, speededness effects varied considerably by ability level. Effects were greatest for moderately high scores at about 30 points in the first study and 20 points in the replication study a year later. Effects for the lowest ability level (below 400) were very small or negative. It might seem odd that extra time could actually produce lower scores, but this is related to the way SAT scores were computed at the time (i.e., with a correction for guessing that penalized an incorrect answer more than an omitted answer) and the way forms were assembled (with the most difficult items near the end). With extra time, low-ability examinees would attempt difficult items, select attractive but incorrect distractors, and therefore get a lower score than they would have gotten if they simply ran out of time and omitted these difficult questions. Consistent with findings from earlier studies (e.g., Evans & Reilly, 1972; Wild, Durso, & Rubin, 1982), speededness did not appear to contribute to racial/ethnic or gender differences. Indeed, with a total sample size of over 8,000 per form, *not* finding any significant (p <0.05) interactions of ethnicity or gender with timing condition is quite remarkable.

A final part of the Bridgeman et al. study examined the validity of scores from speeded and less-speeded forms by correlating the scores from these forms with grades assigned in high school math courses. Because speed was not supposed to be part of the construct assessed by the SAT Math Test, it could be argued that the test should be better able to predict valued outcomes as the irrelevant effects of test speededness were reduced. But on the other hand, it could be argued that extended time allows students to use strategies, such as working backward from the answer choices, which would result in poorer measurement of math reasoning skills. Results indicated essentially equal correlations for the speeded and less-speeded tests.

The timing study for the computerized ACT (Li et al., 2016), mentioned above with respect to item-level results, noted a decrease in the number of omitted items with extended time but did not evaluate the impact on total scores with different timing conditions. For all four scales (English, Mathematics, Reading, and Science), there was a substantial reduction in omit rates with an extra 10 minutes of testing time. In the Reading test, for example, 36% of the students omitted three or more items with standard timing and this dropped to 18% with the extra 10 minutes; the difference was even greater for the Science test; omit rates went from 27% with standard time to 5% with extended time. Although the results strongly suggest the presence of a speed factor, they are difficult to fully understand because there should be no omitting for a test with no guessing correction. Furthermore, as acknowledged by the study authors, time-related effects can be especially difficult to interpret on a test with no stakes for the examinees.

Research on score gains with extended time for other high-stakes admissions tests—such as the Medical College Admissions Test (MCAT)—is rare. MCAT timing research appears to be limited to effects of extra time as an accommodation for students with disabilities (e.g., Searcy, Dowd, Hughes, Baldwin, & Pigg, 2015). See Chapter 4 for a discussion of the effects of extra time as an accommodation.

Computer-Adaptive Tests

Speededness can have an especially dramatic effect on scores on a CAT. In a CAT, the difficulty level of the next item presented depends on the correctness of the response to the current item. An incorrect response reduces the running ability estimate (theta) and means that the next item should be a little easier. Now imagine that an examinee runs out of time and has to start blindly guessing without even reading the test item. If the examinee is not especially lucky, the guess will be incorrect, the theta estimate will be reduced, and an easier item will be administered next. And when the blind guesses on the next few progressively easier items are also incorrect, the theta estimate can start to plunge as the scoring algorithm especially punishes wrong answers on increasingly easier items.

When the CAT version of the GRE was introduced, there were some students who ran out of time and therefore did not receive a score (only scores for complete tests were reported). In order to provide a score for these somewhat slower examinees, a rule was introduced that allowed a score to be reported if at least 80% of the test was completed. This seemed to be reasonable, as the student's estimated ability level (theta) at the 80% point was typically very close to the theta estimate for a complete test and nearly all examinees completed 80% of the questions. But there was an unfortunate unintended consequence of this rule: Well-coached, test-wise examinees realized that they could have more time per item if they intentionally aimed to answer only 80% of the items on the test. On the GRE-Analytical section (GRE-A), for example, examinees who answered all 35 items in the section had 103 seconds per item and examinees who answered 80% of the items (i.e., 28 items) had 129 seconds per item—an additional 26 seconds for every question. Because presumably well-coached students were more likely to be White than African American, and because NOT answering all questions was the optimal strategy, completion rates were higher for African American examinees. In order to discourage this benefit to well-coached students, a stiff penalty for failing to complete the test was introduced. Completion rates across racial/ethnic groups then became very similar, and the score gap between African American and White examinees was slightly reduced (Bridgeman, 1998).

Although the elimination of the 80% rule had some positive consequences, it introduced additional problems: Students resorted to guessing at the end of their CATs in order to avoid the penalty for incomplete tests. For most students, this guessing had relatively little impact on their scores, but, as suggested above, a string of incorrect guesses could have dire consequences. In some cases, if they made four or five unlucky guesses at the end of the test, scores could drop hundreds of points from what their estimated scores were five items from the end of the test. The precipitous score drop problem was greatest on the GRE-A on which there were two question types: Logical Reasoning (LR) and Analytical Reasoning (AR). The LR items were similar to items on the Verbal scale and indeed loaded on the Verbal scale in factor analyses (Wilson, 1984). The AR questions in this section provided a scenario with a set of specified relationships. For example:

A school orchestra conductor is arranging to judge six violinists—R, S, T, U, V, and W. One will play each day from Monday through Saturday. And the schedule must meet these conditions:

R must play earlier in the week than W

S must play on Thursday

V cannot play on Tuesday

Following this setup there would be a series of multiple-choice items, asking questions such as, "If R must play on the day immediately after the day on which V plays, who can play on Friday?"

The test developers assumed that for this type of AR question almost anyone could get it right given enough time, so there was a fairly strict time limit for this section. Even with the penalty for an incomplete test, about 20% of examinees failed to finish the GRE-A section. Item timing studies revealed that items in this section typically took about 2 minutes to answer, but half of the examinees had fewer than 6 minutes to answer the last six questions and one-quarter of the examinees had fewer than 2 minutes for these final questions. Although there might be an expectation that lower-ability examinees would be especially likely to run out of time, this was not the case. Bridgeman and Cline (2004) used a regression equation to estimate GRE-A scores from the Quantitative and Verbal scores on the test and found that about 24% of the students with predicted scores in the 450–540 range had fewer than 2 minutes to answer the final six questions compared to 28% of the students in the 650 or higher predicted score range. The problem for higher-ability examinees was that the CAT algorithm was giving them progressively more difficult items. Unfortunately, the way GRE-A items were made more difficult was by adding more constraints in the problem setup; balancing many constraints is both more difficult and more time consuming than balancing fewer constraints.

In a CAT, different examinees get different questions to answer, and some questions may take more time to answer than others. An important research question therefore was whether examinees who, by chance, were given tests that contained more questions that required more time to answer were at a disadvantage relative to examinees who got tests with more questions that could be answered quickly. This concern led to a study comparing students who had taken more or less time-consuming tests but who had reached an identical score level several items before finishing the test. Specifically, a comparison was made between examinees who had identical GRE-A scores (a theta estimate of 1.0 at item 29 out of 35) but differing amounts of time available to answer the last five questions on this speeded test. Final scores on the test might then be higher for the examinees who had more time for these final questions. Table 5.1 shows the results. Consistent with expectations, students who had less time to complete the GRE-A section received lower final scores than students who had more time. Of course, the luck of the draw in getting more or less time for the GRE-A could have no impact on GRE Quantitative section (GRE-Q) or GRE Verbal section (GRE-V) scores; this would be the case if both time groups were found to have virtually equal means. Though the results show that this finding is clearly seen for GRE-Q, it is *not seen* for GRE-V, indicating that more than the luck of the draw must be involved in getting more testing time. The causal arrow could go in the opposite direction so that somehow students with high verbal skills would get more time-consuming GRE-A questions. The verbally loaded LR items came first; examinees with high verbal skills would do well on these items, and the CAT algorithm would select more difficult (and more time consuming) AR items for these examinees. So long GRE-A tests could not cause higher verbal skills, but high verbal skills could cause examinees to get more time-consuming GRE-A tests. Given the problems in designing a GRE-A test that was fair for all and that still would need to have a fairly strict time limit to adequately assess the intended construct, ETS decided to drop the GRE-A and instead use an essay-based analytical writing measure (GRE-AW).

GRE-A was not the only section with fairness concerns related to speededness; some questions on GRE-Q took longer to answer than others. For example, a question on simultaneous equations took an average of about 90 seconds to answer but a question at the same difficulty level on negative exponents took only 30 seconds to answer (Bridgeman & Cline, 2000). Because with a CAT different examinees get a different mix of questions, some examinees might get a test with a disproportionate number of questions that took a longer amount of time to answer and others might get more short questions. Despite this concern, when examinees who got long GRE-Q tests were compared to examinees who got less time-consuming tests, no evidence of an impact on total scores was found (Bridgeman & Cline, 2000). As long as the test is not too speeded, it may not matter that some students get a more time-consuming test.

Research using an experimental section at the end of an operational test also suggested that time was not a major issue for GRE-V and GRE-Q scores (Bridgeman, Cline, & Hessinger, 2004). Student volunteers were randomly assigned to take a Verbal or Quantitative section with either regular time or time-and-a-half. The extended time raised scores by only 7 points for both the Verbal and Quantitative sections on the 200–800 scale, and again there were no interactions with ethnicity or gender.

The GRE was revised to become a multistage test (MST) in 2011. In an MST as compared to a CAT, branching is based on a group of items rather than on individual items. This branching strategy allows examinees to revisit items within a group and to change answers if they want to. When time limits were set for the MST, an effort was made to allow enough time not only to reach the last item but also to permit possible item revisits (Robin & Steffen, 2014). For the operational MST sections administered in 2012, the average number of item revisits was six or more across ability groups. Item timing information was used to further evaluate the extent to which the test was speeded. Rapid responding (defined as answering in less than 10 seconds) on a multiple-choice item suggests that the examinee did not have time to read the item and fully consider the answer choices. On the operational MST, the average number items with rapid responding was less than 0.5 on the Verbal section and less than 0.2 on the Quantitative section, suggesting that the test is not speeded.

The GMAT is a CAT in which examinees must answer questions in the order presented and there is no opportunity for revisiting items. Under these conditions, completion rates may be a useful speededness index. Completion rates were used to study possible differential speededness for international examinees for whom English was not their native language (Talento-Miller, Guo, & Han, 2013). The completion rate for English speakers on the Verbal scale was 95%; of the 15 foreign language groups studied, completion rates were at least 89% in all but two language groups—83% for Mandarin and a very low 47% for Korean. For the Quantitative scale, completion rates were somewhat lower; though 86% of English speakers responded to all items, in two of the language groups—French and Korean—fewer than 80% completed all items (completion rates were 79% for French speakers and 75% for Korean speakers). Mandarin speakers, who had one of the lowest completion rates for the Verbal scale, had one of the highest completion rates for the Quantitative scale: 88%. Although the Quantitative scale does appear to be very speeded for most examinees, these data cannot shed any light on whether the apparent speededness observed for certain groups is construct relevant.

Up to this point, the focus of this review has been on high-stakes admissions tests, as these typically are the tests with strict time limits and therefore are the focus of the majority of research related to speededness. Nevertheless, there has been some research on K-12 tests used in state accountability assessments, and this research is described next.

K-12 Tests

In contrast to high-stakes admissions tests, most K-12 tests have no clearly specified time limits. In Michigan, for example, the following guidance is given for test timing:

> Spring 2018 M-STEP tests are untimed and student-paced. Therefore, students must be given as much time as they need to complete each session or part of the test .... Some students will complete the test in less time than estimated, while others may require additional time. Be sure to plan for both contingencies. (Michigan Guide to State Assessments, 2017)

Similarly, tests from the Smarter Balanced Assessment Consortium are not timed. Approximate times are provided for planning purposes, but the instructions indicate, “Smarter Balanced assessments are designed as untimed tests; some students may need and should be afforded more time ...” (Smarter Balanced, 2017). Some state testing for elementary and high school students does have strict time limits. The Partnership for Assessment of Readiness for College and Careers (PARCC) Test Coordinator Manual notes, “PARCC tests are strictly timed, and no additional time may be permitted (with the exception of an extended time accommodation ....” (PARCC, 2018, p. 9)

Directions for the Stanford Achievement Tests (grades 1–10) were similar to the Smarter Balanced instructions with the *Directions for Administering* indicating, “Times are included for planning purposes only. Stanford 10 is to be administered so that all children have sufficient time to complete it ... if necessary, additional time must be provided for a student to complete the tests” (as quoted in Brooks, Case, & Young, 2004). These authors conducted a study designed to determine whether performance would differ for students tested with the "suggested" time strictly enforced compared to students tested under the standard untimed conditions. There were 360,000 students in the untimed standardization sample, and about 150 classrooms were selected at random to take the test with the suggested times actually enforced. Demographic characteristics were similar across both groups. In all subject areas and grades, effects (average differences between timed and untimed groups) were less than one raw score point. Although all differences were small, the direction of the differences changed across grade levels. In grades 1–6, the scores were higher for the untimed group; in grades 7–10, scores were higher in the timed group. The authors did not speculate on the reasons for this reversal, but it could be related to the extra anxiety timing might create for older students. Research is needed to confirm the speculation that timing anxiety is greater in older students. Although extra anxiety can have negative effects on a high-stakes test, on a no-stakes test it can be positive as it simply enhances motivation, reflecting the well-known "inverted U" function in which too little or too much arousal (motivation) can be detrimental to performance. Given the legitimate concerns with speededness in both admissions tests and accountability assessments, additional research is needed in this area.

Practical Considerations for Designing and Conducting Timing Research Studies

As in other areas of educational and psychological research, random assignment studies should be the gold standard for studies on the effects of time limits. Examinees can either be randomly assigned to tests of equal lengths with different time limits or to tests of different lengths with a fixed time limit. An advantage of the latter approach is that it can be less obvious to the examinees that they are part of a research study, and motivational factors related to knowing that a test is merely for a study with no individual consequences could be especially problematic in studying effects of time limits. Although there may be a temptation to estimate speed effects without an experimental manipulation, such studies often require assumptions that are known to be untrue. As we have seen, estimating speededness by evaluating completion rates requires the assumption that examinees consider each item in order with no skipping and returning to items if time permits. And with nearly all tests now just counting the number of right answers, all items should be answered even if random guessing is required because of lack of sufficient time to fully consider each item. Computer timing of individual items can be helpful in identifying cases of extreme speededness in which it can be obvious that an examinee answered without enough time to seriously consider the item, but when an examinee exceeds this minimum threshold it is impossible to know if a correct answer would be more likely with more time. If every examinee answers every item and does not show evidence of very rapid responding on any item, it is still quite possible that scores would be higher if examinees had more time to consider each item. The ideal random assignment study may be impossible because of sample size and/or cost considerations. In such cases, an imperfect study is preferable to no study. Observational studies can be of some value in estimating the approximate amount of time needed to answer questions of different types. When new question types were being considered for the next version of the SAT, a study was conducted in which students were observed as they tried to answer these new question types (Bridgeman, Laitusis, & Cline, 2007). The time taken to answer each question then could be used to estimate the amount of time that should be allowed for each section of the revised test, and the new version of the test had somewhat more generous time limits. Another type of useful observational study for computer-based tests is to use the computer to record time to respond to individual items. Although the lack of rapid responding on every item cannot assure the lack of speed effects, it is nevertheless useful to get item response times, if possible, because the presence of rapid responding can still provide solid evidence that a test is speeded.

Conclusions

Test publishers have a responsibility to evaluate the effects of time limits on their tests, and test users have a responsibility to demand evidence on the speededness of tests that they use. The nature of the required evidence has shifted over time, as the problems with some of the early indices have become better known. Speededness studies and guidelines from the early days of standardized testing assumed that completion rates were a reasonable way of assessing test speededness. At the time this may have been a reasonable assumption, but with growing sophistication in test-taking strategies and advice to skip time-consuming items and return to them if time permits, completion rates have been rendered essentially useless as a measure of test speededness. Experimental manipulation of timing is a viable alternative, either by directly manipulating time limits for a given number of items or by adjusting the number of items within a given time limit. Computer delivery allows for nonexperimental approaches that can evaluate rapid responding rates, but rapid responding can have different interpretations; it can indicate serious attempts to answer all items with a time limit that is inadequate, or, particularly with low-stakes tests, it simply may indicate low motivation.

For any new or revised test in which speed is not intended to be a part of the construct being assessed, it is essential that test publishers determine the extent to which test timing affects test scores. At the early stages of test development, small-scale observational studies can be useful in determining reasonable time expectations for each item; when final forms are assembled, experimental manipulations of time limits may be needed. For smaller testing programs, these studies should not require the huge samples that are feasible with large-scale national testing programs.

Different testing models can be more or less sensitive to time limits. Although running out of time can have negative consequences with any testing format, CATs with item-level branching can be especially vulnerable as estimated ability declines with every answer that is wrong because the examinee ran out of time. Effects of time limits need to be assessed for both multiple-choice and constructed-response formats. For essay tests, extra time generally may be beneficial, but this is not always the case. Extra time on an essay may just give weaker examinees a greater opportunity to demonstrate their shortcomings.

The most reasonable generalization to characterize the effects of time limits on test scores is that generalizations are not possible. Extra time under some circumstances has large effects; in other cases it has trivial effects and can even have a negative impact on scores. Effects may be stronger for lower-ability examinees or for higher-ability examinees. One generalization that certainly is true is the familiar mantra that more research is needed. This is especially critical when new tests are introduced or major revisions are made to existing tests.

References

ACT. (2019). *Technical Manual*. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf

Bridgeman, B. (1998). Fairness in computer-based testing: What we know and what we need to know. In Niogi Shilpi (Ed.), *New direction in assessment for higher education: Fairness, access, multiculturalism, & equity* (The GRE, FAME Report Series, Vol. 2, pp. 4–11). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Cline, F. (2000). *Variations in mean response times for questions on the computer-adaptive GRE General Test: Implications for fair assessment* (GRE Report No. 96-20P; ETS RR-00-7). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2000.tb01830.x

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement*, *41*, 137–148. doi: 10.1111/j.1745-3984.2004.tb01111.x/abstract

Bridgeman, B., Cline, F., & Hessinger, J. (2004). Effect of extra time on verbal and quantitative GRE scores. *Applied Measurement in Education*, *17*, 25–37. doi: 10.1207/s15324818ame1701_2

Bridgeman, B., Laitusis, C. C., & Cline, F. (2007). *Time requirements for the different item types proposed for use in the revised SAT* (ETS RR-07-35; College Board Research Report No. 2007-3). New York, NY: College Entrance Examination Board. doi: 10.1002/j.2333-8504.2007.tb02077.x

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41*(4), 291–310.

Brooks, T. E., Case, B. J., & Young, M. J. (2004). *Timed versus untimed testing conditions and student performance.* (Pearson Assessment Report). Retrieved from https://images.pearsonassessments.com/images/tmrs/tmrs_rg/TimedUntimed.pdf?WT.mc_id=TMRS_Timed_Versus_Untimed_Testing

College Board (2018a). *Test specifications redesigned SAT*. Retrieved from https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf

College Board (2018b). *Official study guide.* Retrieved from (https://collegereadiness.collegeboard.org/pdf/official-sat-study-guide-read-keys-doing-your-best.pdf)

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*, 123–131. doi: 10.1111/j.1745-3984.1972.tb00767.x/abstract

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrica*, *15*, 259–269.

Harik, P., Feinberg, R. A., & Clauser, B. E. (2020). How examinees use time: Examples from a medical licensing examination. In M. J. Margolis & R. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp.73–89). Abingdon: Routledge.

Li, D., Yi, Q., & Harris, D. (2016). *Evidence for paper and online ACT comparability* (ACT Working Paper 2016-02). Iowa City: ACT.

Lord, F. M. (1950). *A method of estimating from speeded test data the power condition scores and item difficulties* (ETS-RB-50-24). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1950.tb00211.x

Lord, F. M. (1954). *A study of speed factors in tests and academic grades* (ETS-RB-54-24). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1954.tb00251.x

Lovett, B. J. (2020). Extended time testing accommodations for students with disabilities: Impact on score meaning and construct representation. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 47–58). Abingdon: Routledge.

Margolis, M. J., von Davier, M., & Clauser, B. E. (2020). Timing considerations in performance assessments. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 90–103). Abingdon: Routledge.

Michigan Guide to State Assessments. (2017). Guide to state assessments. Retrieved from http://www.michigan.gov/documents/mde/2017-2018_Guide_to_State_Assessments_ada_2_603949_7.pdf

Mollenkopf, W. G. (1949). *Item-analysis data from an experimental study of the effects on item-analysis data of changing item placement and test time limit* (ETS-RB-49-10). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1949.tb00914.x

PARCC. (2018). *PARCC test coordinator manual.* Retrieved from https://parcc.pearson.com/manuals/

Robin, F., & Steffen, M. (2014). Test design for the GRE revised general test. In C. Wendler & B. Bridgeman (Eds.), *The research foundation for the GRE Revised General Test: A compendium of studies.* Princeton, NJ: Educational Testing Service.

Searcy, C. A., Dowd, K. W., Hughes, M. G., Baldwin, S., & Pigg, T. (2015). Association of MCAT scores obtained with standard vs. extra administration time with medical school admission, medical student performance, and time to graduation. *JAMA*, *313*, 2253–2262. doi:10.1001/jama.2015.5511

Shao, C., Li, J., & Cheng, Y. (2015). Detection of test speededness using change-point analysis. *Psychometrika*, *81*, 1118–1141. doi: 10.1007/s11336-015-9476-7

Smarter Balanced. (2017). *Estimated testing times*. Retrieved from https://portal.smarterbalanced.org/library/en/estimated-testing-times.pdf

Swineford, F. (1949). *Law School Admissions Test-WLS* (ETS-RB-49-12). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1949.tb00915.x

Swineford, F. (1956). *Technical manual for users of test analyses* (ETS-SR-56-42). Princeton, NJ: Educational Testing Service.

Talento-Miller, E., Guo, F., & Han, K. T. (2013). Examining test speededness by native language. *International Journal of Testing*, *13*, 89–104. doi:10.1080/15305058.2011.653021

Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement*, *19*, 19–28.

Wilson, K. M. (1984). *The relationship of GRE General Test item-type part scores to undergraduate grades* (ETS RR-84-38). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2330-8516.1984.tb00078.x.

6

How Examinees Use Time

Examples from a Medical Licensing Examination

Polina Harik, Richard A. Feinberg, and Brian E. Clauser

The widespread use of computers in test delivery has made it possible to collect extensive data about how examinees use available time as they complete a test. This type of timing data can provide important evidence related to the validity of the inferences we make based on test scores. For example, timing data can provide insight into the level of examinee engagement as they complete a test (see Chapter 11), can help to identify examinees who have had prior access to test material (see Chapter 12), and can provide insight into the cognitive processes examinees use in responding to test items (see Chapter 9). These are all important considerations that provide useful information about critical aspects of the testing experience. The present chapter addresses a different aspect of the use of timing data: it provides a framework for understanding how an examinee's use of time interfaces with time limits to impact both test performance and the validity of inferences made based on test scores. The content of this chapter focuses primarily on examinations that are administered as part of the physician licensure process. The reason for this focus is twofold. First, examinees completing these high-stakes examinations tend to be highly motivated; changes in patterns of responding across the test that result from time constraints therefore are not likely to be confounded with changes in the examinees' level of engagement. Second, there is relatively little published research on this topic from other assessment contexts. When such research was available, it has been included to the greatest extent possible.

We begin by examining the extent to which timing data can provide an indication that examinee scores are impacted by test time limits. Drawing inferences about speededness based on observational evidence has been a matter of interest for generations, and the data that have become available from computer-based test delivery have created new possibilities for further exploring this area. After considering different approaches to assessing speededness based on observational timing data, we examine three additional areas. The first focuses on how examinees allocate time for review and how this review impacts their scores. We then consider how the response time for an item relates to the probability of answering the item correctly. This is followed by an examination of the characteristics of test items that impact response time. We conclude with practical recommendations and consideration of how these results and our recommendations might generalize to assessment contexts outside of physician licensure.

Inferences about Test Speededness Based on Observational Data

It is clear that time limits that significantly impact test scores represent a threat to valid score interpretation in many assessment contexts. As a practical matter, however, it is necessary to balance this threat against the costs associated with more generous time limits. Time that is used to test students is time that is not available for other instructional activities. The cost of testing time is even more apparent for tests delivered in commercial test centers, because the cost of the scheduled seat time contributes directly to the cost of the test; more generous time limits result in higher costs. Balancing these priorities is challenging and clearly requires evidence about the extent to which time limits impact scores.

As Chapter 5 makes clear, definitive information about how time limits impact examinee scores requires data from structured experiments. These experiments typically involve manipulating the number of test items delivered in a fixed amount of time or manipulating the amount of time allowed to complete a fixed number of items. Such experiments are the gold standard for inferring causal relationships between time limits and performance. Unfortunately, as Cronbach and Warrington (1951) pointed out more than half a century ago, such experiments tend to be expensive and difficult to implement. These practical limitations of experimental methods have led researchers and test administrators to look for observational approaches that allow for making inferences based on data from operational administrations of the examination. One of the simplest of these approaches is often referred to as Swineford's rule of thumb or simply the Swineford rule. This rule states that essentially all examinees should complete at least the first 75% of the items and at least 80% of the examinees should respond to the final item for a test to be considered non-speeded (Swineford, 1956, cited in Rindler, 1979). Similarly focusing on items that are not reached, Gulliksen (1950) proposed an index based on the ratio of the standard deviations of number of items answered incorrectly to the total number of items answered incorrectly and items not reached. Stafford (1971) provided a simpler index based on the ratio of the sum of items answered incorrectly to the total of items answered incorrectly and items not reached.

In the context of multiple-choice tests (particularly in the absence of a penalty for guessing), these simple observational approaches are likely to be limited because as time runs out examinees may select answers at random hoping that by chance at least a few of the responses will be correct. The first significant step forward in interpretation of examinee behavior based on timing data collected with computer administration was provided by Schnipke (1995; Schnipke & Scrams, 1997). She suggested that both (1) items that were not reached, and (2) items that were responded to with rapid guesses should be considered in evaluating the impact of time limits. The logic of this approach is straightforward. As Schnipke and Scrams state, "We assume that examinees choose to engage in either "solution behavior" or "rapid-guessing behavior" on each item. Further, we assume that examinees can switch strategies at any point and that they do so in response to time constraints on the test" (p. 214). In this context, rapid guessing is defined empirically by identifying the amount of time used by examinees who are responding at chance level (Kong, Wise, & Bhola, 2007; Harik et al., 2018).

This two-state solution is an important step forward in understanding how examinees use time, but it falls far short of fully describing the complex ways that examinees respond to time pressure. The possibility that examinees employ other strategies in response to time constraints is ignored by limiting the model to rapid guessing and solution behavior. For example, examinees may substantially speed up near the end of a timed section; this could impact performance on those items but not reach the response rate required to categorize the responses as rapid guesses. Similarly, examinees may maintain a speeded pace in responding to all items on the test, uniformly lowering the probability of a correct response across the item set (Harik et al., 2018). This speed-accuracy tradeoff that occurs when examinees attempt solution behavior under time constraints is well known (e.g., Heitz, 2014).[1] The fact that examinee performance will deteriorate as the time constraint becomes more extreme is almost unavoidable. The practical question then is: *To what extent does examinee behavior on a single test administration with a fixed time limit provide evidence about the impact of the time limit on performance?* Or put another way, *can we infer whether or not a test is speeded by examining response time data?*

The obvious answer to the latter question is that we *can* make such inferences, at least at the extremes. If all examinees complete the test well before the time limit, it is likely safe to infer that more time will not result in higher scores. Similarly, if many examinees either fail to reach numerous items near the end of the test or make rapid guesses on these items, it is likely that the test is speeded. In other circumstances, we will be less certain about our conclusions. In the next sections, we review several types of evidence that can be collected to make such inferences and evaluate the usefulness of inferences made based on each type of evidence.

Metrics Based on Rapid Guessing

Schnipke's suggestion that rapid guesses should augment measures of items that are not reached in evaluating speededness represented an important step forward in evaluating examinee behavior. In many settings, examinees are likely to have learned that as time runs out they should respond to all items, even if there is not enough time to read the questions. As noted, however, there is a potential problem with this measure; rapidly guessing the answers to the final items in a timed section is only one strategy that examinees may use to adjust their test-taking behavior to time constraints. Another strategy is for examinees to pace themselves throughout the timed section so that they use relatively similar amounts of time across the item sequence. This approach—or a variation of this approach—seems likely in the context of high-stakes standardized tests for which practice materials are available. When examinees take this approach, we are likely to say that they paced themselves well. Whether this strategy optimizes the examinee's score is a separate question, but it is clear that if examinees are able to pace themselves "well," they will reduce or eliminate instances of rapid guessing that might otherwise provide evidence that time constraints are impacting scores.

The extent to which measures of rapid guessing provide useful evidence about the impact of time constraints on test scores is an empirical question, and there is relatively little evidence reported in the literature that provides insight into the answer. The work by Schnipke (1995) and Schnipke and Scrams (1997) makes it clear that the phenomenon does exist. Numerous studies by Wise (e.g., Wise, 2017; Chapter 11, this volume) also report on rapid guessing, but these studies were carried out to examine engagement on tests with generous time limits and so are not directly relevant to the question of time constraints.

More recently, Harik et al. (2018) examined a measure of rapid guessing in a randomized experiment conducted in the context of the Step 2 Clinical Knowledge component of the United States Medical Licensing Examination (USMLE®; a test that physicians with an M.D. degree must pass in order to be licensed to practice medicine in the United States). At the time of the experiment, the test comprised 8 hour-long timed sections with approximately 44 items in each section. For the experiment, the last of the hour-long sections was manipulated to have 28, 32, 36, 40, or 44 items. The sections were constructed so that each set of 28 items was a subset of the set of 32 items; each set of 32 items was a subset of the set of 36 items, etc. Within each section, the item sequence was randomized for each examinee. One of these experimental sections was randomly assigned to each examinee in the study. The results generally showed a modest improvement in performance as the number of items per section decreased from 44 to 32; reducing the number of items from 32 to 28 provided no additional score increase.

To examine the extent to which measures of rapid guessing and not-reached items are sensitive to the impact of time constraints, graphs were produced showing these measures as a function of the item presentation sequence. The results showed extremely low levels of rapid guessing and not-reached items across the item sequence for the condition with 28 items per section. These rates increased noticeably for the condition with 36 items and substantially for the most time-intensive condition (44 items per section). Across all conditions, these measures remained low for items early in the sequence and increased for the final items. Harik et al. (2018) also report evidence that makes it clear that many examinees were pacing themselves in a way that reduced their performance but did not show up in these measures. They report that the reduction in performance for the final position in the item sequence relative to performance earlier in the sequence could not be accounted for by items not reached and rapid guesses. Specifically, they note that the mean performance for the first 32 (of 44) items in the sequence (before an increase in rapid guesses and not-reached items begins) was 0.71. Accounting for examinees who did not reach the final item position and were scored as incorrect, as well as examinees with a chance probability of success due to rapid guessing (approximately 0.2), the expected proportion correct would have been 0.67. The actual value was 0.64.

These results suggest that the presence of rapid guessing and not-reached items may provide positive evidence that time constraints are impacting test scores. However, concluding that the absence of these behaviors demonstrates that the scores are *not* impacted by time constraints may not be justified. It is clear that some examinees whose scores are impacted by time constraints pace themselves so that they avoid the need to guess at random or leave items unanswered. Again, in situations where practice materials are available, examinees may spend considerable practice time learning to pace themselves; such pacing may optimize their performance under the specific time constraint, but it cannot make an insufficient amount of time sufficient.

Metrics Based on Changes in Response Time across the Item Sequence

The previous section provides evidence that an approach that treats examinees as either engaging in (undifferentiated) solution behavior or rapid guessing might oversimplify actual examinee behavior. This raises the question of whether it might be more useful to employ a measure that recognizes that the speed-accuracy tradeoff is continuous rather than discrete. One obvious approach to addressing this question would be to examine the amount of time used on items as a function of item sequence. If examinees are running out of time, they may need to increase their pace as they approach the end of a timed section. Again, if examinees pace themselves well, they will hide this effect, but the approach has the potential to identify examinees who increased their pace in a manner that materially impacts their scores without reaching the criterion for rapid guessing.

Harik et al. (2018) also provided experimental evidence related to this question. Using the data set described in the previous section, they examined whether the extent of the drop in seconds per item near the end of a timed section was predictive of the level of impact associated with the time limit for the specific condition. Although they did not provide a quantitative measure of the change in slope, they made two general statements: (1) an apparent drop in seconds per item (slope) was present for the three more time-intensive conditions in which time constraints were shown to impact scores, and (2) the magnitude of the change did not appear to be proportional to the magnitude of the impact on scores.

To follow up on the Harik et al. (2018) study, we examined an additional experimental data set from Step 2 of the USMLE. The data set used in the original study was collected in 2012; this additional data set was collected in 2015 (for simplicity, we will refer to these data sets as the 2012 study and the 2015 study). The basic structure of this experiment was the same as that already described. Examinees were randomly assigned to hour-long experimental test sections containing different numbers of items. The order of the items within the section was randomized for each examinee and the experimental sections were administered as part of

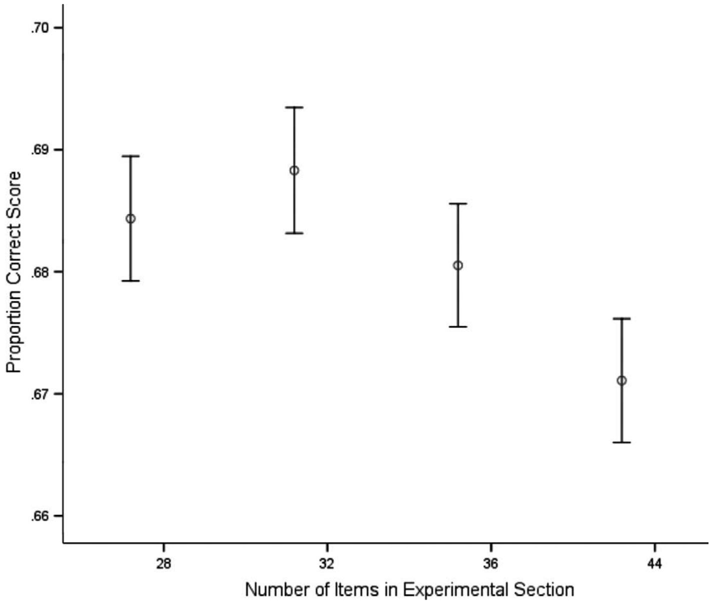


Figure 6.1 Average examinee performance on common items by experimental condition with 95% confidence intervals.

an operational administration so that examinees were performing under high-stakes conditions. The structure of this experiment was identical to that described by Harik et al. (2018), except that the more recent study did not include the 40-item timing condition: examinees saw 28, 32, 36, or 44 items in each hour-long experimental block.

Figure 6.1 presents results showing how performance was impacted by the number of items per block. The findings are similar to those presented by Harik et al. (2018), indicating that performance was similar with 28 or 32 items per block (in fact, in both studies, examinee performance for the 32-items-per-block condition was slightly higher than that for the condition with 28 items per block). Performance then declines when 36 or 44 items are administered in each block. Figure 6.2 shows response time as a function of presentation sequence for each of the experimental conditions. For all conditions, there appears to be a slight downward trend across the presentation sequence, with examinees responding slightly more quickly to items at the end of the sequence than they did at the beginning. Perhaps more noteworthy is the finding that for the 36- and 44-item conditions the slope changes for the last items in the sequence with the time per item falling off more rapidly near the end. Wainer (1971) provided a simplified procedure for evaluating whether such a change in slope is statistically significant. The results of applying Wainer's procedure to the data presented in Figure 6.2 show that there is a significant change in slope for the 36- and 44-item conditions but not for the 28- or 32-item conditions. A review of the results presented by Harik et al. (2018) reveals the same pattern: there is a change in slope for the conditions shown to be impacted by time constraints. Based on these results, it is tempting to speculate that this change in slope might act as an indicator that examinee performance is impacted by the time limit. More work in this area is warranted.

Metrics Based on the Proportion of Examinees Using all Available Time

We have already noted that if essentially all examinees complete the test well before reaching the time limit, we might conclude that a more generous time limit would have little impact

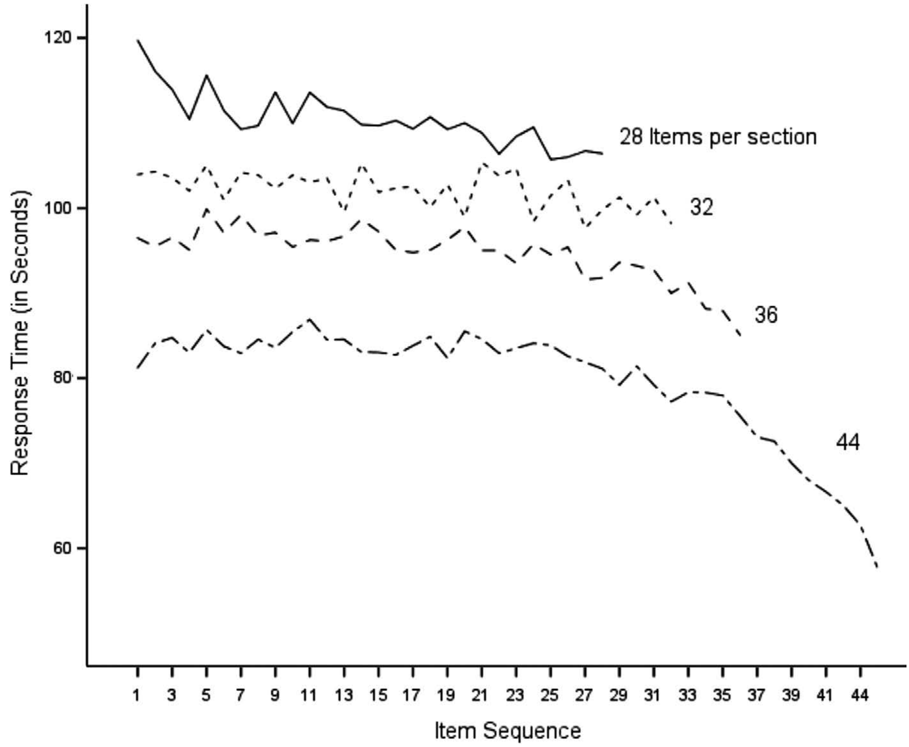


Figure 6.2 Item response time by item presentation sequence for the five experimental conditions.

on performance. This conclusion raises the question of whether the proportion of examinees using essentially all of the allotted time to complete the test is a marker that the test is at least somewhat speeded. This would be a reasonable conclusion if we are willing to make two assumptions: (1) there is some meaningful variability in the rate at which examinees are able to complete the activities required to respond to test items (e.g., reading, calculation, problem solving), and (2) examinees do not use more time than is needed to achieve the highest score they are capable of earning. Under these circumstances, if most examinees used all of the available time, we could conclude that (at least) some had less time than was needed for optimal performance.

There are two papers that provide empirical evidence related to the credibility of the second of these assumptions. Wise (2015) reports results from a large-scale K-12 achievement testing program. In that context, examinees reduced the time they spent across the last 40 items (on a 50-item test) from an average of 52 seconds per item to 33 seconds per item. This change was not associated with any noticeable change in the probability of a correct response. Although there are other possible explanations, this strongly suggests that examinees were working at an unnecessarily leisurely pace earlier in the test. Wise commented in the same paper that he had observed a similar pattern of performance in a medical licensing examination.

The Harik et al. (2018) paper cited in previous sections also provided evidence relevant to this question. Again, the paper described an experimental study in which examinees were randomly assigned to conditions in which they were presented with 28, 32, 36, 40, or 44 items in an hour-long timed section. In general, the results clearly showed an improvement in performance as the number of items per section decreased; reducing the number of items from 32 to 28, however, provided no advantage (in fact, performance dropped modestly for the 28-item condition).

Nonetheless, examinees used more time when they were given more time, and this increase in the use of time per item continued even when the number of items was dropped from 32 to 28. The results provide compelling evidence that at least in some circumstances examinees will use more time than is necessary to optimize their scores. The results presented in Figures 6.1 and 6.2 replicate this pattern of behavior with examinees again using more time per item in the 28-item condition than the 32-item condition even though it resulted in no score increase.

Taken together, the results reported by Wise (2015) and those based on the USMLE studies suggest that the amount of time examinees use may be a questionable indicator that test scores are impacted by time constraints. In the case of the USMLE studies, the examination has very high stakes for examinees and caution in checking and rechecking answers is understandable. The results reported by Wise are more difficult to explain because they occurred in what is likely to be viewed as a relatively low-stakes setting where that level of caution may not be expected.

Examinee Behavior and Item Characteristics

The approaches discussed to this point all focus on indicators that examinees are running out of time. We now take a more focused look at what timing data can tell us about how examinees use testing time. We begin by examining time spent on item review. In this context, we consider both the prevalence of review and the impact of review on scores. We then examine the relationship between response time and the probability of answering correctly. Finally, we consider the characteristics of items that relate to response time.

Time Spent on Item Review

We have already discussed the fact that in high-stakes testing situations, examinees want to be sure that they have not answered incorrectly because of a careless oversight. This concern with careful review may explain why examinees appear to use more time than is required to achieve their maximum score. To better understand this behavior, we begin by considering the prevalence of item review.

Kahraman, Cuddy, and Clauser (2014) examined item review behavior in the context of the USMLE. They reported that a substantial majority of examinees reviewed at least some of the items. A more recent analysis of item review based on data from the 2012 timing study showed that item review increased as the number of items in the hour-long section decreased, but even for the most time-intensive condition (44 items per section), 5% of examinees revisited all of the items. For the 28-item-per-section condition, more than 27% of examinees revisited all 28 items at least once. By contrast, the number of examinees who did not revisit a single item was generally around 2% across conditions.

With these relatively high rates of item review, we might predict that the rate at which scores increase would drop near the end of a timed section—with many examinees reviewing previously answered items and making relatively few answer changes. If this is true, the amount of time examinees spend in a section after achieving their highest score could tell us something about the prevalence of this sort of careful checking. To examine this phenomenon, we returned to data from the 2012 study. The results are presented in Figures 6.3 through 6.6. Figure 6.3 shows the cumulative proportion of examinees achieving their maximum score and the proportion of examinees completing the section, both as a function of testing time. Figure 6.4 shows the distribution of the amount of time used after examinees achieved their maximum score but before completing the section. Figures 6.3 and 6.4 present results for the 28-items-per-section condition (in which most examinees had more time than they needed for

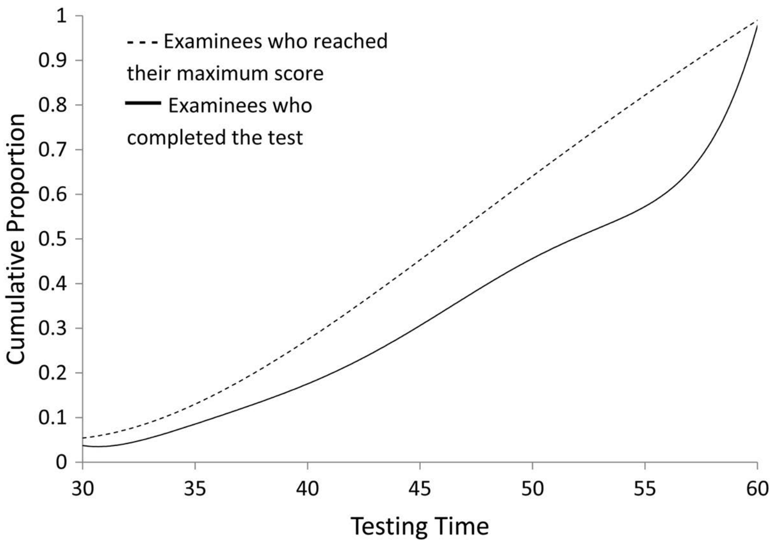


Figure 6.3 Cumulative proportion of examinees at different testing times for the 28-items-per-block condition.

optimal performance). Figures 6.5 and 6.6 provide analogous results for the 44-items-per-section condition.

Figures 6.3 and 6.5 present information about the absolute amount of time examinees use to achieve their highest score on the section and to complete the section. Figures 6.4 and 6.6 show how much time examinees spend reviewing items after achieving their highest score.

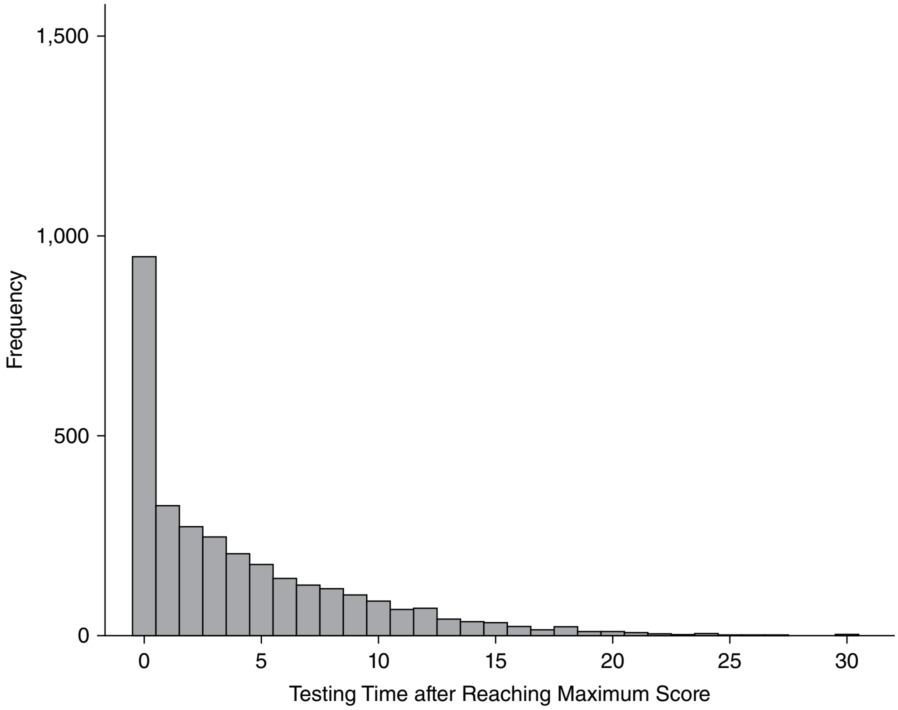


Figure 6.4 Distribution of testing time used by examinees after reaching their maximum score.

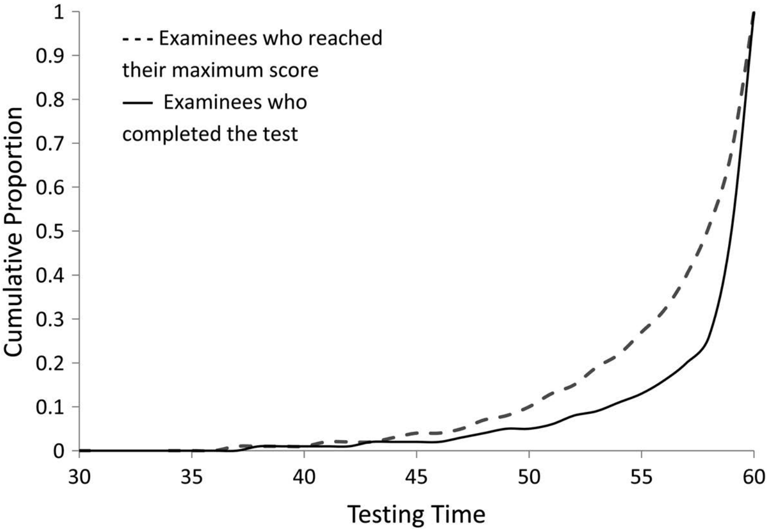


Figure 6.5 Cumulative proportion of examinees at different testing times for the 44-items-per-block condition.

(That is, these figures show how much time examinees spent on item review that did not result in any score improvement.) Taken together, these figures show that more than 50% of examinees spend 5 minutes or more on such item review. This percentage is greater in the less time intensive 28-items-per-section condition. Moving from 28 to 44 items per section increases the percentage of examinees using 2 minutes or fewer for review (after achieving their

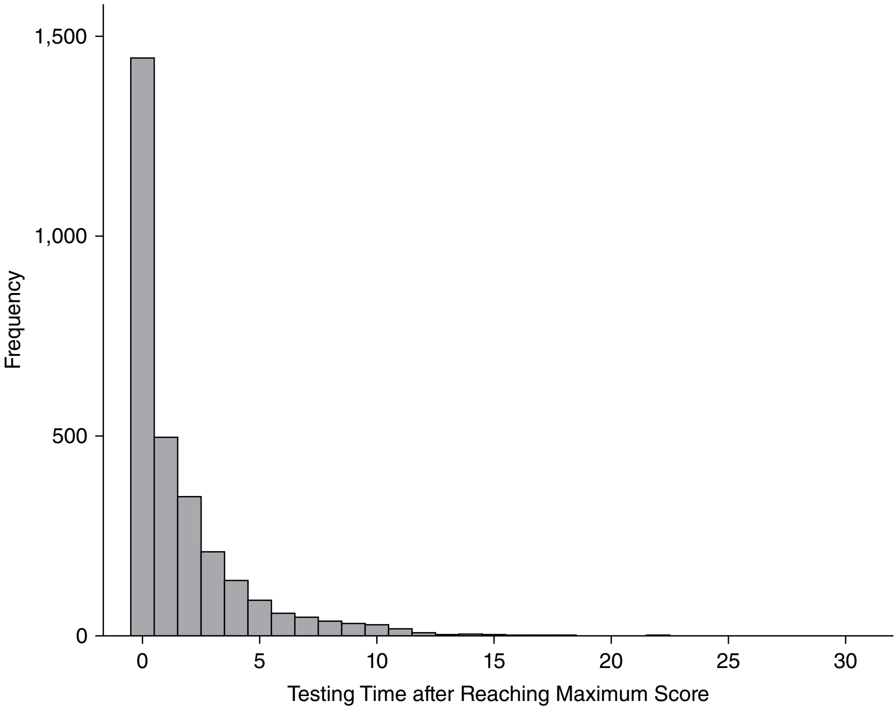


Figure 6.6 Distribution of testing time used by examinees after reaching their maximum score.

highest score on the section) and decreases the percentage using 3 minutes or more. These results clearly suggest that, when given sufficient time, examinees will use significant amounts of time to double-check their answers.

The results presented in the previous paragraph make it clear that a nontrivial amount of time is spent reviewing items without an impact on scores. This raises a more general question about the overall impact of item review; clearly item review must occasionally identify errors and result in score improvement. In a study investigating answer changes resulting from item review, Ouyang, Harik, Clauser, and Paniagua (2019) reported that 99% of examinees reviewed at least one item and 68% changed at least one answer. Approximately 45% of examinees increased their scores and 28% decreased their scores by changing answers. On average, examinees reviewed 16 items (in a 44-item section) but made changes on only 1.4 items. The average score change was positive but extremely small.

The Relationship between Response Time and the Probability of a Correct Response

The next question of interest is *how does the relative difficulty of an item relate to the amount of time examinees allocate to that item?* This relationship has been examined in two previous studies using data from the USMLE. The answer from both studies might be summarized as *it takes more time to be wrong*. Swanson, Case, Ripkey, Clauser, and Holtman (2001) used a hierarchical model to examine the relationship between item characteristics and response time and showed that examinees generally spent more time responding to more difficult items (for which the probability of a correct response is lower). Beyond this simple relationship, they reported the presence of an interaction with examinee proficiency: lower proficiency examinees spent more time on easier items and higher proficiency examinees spent relatively more time on more difficult items.

To further explore the interaction effect reported by Swanson et al. (2001), we examined response time as a function of both item difficulty and examinee proficiency. Figures 6.7 and 6.8

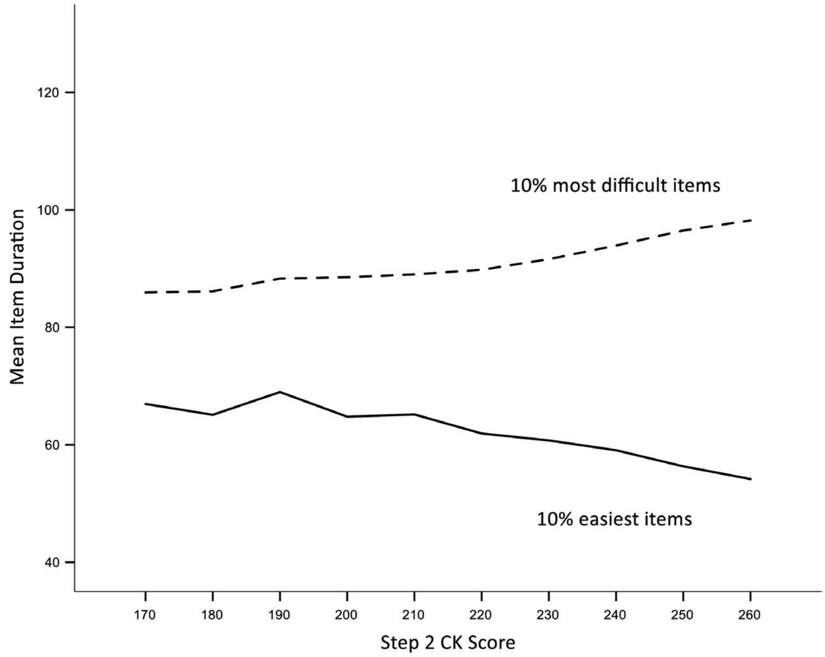


Figure 6.7 Item duration for examinees of different ability for the 44-items-per-block condition.

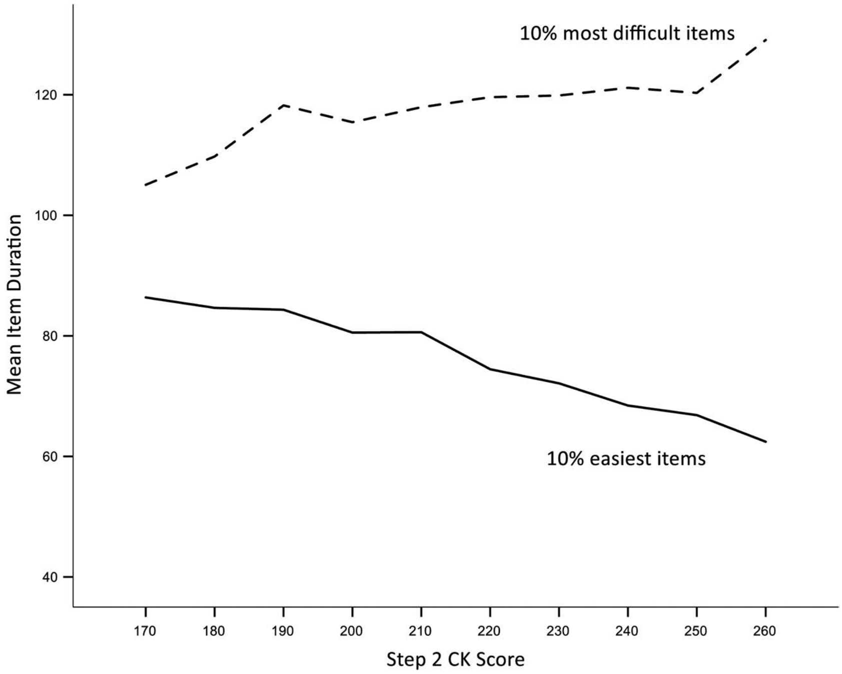


Figure 6.8 Item duration for examinees of different ability for the 32-items-per-block condition.

show the amount of time used on the easiest and most difficult 10% of items as a function of examinee proficiency (measured by the reported total test score). The data were collected as part of the 2015 study; Figure 6.7 presents results for the 44-items-per-section condition and Figure 6.8 presents results for the 32-item condition. Although the results are more extreme for the 32-item condition, both graphs show that for the easier items there is a decrease in seconds per item as proficiency increases. For the more difficult items, there appears to be a modest increase in time spent per item as a function of proficiency. The results represented in these figures are consistent with those reported by Swanson et al. (2001), showing that although all proficiency groups tend to spend more time on difficult items than on easy items, the extent of this disparity is greater for more proficient examinees who use disproportionately less time on easy items.

More recently, Feinberg and Jurich (2018) also reported results related to the amount of time examinees used on an item and the corresponding probability of a correct response, but rather than looking at time use as a function of proficiency they looked at the probability of a correct response as a function of time use. They reported that, as expected, very short response times were associated with chance levels of performance. As examinees used more time on an item, the probability of a correct response rapidly increased and then more slowly declined. The time associated with maximum performance was well below the average time spent on an item and well below the time available. Figure 6.9 provides an example of this relationship, again based on an annual cohort of examinees completing the USMLE Step 2 examination. The results do not provide evidence about the direction of causality (or even demonstrate causality), but it seems likely that when examinees are immediately sure that they have identified the correct answer to an item they are both more likely to be correct and less likely to spend additional time considering alternative options or returning to the item for review.

In spite of the seemingly clear message from Figure 6.9, it is important to keep these results in perspective. It is evident that examinees who respond to items quickly (but not too quickly)

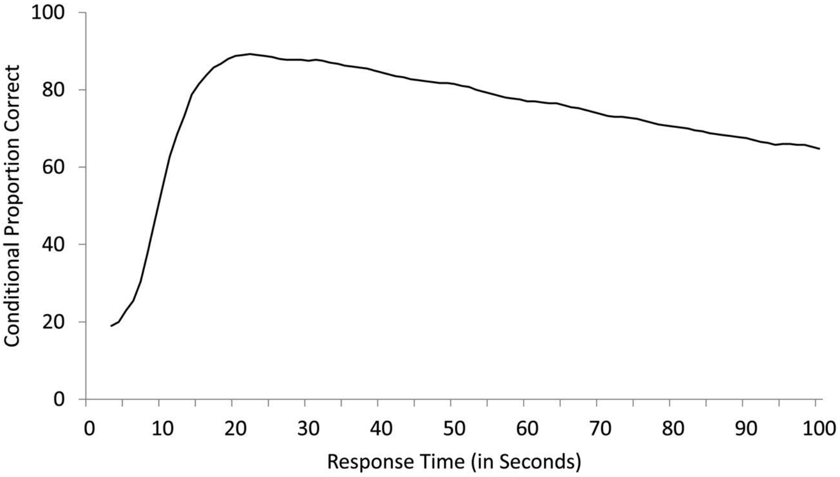


Figure 6.9 Conditional proportion correct by response time.

have a high probability of responding correctly. The temptation in viewing these results is to reverse the directionality and to infer that more proficient examinees uniformly respond more quickly. Figure 6.10 (based on the same data set used for Figure 6.9) shows the distribution of response rates for the highest, middle, and lowest quintiles of examinees based on the total test score. The results do confirm that both the mean number of seconds per item and the modal response time are in fact modestly lower for more proficient groups; however, the within-quintile variability is much greater than the between-quintile variability.

Characteristics of Items that Relate to Time Intensity

Clearly, item difficulty relates to how much time an examinee will spend responding to an item. There also are more superficial characteristics of items that relate to response time. In the previously referenced study by Swanson et al. (2001), the authors reported on surface characteristics

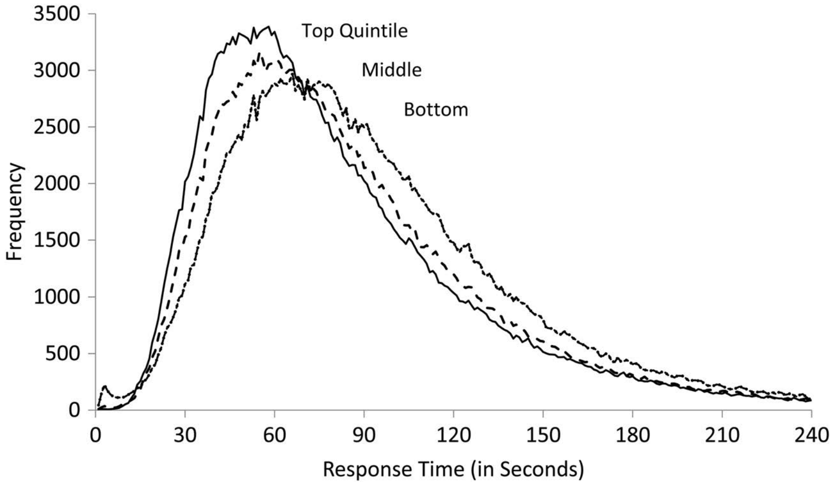


Figure 6.10 Density of response times by examinee proficiency quintile.

of items that predicted time intensity in addition to examining time usage as a function of examinee proficiency and item difficulty. They reported both a linear and quadratic relationship between word count and response time, with each additional word adding approximately half a second to the time required to respond. They also reported that inclusion of a picture as part of the stimulus material added approximately 12 seconds to the response time.

Swanson, Holtzman, Albee, and Clauser (2006) presented results relating the number of options presented in the item to testing time. The results showed a clear increase in response time associated with an increased number of options; there was no explicit control for word count, however, and the authors concluded that much of the increase in response time was associated with the increase in word count.

More recently, Ha, Marsic, and Yaneva (2017) examined the relationship between linguistic features of USMLE items and mean response times. They found that the strongest predictors were counts of nouns (and noun phrases) and counts of rare words (those that are not in the most commonly used 2,000 or 3,000 words). They produced predictive models that had a correlation of approximately 0.60 with actual response times.

Conclusions

In this section, we attempt to draw conclusions and provide recommendations for practice based on the timing research reported in this chapter. We begin with recommendations on the use of observational data as a basis for making inferences about whether time limits are impacting test scores. We then consider the implications of research on (1) item review, and (2) the relationship between response time and proficiency. This is followed by a discussion of how timing information can improve test construction. Finally, we consider the extent to which the results reported in this chapter are likely to generalize to contexts beyond medical licensing.

Using Timing Data to Identify Speededness

As stated previously, it is clear that at the extremes, timing data can provide answers about whether or not time limits are impacting test scores. If essentially all examinees complete the test well before time runs out, it is unlikely that time limits are problematic. Similarly, if many examinees fail to reach a substantial number of items near the end of the test or respond to those items by rapidly guessing, it is likely that time constraints are significantly impacting test scores. Outside of these extremes, the evidence suggests that timing data may provide evidence that time constraints are impacting scores. The absence of such evidence, however, only means that examinees are maintaining a constant pace across the item presentation sequence. This pattern of behavior may result from the fact that they have no need to change pace because the allotted time limit is sufficiently generous; it also may indicate that examinees have practiced completing the test in the allotted time limit and have learned to maintain a speeded but consistent pace.

Based on the results presented in this chapter, it seems that patterns of not-reached items and rapid guesses are a potentially useful but fairly insensitive measure of the impact of time constraints. Examining patterns of drop-off in seconds per item later in the item presentation sequence may be more sensitive. Both of these approaches may raise a flag suggesting that closer examination is needed, but they are unlikely to provide direct evidence about the magnitude of the effect. The main reason for this is that there is ample evidence to suggest that when faced with time constraints examinees will increase their pace across all items to minimize the need for highly speeded responses near the end of the timed section. Although they may not be successful in maintaining a steady pace, they are likely to make an attempt to answer at a similar rate across the entire sequence.

In this context, it is interesting to reconsider the results reported by Swanson et al. (2001). Their results showed that less proficient examinees spend more time on relatively easy items and more proficient examinees spend less time on easier items. This may be evidence of a strategy on the part of the less proficient examinees to answer more difficult items quickly because they recognize that they do not have sufficient time to devote to these more complex items. Apparently, the more proficient examinees can answer the relatively easy items quickly and so have more time to devote to the more difficult items. There is every reason to believe that even if the strategies are more implicit than explicit examinees do use fairly sophisticated strategies to optimize their scores. It is unlikely that most examinees will engage in undifferentiated "solution behavior" until they discover that they have only a few seconds left and then rapidly guess the answers to the remaining items.

The Use of Timing Data to Improve Score Interpretations

There are a number of conclusions that can be drawn from the research about item review and the relationship between proficiency and response time. The results related to item review dovetail with those reported by Harik et al. (2018) and those displayed in Figures 6.1 and 6.2 to make it clear that examinees will use more testing time than is necessary. Some—and perhaps most—of this extra time is spent reviewing questions that already have been answered. In general, this review appears to have little impact on examinee scores. The results also show that examinees spend more time on items that they answer incorrectly, in part, it would seem, because they are more likely to review items when they are unsure of the answers.

We previously summarized the first of these results as: *When extra time is available, examinees are likely to use more time than they need to achieve their highest score.* We summarized the second result as: *It takes more time to be wrong.* These statements are simple descriptions of what was observed, but the second statement might suggest to practitioners that response time can tell us something about proficiency. It is well established that fluency (the ability to respond quickly) may be an indicator of a higher level of mastery than simple accuracy (see Chapter 4). Even if the argument for incorporating response time into the score is not based on expanding the construct to include fluency, there are more purely statistical reasons for incorporating response time: scores that incorporate accuracy and response time may simply be more precise than those based on accuracy alone (van der Linden, 2007; van Rijn & Ali, 2017). These recently proposed models have the potential to increase the precision of scores without increasing the number of items administered.

The problem (or perhaps more correctly, one of the problems) with drawing inferences based on response time data is that fluency is only one factor that impacts response time. Fluency with regard to the proficiency of interest is unlikely to impact an examinee's propensity to review items. Highly proficient (and fluent) examinees still may feel a compulsion to check and recheck their work. If examinees use more time than they need, response time will not provide a direct measure of fluency. That is not to say that response time is uncorrelated to proficiency; clearly it is. The problem is that other factors that impact the propensity to use more time to review items (e.g., self-confidence) are not sources of random error; they are systematic. As such they do not simply reduce the usefulness of using response time as collateral information in producing scores; they introduce bias in the scores.

If response time were used as collateral information under the current administration conditions, it would penalize examinees who are more cautious and spend more time reviewing responses. This certainly would be unfair and could be counterproductive as well (as cautious attention to detail very well may be an attractive characteristic for a physician). If, alternatively, the conditions of administration were changed by telling examinees that response rate will impact their scores, examinees likely would change their behavior. However, two examinees who are capable of completing the test with the same speed and accuracy may still respond at different rates (and receive different scores) if the scores are influenced by the ability to make good judgments about how to optimize pacing with respect to the scoring algorithm.

In considering the contribution that speed might make to an examinee's score, it is worth noting that there may be contexts in which speed is clearly a part of the construct of interest. In emergency situations or other critical care settings, rapid responses may be important. If measures of both speed and accuracy are central to the construct of interest, the logic related to the use of information about response time is changed. Such items could be presented in a separately timed section with separate instructions for examinees. Even under these conditions, however, it will be necessary to demonstrate that the characteristics that impact response time on the items are the same characteristics present in these time-critical practice settings. This is likely to be challenging.

Using Timing Data to Improve Test Development

Perhaps the most obvious use of timing data to improve test development is to use it to establish time limits that will be consistent with the intended inferences that are to be made based on the test scores. In an optimal scenario, such data would come from randomized experiments, but the results reported in this chapter make it clear that useful information can also be collected in observational studies.

Standardized testing is a process for collecting information to make an inference or decision (usually about examinees). Optimizing test development with respect to time means collecting the maximum amount of information in a fixed amount of time, given the construct(s) of interest. It also means collecting data that are maximally useful (per unit) for the decision or inference that is to be made. The research already discussed by Swanson et al. (2001) relates directly to this question. For example, items with more words will generally take longer to complete. Test developers therefore should be cautious about creating complex (and lengthy) scenarios unless there is evidence that there is a payoff in terms of improved measurement. This improvement could be in the form of items that are more discriminating (providing increased statistical information), or it could be in the form of items that are viewed as more directly measuring the construct of interest. In either case, it is important to be aware of the tradeoffs that may be required to increase authenticity. More complex scenarios *may seem* more realistic, but they almost certainly *will* require more testing time. Empirical evidence should be collected whenever possible to demonstrate that this additional time is justified.

The same logic applies when considering the addition of other types of stimulus materials. It is easy to understand why a test developer might make the decision to require that the physician identifies the correct diagnosis after listening to a recording of heart sounds rather than simply reading a description of the heart sounds, even if the audio version of the item requires more time. Nonetheless, audio, video, and graphical stimulus materials are likely to add to response time (Holtzman, Swanson, Ouyang, Hussie, & Albee, 2009). Even if it is impossible to directly evaluate the value added by the more complex stimuli, the additional time requirements should be evaluated.

As explored in Chapter 3, regardless of the decisions that are made in selecting the types of items to include on the test, it will be important to construct test forms (or individually timed sections) so that any impact resulting from time limits will be similar across test forms. Model-based approaches such as that proposed by van der Linden (1998, 2005) represent one strategy. In contexts in which prior information about the time intensity of individual items is not available (e.g., when items have not been pretested), characteristics of items such as word count or other linguistic characteristics might be used to build forms that will have similar timing requirements.

Implications for Other Testing Contexts

Most of the results discussed in this chapter come from data collected in the context of medical licensing examinations. We focused on these results in part because numerous studies have been published based on data collected since these examinations moved to computer administration. Additionally, the high-stakes nature of the tests reduces the potential confounding effects of variable levels of engagement. Nonetheless, it is sensible to consider the extent to which these results are likely to generalize to other settings.

As we have already discussed, licensing examinations are achievement tests. They also are administered under high-stakes conditions. The results of these tests determine whether or not a physician can practice. In some sense that makes these tests similar to university entrance examinations; in both cases, examinees are likely to make substantial efforts to maximize their scores. This includes both organized test preparation with respect to the content of the examination and practice with pacing to minimize the impact of time constraints. There are, however, other aspects of achievement tests that may differ substantially from the tests described in this chapter. For example, much of the research on the SAT (the current form of which may reasonably be viewed as an achievement test) was carried out during a time in which there was a penalty for guessing. Items on the SAT also were administered at least partially in difficulty order. These characteristics could affect the impact of time constraints substantially. Bridgeman, Trapani, and Curley (2004) reported that additional time on the math section of the SAT benefitted more proficient examinees. This likely occurred because the order of items and a penalty for guessing led to a circumstance in which less proficient examinees ran out of knowledge before they ran out of time. Changes in either the conditions of administration or the motivation of examinees may affect the impact of time constraints. For example, it is unlikely that the results reported in this chapter would generalize to low-stakes (for examinees) testing contexts like those described in Chapter 11. These characteristics should be taken into consideration in interpreting the results presented in this chapter.

Note

[1] The speed-accuracy tradeoff has been studied in numerous performance-related contexts and has even been observed in insects (Chittka, Dyer, Bock, & Dornhaus, 2003).

References

Bridgeman, B. (2020). Relationship between testing time and testing outcomes. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 59–72). Abingdon: Routledge.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I Scores. *Journal of Educational Measurement*, *41*, 291–310.

Chittka, L., Dyer, A., Bock, F. & Dornhaus, A. (2003). Bees trade-off foraging speed for accuracy. Nature, *424*, 388.

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, *16*, 167–188.

Feinberg, R. A., & Jurich, D. (2018, April). Using rapid responses to evaluate test speededness. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Gulliksen, H. (1950). *Theory of mental test scores*. New York, NY: John Wiley and Sons.

Ha, L. A., Marsic, G., & Yaneva, V. (2017, October). Predicting item response time using linguistic features. Paper presented at the Timing Impact on Measurement in Education Conference, Philadelphia, PA.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., … Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement*, *55*, 308–327.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, article 150.

Holtzman, K. Z., Swanson, D. B., Ouyang, W., Hussie, K., & Albee, K. (2009). Use of multimedia on the Step 1 and Step 2 Clinical Knowledge components of USMLE: A controlled trial of the impact on item characteristics. *Academic Medicine*, *84*, Suppl, S90–S93.

Kahraman, N., Cuddy, M., & Clauser, B. E. (2014). Modeling pacing behavior and test speededness using latent growth curve models. *Applied Psychological Measurement*, *37*, 343–360.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606–619.

Kyllonen, P., & Thomas, R. (2020). Using response time for measuring cognitive ability illustrated with medical diagnostic reasoning tasks. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 122–141). Abingdon: Routledge.

Lee, S., & Wollack, J. A. (2020). Concurrent use of response time and response accuracy for detecting examinees with item preknowledge. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 165–175). Abingdon: Routledge.

Lovett, B. J. (2020). Extended time testing accommodations for students with disabilities: Impact on score meaning and construct representation. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 47–58). Abingdon: Routledge.

Ouyang, W., Harik, P., Clauser, B. E., & Paniagua, M. A. (2019). An investigation of answer changes on the USMLE® Step 2 Clinical Knowledge examination. *BMC Medical Education*, *19(1),* 389. doi:10.1186/s12909-019-1816-3.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*, 261–270.

Schnipke, D. L. (1995, April). Assessing speededness in computer-based tests using item response times. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Sireci, S. G., & Botha, S. M. (2020). Timing considerations in test development and administration. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 32–46). Abingdon: Routledge.

Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. *Journal of Educational Measurement*, *8*, 275–278.

Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E., & Holtman, M. C. (2001). Relationships among item characteristics, examinee characteristics, and response times on the USMLE Step 1. *Academic Medicine*, *79* (10 October Suppl), S114–S116.

Swanson, D. B., Holtzman, K. Z., Albee, K., & Clauser, B. E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine*, *81* (10 October Suppl), S52–S55.

Swineford, F. (1956). Technical manual for users of test analyses (Statistical Report 56–42). Princeton, NJ: Educational Testing Service.

van der Linden, W. J. (1998). Optimal assembly of educational and psychological tests. *Applied Psychological Measurement*, *22*, 195–211.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308.

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*, 317–345.

Wainer, H. (1971). Piecewise regression: A simplified procedure. *British Journal of Mathematical and Statistical Psychology*, *24*, 83–92.

Wise, S. L. (2015). Response time as an indicator of test taker speed: Assumptions meet reality. *Measurement: Interdisciplinary Research & Perspective*, *13*, 186–188.

Wise, S. L. (2017). Rapid guessing behavior: Its indication, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61.

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150–164). Abingdon: Routledge.

7

Timing Considerations for Performance Assessments

Melissa J. Margolis, Matthias von Davier, and Brian E. Clauser
National Board of Medical Examiners

For well over a century, researchers across disciplines have studied different aspects of the relationship between speed and performance. The effect of speed on decision-making has been studied and reported in the psychology literature since the 1800s (e.g., Donders, 1868); concern about the impact of time limits on test performance has been documented since at least the 1920s (e.g., Yerkes, 1921). More recently, educational measurement researchers have devoted considerable attention to understanding how timing and time limits impact test scores. Much of this research has been conducted within the context of multiple-choice examinations, and numerous large-scale studies have examined both the impact of time limits (Bridgeman, Cline, & Hessinger, 2004; Bridgeman, Trapani, & Curley, 2004; Evans & Reilly, 1972; Harik et al., 2018; Wild, Durso, & Rubin, 1982) and how examinees use available time (Kahraman, Cuddy, & Clauser, 2014; Schnipke, 1995; Schnipke & Scrams, 1997) in this context. Despite the efforts dedicated to research in this area, relatively little work has been done in other testing contexts. The purpose of this chapter is to extend the discussion of timing-related issues to an area that has received considerably less attention in the literature: performance assessment. We begin by discussing the (somewhat limited) published literature that does exist in this area; this work mainly relates to writing tasks. We then address timing considerations in the context of other types of performance assessments and report on a previously unpublished experiment examining timing with respect to performance on computer-based case simulations that are used in physician licensure. Finally, we discuss literature on psychometric models that take into account both time limits and working speed in addition to performance. We end with a discussion of how research around time limits can be enriched by looking at log files and process data.

Time Limits in Performance Assessment

Essay Examinations

It is perhaps surprising to realize that, two decades into the 21st century, writing tasks remain the primary focus of research dedicated to understanding timing considerations in non-multiple-choice assessments. More studies addressing the effects of time limits on complex simulation- and game-based assessments are likely to emerge in the future, as those types of assessments are being used with more frequency in both formative and summative assessment contexts. The current state of the literature, however, is that relatively little published research exists that examines timing in the context of performance assessment. The research area that has been the exception and that has received continued attention over the years is the impact of time limits on scores for essay questions.

Much of the work around time limits on essay tests focuses on the evaluation of timing with respect to performance on specific tests, but there are also some studies that address the impact of time limits on essay scores more abstractly. We will begin our review by discussing this latter category of research and then consider the research projects that are specific to individual testing programs.

In a study described by Biola (1982), 96 freshman at Georgia State University were required to complete an essay test. Students were assigned to one of two topics and each topic was administered under one of two timing conditions (45 or 120 minutes). Analyses investigated the impact of time on the resulting scores as well as whether there were differences associated with the topics or the interaction between topic and time limit. The findings were straightforward; more time was associated with higher scores and there were no significant effects for topic or for the interaction between topic and time limit.

Caudery (1990) reported on a similar, but less structured, small-scale study implemented in the context of a class for Cypriot students preparing for the General Certificate of Education Ordinary Level (O-level) English examination. In this study, 24 students completed two different essays under two different conditions. In the first condition, students had a 40-minute time limit to complete the essay in the classroom. In the second condition, students had 1 hour to work on the essay in class and then had the next 2 days to complete the essay at home. Three criteria were used to evaluate the essays: organization, language, and overall impression. The only score that differed significantly across the two timing conditions was for language. The researcher was surprised that organization was not improved when more time was provided, although he noted that the students had specifically practiced writing essays within a 40-minute timeframe in preparation for the O-level examination. The findings from these two studies do little more than demonstrate the obvious, which is that time limits may impact scores on essay examinations. The small sample sizes and lack of information about how the specific examinee groups might have impacted the results make it difficult to draw other general conclusions from this research.

The remaining studies that we reviewed were motivated by an interest in collecting validity evidence for specific assessments. Livingston (1987) reported on two studies that used a counterbalanced design to examine the impact of additional time on essays completed as part of the New Jersey College Basic Skills Placement Test. Two groups of examinees participated in the study, all of whom took the test as volunteers (i.e., their scores were not used to make placement decisions). The first group of examinees completed two essays under different timing conditions: the standard 20-minute time limit and an extended 30-minute time limit. Both the order of the essay prompts and the order of the timing conditions were counterbalanced. In the second group, examinees were assigned to the same two timing conditions except that in the 30-minute condition the additional time was provided for planning the essay; examinees either had 10 minutes to plan the essay and 20 minutes to write or they had 20 minutes to write with no planning time. The results indicated that the most significant difference was between essay prompts; one prompt resulted in higher scores than the other. There were no significant differences between timing conditions, suggesting that additional time (whether for planning or writing) was likely to have relatively little impact on scores.

Klein (1981) reported on the results of a randomized experiment examining timing issues on the California Bar Examination. For one administration of the examination, a special section was added. Applicants were told that the section was optional, but that if they failed based on the regular examination, their performance on the special section would be included in a composite score and could result in a passing decision. If examinees passed based on the regular examination, the special section would not be considered. The special section contained two essays (one related to "business law" and one related to "trial law"), and a counterbalanced design was used to assign one essay to a shorter (55-minute) timing condition and one to a longer (90-minute) timing condition. The results showed that examinees scored higher on average when given more time, but there was no evidence that additional time differentially impacted groups defined by age, sex, race, type of law school attended, or repeater status. There was similarly no significant relationship between the improvement associated with having more time and scores on the full examination.

In 1992, Hale reported on a study that was similar to Livingston's (1987) study; 820 international students served as paid volunteers and were tested to evaluate the impact of time limits on the Test of English as a Foreign Language (TOEFL) test of Written English. Two essay types were examined: prose topics and chart/graph topics. For each essay type, two prompts were used along with two time limits (standard 30 minutes and extended 45 minutes). Time limits and topics were counterbalanced to control for order effects. To allow for estimation of parallel forms reliability, a group of examinees responded to both prompts with the same time limit (either 30 minutes or 45 minutes). Finally, an additional group of examinees participated in a supplemental condition in which one prompt was completed with the standard time limit and the other was completed with a 15 minute planning period followed by 30 minutes for writing. Results were reported both in terms of correlations and mean performance across timing conditions. Correlations between scores on essays written with a 30-minute and 45-minute time limit were similar to correlations between essays written with the same time limit. For the prose topics, the mean correlation between time limits was 0.77 and the mean correlation between topics with the same time limit was 0.75. For the chart/graph topics, the mean correlation between time limits was 0.69 and the mean correlation between topics with the same time limit was 0.74. Generally speaking, the results showed improvements in scores when examinees had additional time to complete the essay. The effect was statistically significant and resulted in an improvement of slightly more than a third of a standard deviation. Additional analysis indicated that the magnitude of the improvement associated with having an additional 15 minutes was similar for high- and low-proficiency examinees. The supplementary condition in which examinees had an additional 15 minutes to plan their essays before they began writing did not lead to a significant improvement in essay scores.

Powers and Fowles (1997) conducted a similar study in which they recruited volunteers to examine the impact of time limits on the Graduate Record Examination (GRE) Writing Test. Four essay prompts were used with test forms spiraled so that half of the examinees completed prompt A followed by B and half completed prompt C followed by D. Two separate administrations were required to counterbalance the timing conditions: for one administration, there was a 40-minute time limit for the first essay and 60 minutes for the second; for the other administration, the timing was reversed. Examinees were also administered a questionnaire to gather information about their self-perception related to their ability to write quickly and their level of frustration with timed writing tests. Both the time limit and the slowness/quickness variable extracted from the questionnaire were significantly related to essay scores; there was, however, no interaction indicating that examinees who perceived themselves to be slower or quicker benefitted more from additional time.

The fact that Powers and Fowles (1997) showed no differential impact associated with providing additional time on the GRE analytic writing assessment opened the way to efforts to reduce the overall time allotted to essays for the redesigned GRE. Robin and Zhao (2014) describe an experiment designed to evaluate the best way to divide a reduced amount of time between the issue and argument essays that comprise analytic writing. The results may be most noteworthy because they highlight the complexity of making interpretations based on data collected from volunteers taking the test under low-stakes conditions. Less than half of the 1,183 volunteers who had previously taken the GRE could be used in the final analysis because they either did not complete both the issue and argument tasks or because the required scores from the operational test were not available. The original operational timing for the issue and argument tasks was 45 and 30 minutes, respectively. In the experimental conditions, the timing was 40 minutes and 20 minutes, 35 minutes and 25 minutes, or 30 minutes and 30 minutes. Although the volunteer group had scores from the operational test that were above the national average, their performance on the experimental essays was approximately a standard deviation below that average (even when the timing was the same as the original operational condition). Additionally, with the issue essay, providing more time (30, 35, and 40 minutes) led to respectively lower scores. This would seem to provide strong evidence that motivation matters.

These results related to using volunteers to evaluate the impact of time limits on performance are similar to those reported for evaluations of multiple-choice-based tests delivered in low-stakes settings. For example, Bridgeman et al. (2004) reported that 46% of the volunteers they recruited had to be removed from the analysis because their scores on the experimental section were sufficiently different from those on the operational sections that it was reasonable to conclude that they had not given serious effort to the voluntary experimental section. A similar score anomaly was reported by Evans and Reilly (1972) for a study of the Law School Admissions Test in which a subset of the examinees completed the examination at free test centers. Taken together, these results argue for considerable caution in generalizing from low-stakes experimental settings to high-stakes operational conditions. This is troubling, because essentially all of the research on essay examinations is based on low-stakes performance. The one exception is the study by Klein (1981) in which performance on the experimental essay section could help otherwise failing examinees, but even in this study poor performance on the experimental section could not impact the examinee.

If we put this substantial limitation of the research base aside, we might reasonably conclude that providing additional time often—but not always—leads to improved performance. In those instances in which it does lead to improved performance, the improvement appears to be uniform across subgroups, whether those groups are defined by proficiency or in terms of self-reported need for more time. It may, however, be imprudent to put the limitations of the research base aside, because in the context of multiple-choice items, differences in impact for time limits are most clearly apparent in studies implemented in high-stakes conditions (e.g., Bridgeman et al., 2004; Harik et al. 2018).

Simulations Used in Medical Licensing

Standardized-Patient-Based Clinical Skills Assessment

As noted previously, much of the timing research related to performance assessments has focused on essays. Another context in which timing considerations have been studied is the performance assessments that are part of the United States Medical Licensing Examination (USMLE). The USMLE, which is required for granting of a license to practice medicine in the United States, comprises four separate testing events. The examination sequence includes two different performance assessments: the Step 2 Clinical Skills (CS) examination and the computer-based case simulation component that is part of the Step 3 examination. Step 2 CS is a day-long examination designed to assess the ability of physician trainees to gather information from patients, perform physical examinations, and communicate findings to patients and colleagues. Laypeople are trained to portray "standardized patients," each with a specific clinical complaint. During the examination, examinees interact with a series of 12 of these "patients" and are allotted a total of 25 minutes for each encounter: A maximum of 15 minutes is permitted for the actual patient interaction in the examination room, and the remaining time (a minimum of 10 minutes) is used to complete a structured patient note that requires documenting the findings and recommendations resulting from the encounter.

A series of studies was implemented to examine how the amount of time used by examinees relates to examinee characteristics, characteristics of the simulated cases, and scores from the Step 2 CS examination. Swygert and colleagues (Swygert, Scott, Swanson, McKinley, & Boulet, 2008, 2009) used hierarchical linear modeling procedures to examine how total time taken in the patient encounter relates to encounter characteristics. Results indicated that examinees use more time for encounters earlier in the test day and that on average they use the most time on the first encounter; this finding seems to suggest a warm-up effect. The results also show that examinees from U.S. medical schools use less time than those from international schools and that women use less time than men. Case content (i.e., the organ system the case focused on) was not significantly related to use of time.

Two follow-up studies examined how time devoted to different aspects of the encounter related to scores. Swygert, Muller, Swanson, and Scott (2009) examined the relationship between two timing variables (total time spent with the standardized patient and time spent closing the encounter) and the examinee's communication and interpersonal skills score from the same encounter. (The "closing" period is the time after completing the patient history and physical examination and before leaving the room; common activities at this point in the patient encounter include explaining the findings to the patient, answering the patient's questions, and discussing possible next steps.) Results indicated that both time measures were positively related to scores. They also showed that scores improve across the test day, with examinees on average receiving their lowest score on the first encounter of the day.

Finally, Swygert, Muller, Scott, and Swanson (2010) examined how the amount of time examinees spent writing the patient note related to scores on that component of the test. Results indicated that spending more time writing the note was associated with higher scores; each additional minute spent writing was associated with a 0.04 score increase. To put this result in perspective, the 0.04 increase represents well under a tenth of a standard deviation for the note score; a two standard deviation increase in the amount of time an examinee spent on a note would be expected to be associated with a fifth of a standard deviation increase in the score.

Taken together, the studies by Swygert and colleagues indicate that the use of time in these simulations varies across demographic groups and that the amount of time used on a component positively covaries with the scores for that component. What is less clear is whether there are any existing causal relationships between those characteristics and time. The other potentially important result reported in these studies is that examinees use the most time on the first encounter of the day and also have the lowest scores on that encounter. This may be an indication that aspects of the test that are unrelated to the proficiencies of interest are impacting examinee performance.

Computer-Based Case Simulations

The finding of systematic score increases along with a decrease in the amount of time to complete tasks across the testing day closely parallels results of a similar study in the context of the USMLE Step 3 computer-based case simulations (Clauser, Margolis, & Clauser, 2017). These are dynamic simulations of the patient-care environment in which examinees are expected to manage each "patient" by gathering data and making patient-care decisions based on both the data they collect and any changes in the patient's condition. (See Margolis & Clauser, 2006; Harik, Clauser, & Baldwin, 2013 for more information about the simulations and associated scoring procedures.) At the time of this research, examinees completed nine cases and had a maximum of 25 minutes to complete each case. Different sets of nine cases were used on different test forms, and for any given test form the order of case delivery was randomized by examinee. Because of the randomization, results for any specific sequence position were not related to the specifics of the cases; all cases occured in each position a similar number of times. Results showed that examinees moved more quickly through consecutive cases occurring in positions one through eight. The first case took on average about 1.5 minutes longer than the second case, and there were modest reductions in time for subsequent cases. The results also showed a parallel increase in mean score as a function of sequence position.

Overall, the results from studies investigating the relationship between time usage and score indicate that familiarity with the interface, or perhaps with the demands of the simulation, leads to improved performance in less time. It is unclear whether the reduction in time used across sequence position reflects increased confidence on the part of examinees, increased facility with the interface, or both. The increase in scores as a function of sequence would seem to suggest that some aspects of the simulation are associated with construct-irrelevant variance in the scores that is reduced with practice.

One final study in the context of the USMLE Step 3 computer-based case simulations approached the question of examination time constraints from a slightly different perspective than has been typical for this area of research. As described above, much of the research investigating time limits both in essay and in other performance assessment testing contexts has addressed the question *does performance improve if examinees are given (or use) more time?* This study is different in that it was motivated by evidence suggesting that the time allocation for the simulations might be unnecessarily generous. Many examinees finished the simulations with time to spare, and post-examination surveys indicated that relatively few examinees reported the desire for additional time. If research suggested that the standard time (25 minutes per case) could be reduced without impacting examinee performance, a larger number of cases could be administered and would (most likely) positively impact the precision of the resulting measure. A 5-minute (20%) reduction for each case would allow for administering two additional cases. If the time limit could be reduced to 15 minutes (an additional 20%), the test length could be increased from nine to fifteen cases. This increase in test length could have a substantial positive impact on the reliability of this part of the examination. (See Clauser, Harik, & Clyman, 2000, for generalizability results related to increasing the number of tasks on this examination.)

The research was completed in the context of the operational Step 3 examination, and at the time of the study, one or more of the nine cases was unscored in order to collect statistical information for quality control and scaling; examinees were not aware of which cases were not scored. One of these unscored cases was used for this research.

Nine cases representing a range of complexity were selected from the case pool. Examinees were randomly assigned both to cases and to timing conditions within cases. For each case, three timing conditions were examined: 15 minutes, 20 minutes, and 25 minutes (standard time). Examinees were aware of these different timing conditions. The bulletin of information for the examination described the fact that the cases could be administered with differing time allocations, the time limit for each simulation was presented at the beginning of the case, and a timer provided feedback about the amount of time remaining.

Analyses ranged from simple descriptive statistics to analysis of covariance (ANCOVA) and were intended to provide insight into several different issues, including (but not limited to) whether there were differences across timing condition after accounting for a number of covariates, whether there was an interaction between examinee proficiency and time (reflecting a differential impact for time limits on a subset of examinees), and whether there was evidence that changes in timing impact score validity.

Results indicated a nearly perfectly consistent pattern in which the mean scores decreased as time decreased. For all nine cases, differences existed across the levels of the timing variable. For five of the cases, there were significant differences between the variances in the scores across the levels of the timing condition; the variances consistently increased as the allotted time decreased. ANCOVA was used to examine the score differences across timing conditions using examinee characteristics as covariates; these covariates included a proficiency estimate based on the multiple-choice section of the examination, examinee gender, English language status, location of medical school (U.S. or international), and a variable representing whether the examinee was completing the examination for the first time.

Table 7.1 shows the estimated timing-condition means based on the covariates. There were statistically significant differences between scores across the timing conditions for each of the nine cases. For the majority of cases, a reduction in time from 25 to 20 minutes did not result in a significant score reduction; a reduction from 20 to 15 minutes did.

The described results suggest that a reduction in the time limit will impact overall examinee performance; one obvious question following from that finding is whether this impact is uniform across examinees or whether some examinees are affected more than others. The pattern of increasing standard deviations associated with shorter time limits might suggest that the impact is not uniform. Analytic results indicated significant interactions between examinee proficiency and timing condition for four of the nine cases (cases 1, 3, 7, and 9).

Graphic representation of the results provides insight into the specifics of these interactions. Figures 7.1 and 7.2 present plots in which the score on the studied case is plotted against proficiency (based on the multiple-choice portion of the test), with separate lines representing the separate timing conditions. Figure 7.1 presents an example indicating that decreasing the time limit has relatively little impact on performance for high-proficiency examinees and substantially greater impact on low-proficiency examinees. Figure 7.2 presents a different pattern in which it appears that with standard time, the case does not discriminate well between examinees with medium to high proficiency; with reduced time (15 minutes), the case does not discriminate between lower-proficiency examinees. The plots for cases 1 and 3 are similar,

Table 7.1 Estimated marginal mean scores and standard errors by timing condition

	Timing Condition		
Case	15 minutes	20 minutes	25 minutes
1	5.102 (0.048)*	5.729 (0.046)	5.775 (0.040)
2	4.135 (0.055)*	4.479 (0.056)	4.595 (0.053)
3	5.808 (0.060)*	6.253 (0.050)	6.365 (0.051)
4	4.746 (0.047)*	5.161 (0.048)	5.182 (0.054)
5	6.980 (0.047)	6.931 (0.040)*	7.100 (0.045)
6	4.877 (0.037)*	5.168 (0.038)*	5.065 (0.032)
7	4.838 (0.042)*	5.325 (0.041)*	5.600 (0.040)
8	5.314 (0.086)*	5.828 (0.081)	5.898 (0.091)
9	5.287 (0.042)*	5.437 (0.042)	5.459 (0.048)

* Estimated mean is significantly different from the 25-minute condition (p <0.05)

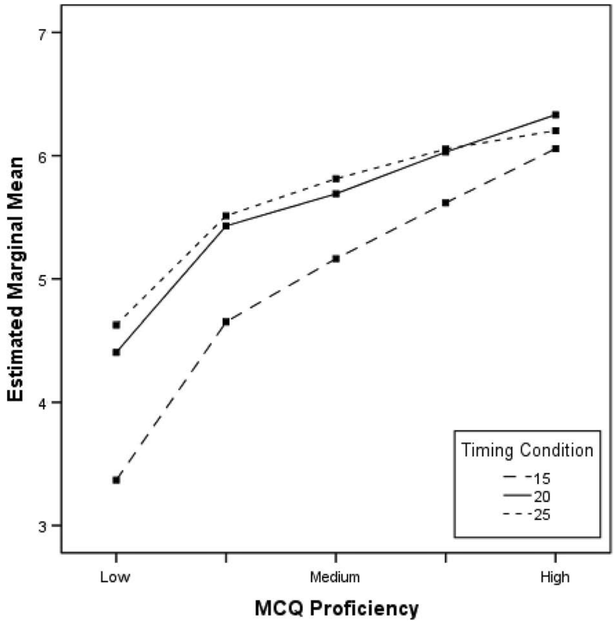


Figure 7.1 Estimated marginal means plotted against examinee proficiency for case 1.

while those for the remaining two cases are different. The plotted values represent a type of empirical item characteristic curve for the studied cases.

The implication of these results is that a 40% reduction in testing time may make a difference; a 20% reduction may not, although a lack of significance does not demonstrate that subgroups within the population are unaffected. Examinee performance tended not to differ

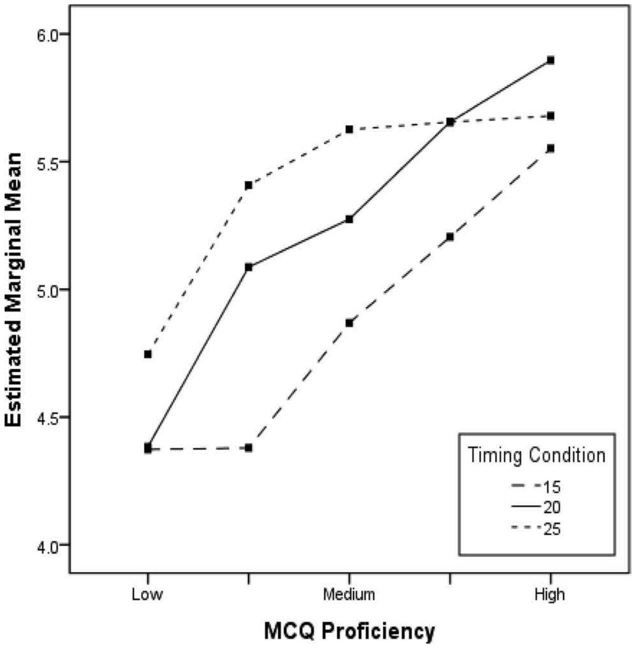


Figure 7.2 Estimated marginal means plotted against examinee proficiency for case 7.

significantly between the 20- and 25-minute conditions, but it did tend to differ significantly between the 15- and 25-minute conditions. Beyond this one generalization, the results were highly case dependent: the more substantial shift in expected performance tended to be associated with a change in timing from 20 minutes to 15 minutes, but for a few cases the change from 25 minutes to 20 minutes was more significant. Similarly, for most cases, a reduction in timing provided a similar performance disadvantage to examinees across the examinee proficiency range, though the expected advantage varied by examinee proficiency level for four cases.

Previous research on the impact of varying time limits has shown that the results may be context specific. Given that finding, generalization of the results of this study should be made cautiously. That being said, this study extends the available results by providing information about the impact of timing on complex constructed-response items; it is one of the few studies that was conducted under high-stakes conditions and also is one of the few studies that shows differential effects for time limits across examinee groups.

Psychometric Approaches to Modeling Data from Time-Limited Tests

The above review has shown that time on task and performance may be related. In the case of performance tasks, this appears to be an example of the well-studied speed-accuracy tradeoff. Complex performances require continued effort, and insufficient time will lead to sub-optimal outcomes. In this section, we discuss why, given this relationship, common approaches to integrating timing data into psychometric models may not be appropriate for analysis of performance tasks. We then discuss how these existing models can be extended to accommodate complex tasks administered in tests with time limits. We close with a section that provides some research directions based on current developments around process and sequence data analysis.

Current Modeling Approaches Combining Time and Response Data

One could argue that most psychometric models ignore time on task. Even if time is included in the model (e.g., Klein-Entink, Fox, & van der Linden, 2009), working speed is the primary concern; time limits are not typically made explicit in the model. In addition, models for time used per task either implicitly or explicitly assume that the association between speed and accuracy is the same for all items. For example, the hierarchical speed-accuracy model (van der Linden, 2007) includes two latent variables and assumes that the relationship between item response and accuracy are positive for all items. The model additionally assumes that the relationship between working speed and time used is strictly monotone and directed the same way for all items. The simplicity of this model may seem attractive, but there is a drawback: the relationship between accuracy and speed is modeled at the latent variable level as a correlation, while in real data (e.g., Yamamoto, Khorramdel, & von Davier, 2013) time on task and probability of success may be positively related for difficult (complex) items and negatively related for easy (simple) items. Appropriately addressing this relationship would require a model that allows negative as well as positive associations of a working speed variable at the item level. It would also reconceptualize the speed variable as “good time management” rather than “working speed,” as it would relate to the ability of an examinee to adjust her/his time use to the difficulty of the task at hand.

Model Extensions for Time Limits

Some recent extensions to this model have explicitly taken into account the fact that most tests are administered with time limits. Whether these are writing tests such as single- or multi-prompt essays, mixed-format tests that contain more traditional item formats, or game- or simulation-based tasks, most come with some explicit maximum time limit.

Item response models have been extended to allow for the effects of time limits in various ways. One of the earliest examples is the Rasch Poisson Count Model. In its simplest form, it does not contain a time limit parameter, but it has been extended to include one. The item scores are assumed to follow the model:

$$P\left(R=r \mid \tau_i,\, \theta_v,\, \beta_i\right)=\exp\left(-\tau_i\theta_v\beta_i\right)\frac{\left(\tau_i\theta_v\beta_i\right)^r}{r!}, \tag{7.1}$$

where $\tau_i = \exp\left(\vartheta_i\right)$ represents a time limit parameter, $\theta_v = \exp\left(-\theta_v\right)$ is the examinee ability, and $\beta_i = \exp\left(b_i\right)$ is the item difficulty. This model allows for analyzing data based on a set of items that is administered under different time limits. In this model, scores represent deviations from an ideal solution (zero is best, counts of errors or omissions lead to higher scores); holistic human-rater-based scores will need to be recoded to align with this model. Research using the Rasch Poisson Count Model suggests that it has good fit for datasets obtained from a variety of tests (e.g., Doebler & Holling, 2015; Verhelst & Kamphuis, 2009); application of the model to scores from the previously described computer-based case simulations is currently being examined.

Yamamoto's Speededness HYBRID model provides another example of how time limits can be included in scoring models (Boughton & Yamamoto, 2006; Yamamoto & Everson, 1997). Under this model for binary item response data, it is assumed that time limits may lead respondents to switch from a skill-based response process that involves active information processing to a heuristic choice or even guessing-based response as they move through the item sequence. This change typically would be expected when time is about to run out. Although the model implicitly assumes that test takers complete test items in the order they are presented, it has been shown to provide good fit. If test takers return to items that were previously skipped or were previously answered by guessing and they subsequently attempt them in a fully engaged mode, this does not contradict the HYBRID model assumption as the model allows for the possibility that some test takers answer all items in an engaged mode.

The Speededness HYBRID model is based on a discrete mixture distribution of respondent groups who switch from a problem-solving behavior to a simpler approach to responding at different item positions. The approach may be based on heuristics or, at its most extreme, test takers may be guessing at random. The probability of a response pattern takes the following form:

$$P\left(x_1,\ldots,x_I \mid \theta\right)=\sum_{i=1}^{I}\pi_{S=i}\prod_{j=1}^{i}P\left(x_j \mid \theta\right)\prod_{k=i+1}^{I}g_k, \tag{7.2}$$

where g_k is the guessing probability on item k, $P\left(x_j \mid \theta\right)$ is the probability of response x_j given ability θ, and $\pi_{S=i}$ is the probability of switching to a random/heuristic strategy after item i.

This model was shown to improve item difficulty estimation and allowed the TOEFL program to assess speededness and shorten the test (Boughton et al., 2006). Extensions to performance items with polytomous scores can be implemented via the polytomous HYBRID model (von Davier, 1996) and the general modeling framework introduced by von Davier (2005, 2008) for diagnostic testing and multidimensional item response modeling.

Another potentially useful approach was described by Lee (2007). In this model, time to completion of a task can be impacted by time limits that may interfere with completing all items. This model assumes dichotomously scored tasks, but extensions to polytomous or count scores are straightforward. (See Lee & Chen, 2011, for a recent review of models for timing data and item responses.) This approach seems promising, because examinees who do not finish their work before the time limit are likely to have lower scores than those who have sufficient time to finish. It is clear that many of these approaches can be used to inform researchers about the effects of time limits. That being said, the fact that time measures are indirect indicators of examinee behavior and the extent to which time limits actually impact their scores highlights the need for continuing research in this area.

Process Data Analysis to Inform Test Time Limits

Time on task and the impact of time limits were studied long before computers became common in test delivery, but these studies were labor intensive and provided limited data. With the advent of computer-based testing, rich process data has become readily available. Process data results from automated recording of each action taken by the examinee along with a time stamp for the action. Actions may include mouse clicks, menu choices, keystrokes, or any other interaction with the test delivery platform. These data are structured in the sense that all entries in the log file where they are recorded are well defined; they are, however, not structured like other response data collected as part of test administration. Process data are, like text or natural language, sequence data (Dong & Pei, 2007; Sukkarieh, von Davier & Yamamoto, 2012), and different techniques are needed to analyze these new types of data (He & von Davier, 2015, 2016).

Process data appear to be another opportunity to study the effect of time limits on performance. For writing tasks, Almond, Deane, Quinlan, and Wagner (2012) and Deane (2014) provide insight into how key-stroke data can be collected as a tool for gaining a deeper understanding of how test takers distribute the available time between planning, writing, and reviewing/revising their work. In particular, Deane (2014) develops concepts around process and product features of the writing task that may be relevant correlates of how respondents utilize the available time for task completion.

Data on how time is allocated between different activities may help to delineate optimal versus suboptimal time usage on performance tasks. Classifying process data into categories such as orientation, production, review, and disengagement may provide a basis for estimates of how much time is spent on different types of activities, and ultimately how this allocation of time relates to the score obtained on the performance task.

Pohl, Ulitzsch, and von Davier (2019) and Pohl and von Davier (2018) describe an approach that combines timing information with data about what part of the assessment was not reached by the examinee. This approach can be useful when looking at process data in complex performance items where respondents are not expected to reach all parts of the assessment. Assessing whether test takers engage in responding to the items or whether they utilize heuristics to select an option quickly, or even guess a response, can be improved by incorporating timing data (Ulitzsch, von Davier, & Pohl, 2019a,b). This allows for estimating ability only based on those responses that are considered engaged, and can also be understood as a way of disentangling careless responding from effortful, ability-based responding.

Conclusions

The ways in which time limits and response speed interact with performance on tests have been the focus of considerable attention. The impact of time limits on standardized tests comprising multiple-choice items has been carefully studied (see Chapter 5), but as we stated at the beginning of this chapter, less attention has been given to time limits for performance assessments. Having fairly exhaustively reviewed the available literature in this area, the limitations of that literature should be apparent. Even in the context of essay-based examinations, where relatively more research has been carried out, interpretable and generalizable results are difficult to come by because so much of the work is based on examinee responses gathered under conditions that differ markedly from those of the actual examination (i.e., differences in examinee motivation).

Even with those limitations, the results of previous studies make it clear that time limits can significantly impact scores on performance assessments. Although there is little evidence that time constraints differentially impact different demographic groups (defined for example by gender or ethnicity), the results reported in this chapter do show that time limits may differentially impact examinees of different proficiency levels.

The literature reviewed in this chapter also makes it clear that, in comparison to multiple-choice formats, relatively little attention has been given to scoring models that incorporate information about response rate or the impact of time limits. This relative lack of evidence to guide development and administration of performance assessments is problematic for a number of reasons. The importance of understanding the impact of time limits on performance assessments is increased because the value of these more complex tasks is typically linked to the view that they more directly reflect the real-world behavior of interest. If artificial time limits result in response patterns that are systematically different than those in the "real world," the advantage of using a more authentic task may be undermined. This problem can be exacerbated by the fact that there are likely to be aspects of the assessments that differ from the real-world challenges they are meant to approximate. Using a computer interface to manage a patient introduces familiarity with the interface as a variable that could impact both the time requirements and performance. Producing an essay using an unfamiliar word processing system could similarly impact time requirements and scores.

As the number of performance tasks that are administered on computer continues to increase, it is hoped that the availability of additional response time data will yield important evidence to advance our understanding of time limits in the context of performance assessments. Advances in this area have the potential to lead to improved test administration, increased efficiency, and scores that better support the intended inferences about examinees.

References

Almond, R., Deane, P., Quinlan, T., & Wagner, M. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (Research Report No. RR-12-23). Princeton, NJ: Educational Testing Service.

Biola, H. R. (1982). Time limits and topic assignments for essay tests. *Research in the teaching of English*, *12*, 97–98.

Boughton, K., & Yamamoto, K. (2006). A hybrid model for test speededness. In M. von Davier & C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models; Extensions and applications* (pp. 147–156). New York, NY: Springer.

Bridgeman, B. (2020). Relationship between testing time and testing outcomes. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 59–72). Abingdon: Routledge.

Bridgeman, B., Cline, F., & Hessinger, J. (2004). Effect of extra time on verbal and quantitative GRE scores. *Applied Measurement in Education*, *17*, 25–37.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I Scores. *Journal of Educational Measurement*, *41*, 291–310.

Caudery, T. (1990). The validity of timed essay tests in the assessment of writing skills. *ELI Journal*, *44*, 122–131.

Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement*, *37*, 245–262.

Clauser, B. E., Margolis, M. J., & Clauser, J. C. (2017). Validity issues for technology-enhanced innovative assessments. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 139–161). Charlotte, NC: Information Age Publishing.

Deane, P. (2014). Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. ETS research report No. RR-14-03.

Doebler, A., & Holling, H. (2015). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson counts model. *Learning and Individual Differences*, *52*, 121–128.

Donders, F. C. (1868). On the speed of mental processes. *Arch Néerland*, 3, 269–317.

Dong, G., & Pei, J. (2007). *Sequence data mining*. Springer: New York.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*, 123–131.

Hale, G. (1992). *Effects of the amount of time allowed on the test of written English* (Research Report RR-92-27). Princeton, NJ: Educational Testing Service.

Harik, P., Clauser, B. E., & Baldwin, P. (2013). Comparison of alternative scoring methods for a computerized performance assessment of clinical judgment. *Applied Psychological Measurement*, *37*, 587–597.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., … Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement*, *55*, 308–327.

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with N-grams. In A. van der Ark, D. Bolt, S. Chow, J. Douglas, & W. Wang (Eds.), *Quantitative Psychology Research: Proceedings of the 79th Annual Meeting of the Psychometric Society* (pp.173–190). New York: Springer. Doi: 10.1007/978-3-319-19977-1_13

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-Grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749–776). Hershey, PA: Information Science Reference. doi:10.4018/978-1-4666-9441-5.ch029

Kahraman, N., Cuddy, M., & Clauser, B. E. (2014). Modeling pacing behavior and test speededness using latent growth curve models. *Applied Psychological Measurement*, *37*, 343–360.

Klein, S. P. (1981). The effect of time limits, item sequence, and question format on applicant performance on the California Bar Examination. A Report submitted to the Committee of Bar Examiners of the state of California and the National Council of Bar Examiners.

Klein-Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48. https://doi.org/10.1007/s11336-008-9075-y

Lee, Y.-H. (2007). *Contributions to the statistical analysis of item response time in educational testing* (Unpublished doctoral dissertation). Columbia University, New York, NY.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359–379.

Livingston, S. A. (1987, May).The effects of time limits on the quality of student written essays. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring for complex tasks in computer based testing* (pp. 123–167). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. Psychometrika. https://doi.org/10.1007/s11336-019-09669-2

Pohl, S., & von Davier, M. (2018). Commentary: On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology*, *9*, 1988. doi=10.3389/fpsyg.2018.01988

Powers, D. E., & Fowles, M. E. (1997). *Effects of applying different time limits to a proposed GRE writing test* (Research Report RR-96-28). Princeton, NJ: Educational Testing Service.

Robin, F., & Zhao, J. C. (2014). Timing of the analytic writing measure of the GRE revised general test. In C. Wendler & B. Bridgeman (Eds.), *The research foundation for the GRE revised general test: A compendium of studies* (pp. 1.8.1–1.8.8). Princeton, NJ: Educational Testing Service.

Schnipke, D. L. (1995, April). Assessing speededness in computer-based tests using item response times. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Sukkarieh, J., von Davier, M., & Yamamoto, K. (2012). From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks. ETS Research Report Series. ETS RR–12-25.

Swygert, K. A., Muller, E. S., Scott, C. L., & Swanson, D. B. (2010). The relationship between USMLE® Step 2 CS patient note ratings and time spent on the note – do examinees who spend more time write better notes? *Academic Medicine*, *85*(suppl 10), 89–92.

Swygert, K. A., Muller, E. S., Swanson, D. B., & Scott, C. L. (2009). The relationship between USMLE® Step 2 CS communication and interpersonal skills (CIS) ratings and the time spent by examinees interacting with standardized patients. *Academic Medicine*, *84*(suppl 10), 1–4.

Swygert, K. A., Scott, C., Swanson, D., McKinley, D., & Boulet, J. (2008, March). How Step 2 CS examinees use their time in the patient encounter. Paper presented at the International Ottawa Conference on Medical Education and Assessment, Melbourne, Australia.

Swygert, K. A., Scott, C., Swanson, D., McKinley, D., & Boulet, J. (2009). An assessment of encounter timing in a high-stakes standardized-patient based examination. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019a). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level nonresponse. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12188.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019b). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*. doi: 10.1080/00273171.2019.1643699.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

Verhelst, N. D., & Kamphuis, F. H. (2009). A Poisson-Gamma model for speed tests. *Tech. Rep*. Arnhem: Cito.

von Davier, M. (1996). Mixtures of polytomous Rasch models and latent class models for ordinal variables. In F. Faulbaum & W. Bandilla (Eds.), *Softstat 95 - Advances in statistical software 5*, Stuttgart: Lucius & Lucius.

von Davier, M. (2005). A general diagnostic model applied to language testing data. Research report RR-05-16. ETS: Princeton, NJ. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), pp. 287–307. https://doi.org/10.1348/000711007X193957

Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement*, *19*, 19–28.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences*. New York, NY: Waxmann.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling outcomes. In *Technical Report of the Survey of Adult Skills* (PIAAC) (pp. 406–438). Retrieved from http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf

Yerkes, R. M. (1921). Psychological examining in the United States army. *Memoirs of the National Academy of Sciences*, *15*, 1–890.

8

Impact of Technology, Digital Devices, and Test Timing on Score Comparability

Wayne J. Camara and Deborah J. Harris

Introduction to the Impact of Technology on Testing

Over the past two decades, significant technical and technological advances have been made across numerous facets of assessment development and administration; among the areas in which these advances can be found are test design, item and content generation, test delivery, and scoring (Drasgow, Luecht, & Bennett, 2006). In fact, technology has so greatly expanded the definition of computer-based testing that it may not adequately describe assessment in the 21st century. As a general term, *computer-based testing* (CBT) describes tests administered on a computer rather than on paper. Today's assessments might more appropriately be termed *digital assessment*, however, because there are many more testing options now than when CBT was introduced. Though modern assessments may comprise a multiple-choice test administered on a desktop computer at a proctored test center, they also may include testing on multiple devices, remote or no proctoring, and technology-enabled content such as simulations, scenarios, or games. Luecht (2016) notes that the millennium ushered in internet-based testing and the possibility of testing anytime and almost anywhere. These variations in the design, delivery, and input of assessments may be intended to increase access to more test takers (Winter, 2010), utilize different devices purchased by schools (Kajeet, 2018), and/or take advantage of new technology (Strain-Seymour, Way, & Dolan, 2009).

The particular way in which a test is delivered is referred to as the test administration *mode*: paper-and-pencil and CBT represent the most frequently cited standardized testing modes. Though there is really only one option for delivering a paper-based test, as noted earlier, *digital assessment* allows for the use of numerous test-delivery options such as tablets, smartphones, laptop computers, and desktop computers; these different hardware options are referred to as testing *devices*. The availability of these new device options allows for "personalization" of the testing experience to both the test taker and test user, and while this personalization may increase access for individuals, at the same time it may eliminate the standardization which has been a guiding principle for ensuring score equivalence: "In today's environment, standardized testing is more likely to be composed of a collection of testing variations than a single controlled venue" (Way, Davis, Keng, & Strain-Seymour, 2016, p. 261). Testing companies now must choose how to administer their tests by considering not only what mode is most desirable for their testing program but, should the selected mode be a digital assessment, what device(s) will be acceptable and appropriate for test delivery and what if any variations should be permitted.

Standardization and Comparability

The purpose of standardization is to maximize comparability, replicability, and interoperability, as well as to improve measurement (ANSI, 2015). In assessment, standardization implies a consistency or uniformity in nearly all facets of test development, administration, scoring and reporting, such that common percentiles, norms, and interpretations can be generated from test scores. Standardization in assessment attempts to eliminate—or at least minimize—differences in the assessment experience to facilitate comparisons of scores across time, location, conditions, scoring, and test takers. In fact, the terms "choice" and "flexibility" largely have been considered antithetical to standardized testing and are commonly viewed as a potential source of invalidity by measurement professionals. Consumers, however, value choice and flexibility in their products, services, and increasingly their assessments, creating apparent conflict between the goals of measurement and the desires of test takers. One's perspectives on the value of standardization or flexibility in the design, presentation, delivery, timing, scoring, and technology associated with assessments are likely related to the type of claims that will be made about test scores. If a testing company wants to ensure that scores from tests administered in different modes and/or across different devices have the same meaning, it is clear that claims of comparability will be important.

Comparability refers to the commonality of score meaning and interpretation across testing conditions. When comparability exists, scores can be considered interchangeable (Drasgow et al., 2006). Standard 4.4 states that test developers should document changes (or variations) to test specifications when different versions of a test are permitted, and they should describe the impact of these differences on the validity of score interpretations, score precision, and score comparability (AERA, APA, & NCME, 2014). Other Standards state that when changes to administration conditions are permitted, all such changes should be examined for construct-irrelevant variance.

Winter (2010) contrasted *score comparability* and *score interchangeability*. She explains that score interchangeability is generally reserved for equated scale scores, but comparability is increasingly thought of as a slightly less fine-grained comparison such as score pass/fail decisions or performance-level classifications. Scores may be comparable if variations in testing conditions can lead to the same score interpretation, but if those scores cannot be equated they are not interchangeable. One testing variation that impacts score-level comparability is the time allowed for the assessment and whether it is the same or varies by the mode of administration. This timing issue is the real focus of this chapter, but we felt it critical to provide the full context before focusing in on this one specific variation and its impact on comparability.

Impact of Technology on Comparability

The infusion of technology into assessment and the available variations to testing conditions based on accommodations and modifications complicate efforts to establish score comparability (Way et al., 2016). Way et al. (2016) identified two focal issues that relate to the impact of technology on comparability. First, does altering the delivery mode or device change the construct? Second, does altering the mode or device introduce construct-irrelevant variance?

Different stakeholders often will prioritize different claims for comparability. Two such claims are illustrated below:

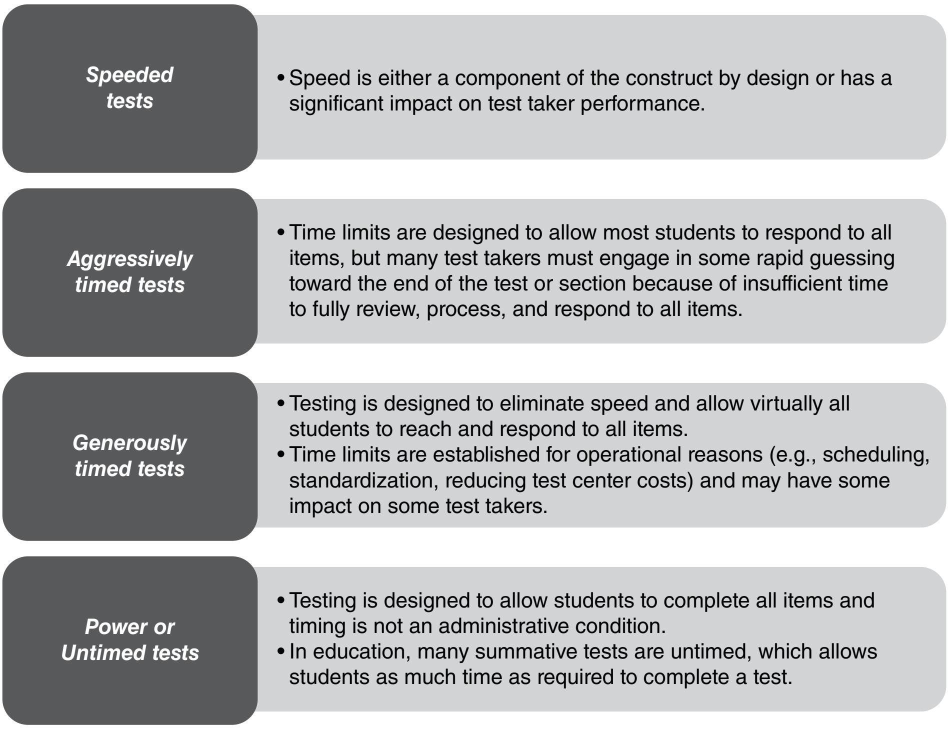
1. If a test taker took the assessment on another device (or mode), he or she would have received the same score.
2. The test taker took the assessment on the device most likely to produce the most accurate estimate of his/her true test score.

The first statement prioritizes score comparability across test takers and claims device neutrality, reflecting a more traditional view of standardization and comparability. Such a claim may be valued in contexts where scores are used for norm-referenced purposes and test takers are compared (e.g., admissions, selection, scholarships). The second statement allows for the possibility that differences exist in devices and testing conditions more generally (e.g., modes, tools, timing) and privileges individual differences as long as construct-irrelevant variance is not introduced (Dadey, Lyons, & DePascale, 2018). The second claim also prioritizes access and the ability to demonstrate maximum performance and may be valued when scores are used in criterion-referenced ways (e.g., qualification, certification, and classification/pass-fail). Such examples create a false dichotomy, however, and there are many conditions and caveats with each interpretation that must be weighed by test users before accepting either comparability claim. For example, standardization often constrains testing conditions for test takers who have disabilities or limitations (e.g., language) in ways that will impact the validity of individual scores. When standardization is privileged to the extent that exceptions or accommodations are not considered, it actually may impact score validity to a greater extent than would providing such exceptions. For example, if a state insists on testing all students on tablets when students in some districts have not had experience in using them (nor with the associated innovative items that require different input methods), it could create a barrier to optimum performance. On the other hand, if the standardized administration is altered for a test that is used to predict future performance (e.g., college success, job performance), but the same accommodations or exceptions are unlikely to be allowed, the precision of the score for predicting future performance may be in doubt. Two examples of such situations are (1) when a test taker with a disability is allowed to dictate an essay during testing but will be required to write in college, and (2) when a job applicant is provided double time on a simulation but will not be given similar extra time to perform that same task when working in the actual job. Many different features of digital assessments have the potential to impact score validity and claims about comparability if the features either vary across devices or if test takers are permitted to choose options even when a single device is employed.

Test Timing and Speededness

To this point, we have introduced the features of digital assessment that have the potential to interact with other variables and lead to performance differences. Test timing is one of the variables that can lead to different outcomes based on how the timing factors interact with the digital assessment features and it is, of course, the main focus of this book. In this section, we provide a brief conceptual background on test speededness in order to begin to explain how these factors impact score comparability. Please see other chapters in this volume (e.g., Chapters 1 and 5) for a more detailed description of the relevant topics.

Measurement research has long held that ability and speed jointly impact response behavior and response accuracy (Lohman, 1989; Thorndike, Bregman, Cobb, & Woodyard, 1926). The terms *speed test* and *power test* are used frequently to differentiate tests relative to speededness.



Speeded tests

- Speed is either a component of the construct by design or has a significant impact on test taker performance.

Aggressively timed tests

- Time limits are designed to allow most students to respond to all items, but many test takers must engage in some rapid guessing toward the end of the test or section because of insufficient time to fully review, process, and respond to all items.

Generously timed tests

- Testing is designed to eliminate speed and allow virtually all students to reach and respond to all items.
- Time limits are established for operational reasons (e.g., scheduling, standardization, reducing test center costs) and may have some impact on some test takers.

Power or Untimed tests

- Testing is designed to allow students to complete all items and timing is not an administrative condition.
- In education, many summative tests are untimed, which allows students as much time as required to complete a test.

Figure 8.1 Common terms used to describe the difference between tests with respect to the speed-power continuum.

A speed test may assume that few test takers will reach all items, and a pure power test generally allows all test takers to attempt all items (Anastasi & Urbina, 1997). If the goal of a test is to measure only ability, the time limits should not impact scores or put test takers under pressure (van der Linden, 2005). In practice, the distinction between speed and power tests is one of degree. Educational assessments are designed as power tests and are either untimed or *generously timed*, with time limits established for operational efficiency. Other assessments—including many admissions and certification and licensure tests—do not include speed within the definition of the measured construct; these tests are likely to have more aggressive time limits than educational assessments for operational reasons such as cost, security, administrative convenience, and event scheduling. The term *aggressively timed* may best represent such assessments, but empirical criteria or precise definitions have not been proposed in the measurement literature for these terms. Figure 8.1 provides descriptions of the commonly described terms that are used to describe tests on the speed-power continuum.

Score Comparability across Modes, Devices, and Items

Research on mode and device differences is still in a nascent stage, and quantitative studies that account for effects on speeded or *aggressively timed* tests are uncommon. This section provides a more detailed discussion of score comparability for large-scale assessment programs that are offered across modes or devices. We begin by reviewing many of the variations in testing conditions associated with different devices and the theoretical, logical, or empirical evidence related to score comparability. We then describe the interaction of mode and device type with testing time.

Finally, we describe research designs, analyses, and some results from studies conducted by ACT on timed tests that are administered with variations across modes, devices, and other conditions.

Comparability within and across Testing Modes

All national tests used for undergraduate, graduate, and major professional programs are offered digitally today (illustrated in Figure 8.2). The ACT and SAT are the major exceptions, offering computer administration to state and district testing programs and relying on paper for the large majority of test takers.

Research on mode effects in assessment generally focuses on paper tests versus digital assessments administered using a single digital device (which may be a laptop, Chromebook, tablet, or other mobile device). Of course, differences across digital devices (e.g., screen size, on-screen or external keyboard, mouse or other pointing device, calculator) can introduce differences when comparing within the digital mode. This section briefly reviews research on mode effects and score comparability. In establishing score comparability within mode (i.e., across two paper forms or two digital forms), test questions themselves may be the source of variability. There may be differences in speededness within the same mode; for example, students with visual impairments who require a large-print test form and answer sheets typically receive additional testing time. Resulting differences in such instances are still attributable to the items, as font size, ink-and-paper or screen contrast, method of indicating responses, and so on are typically identical across different forms. However, when looking to establish comparability *across* administration modes or devices, other factors affect timing even if the exact same items are employed. Differences may exist in item rendering and presentation or in how a test taker responds to an item. Examples of different response methods include using a mouse to select a response versus bubbling an answer sheet, needing to scroll to see all the answer options versus seeing them all together on the same screen, or working a math problem involving a graph directly in a booklet versus not being able to write on the graph on screen.

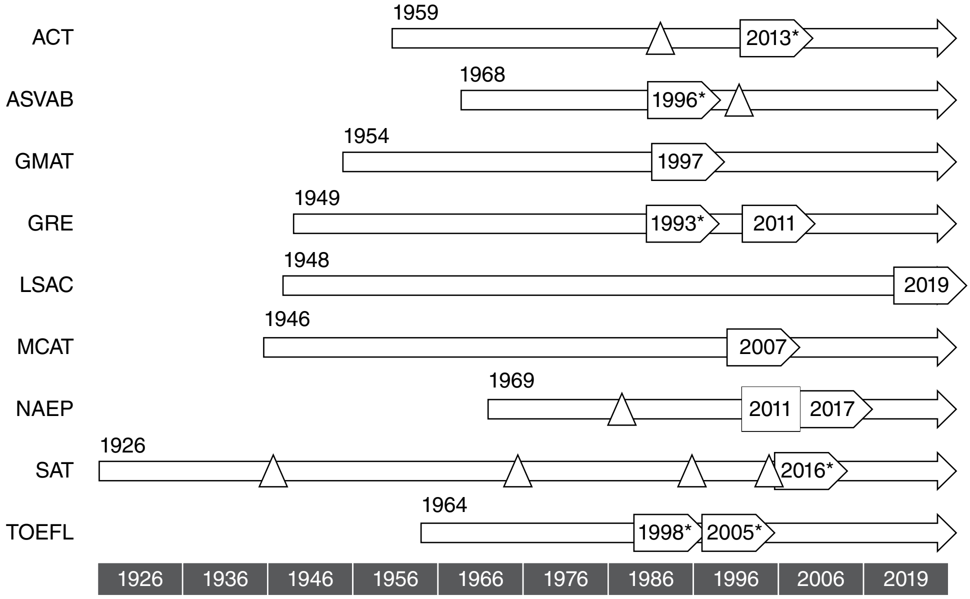


Figure 8.2 Transition of national testing programs from paper to computer. Δ indicates a major revision made to the test beyond CBT and * indicates that computer and paper were offered simultaneously.

Empirical research to date on comparability across modes (and across devices within mode) often differs based on the size of the study; larger studies have tended to focus on score comparability and smaller studies on user experience. Though the smaller studies such as cognitive labs (sometimes called *think-alouds*) may be better able to determine whether timing impacts score comparability (e.g., in observing how much faster a student responds with a keyboard as opposed to a touch screen), the small numbers of students and the observational nature of this type of research typically are factors that limit the generalizability of conclusions across students, test forms, and operational conditions. In addition, cognitive labs have been used prior to conducting large-scale studies. These allow an investigator to observe an examinee responding to an item and to measure how long it takes to scroll through a long reading passage on a tablet, for example, and hear the examinee talk about the experience and whether or not it was confusing. Cognitive labs can help to identify problems with instructions and item rendering, and they can help researchers estimate timing for different experimental conditions in larger studies.

Recent meta-analyses generally have reported small effect sizes for mode effects between paper- and computer-based tests, but findings differ across grades and content areas (Kingston, 2008; Wang, Jiao, Young, Brooks, & Olsen, 2007). Dadey et al. (2018) noted that more than 33% of effect sizes reported by Kingston (2008) are large (>1.00), with 22% reporting higher performance for the paper condition and 13% indicating higher performance on computer. One important and consistent finding is that measurement invariance is generally supported across modes, suggesting that any differences may be due to construct-irrelevant variance. A preliminary conclusion of this research was that mode effects favored computer administration in English language arts and social studies and paper administration in math (Kingston, 2008).

ACT Research on Comparability

The ACT has been offered as a linear computer-based test to a limited number of states and districts conducting school-day testing. In spring 2017, approximately 81,000 students tested on computer, including 43% who used a Chromebook (Z. Cui, personal communication, May 8, 2017). Research on mode comparability primarily has focused on differences in screen size or content displayed within the same device (e.g., laptop, desktop). In an early study comparing monitor size differences, Bridgeman, Lennon, and Jackenthal (2003) found that verbal scores were 0.25 standard deviations lower when the amount of reading content displayed on screen was reduced. When less content is displayed on the screen, more scrolling is required; this may increase the demand on short-term memory and cognitive load (Sanchez & Goolsbee, 2010) and require additional time for similar performance. This issue may be most prominent when dual passages or multiple graphics (e.g., tables, figures) are present or where items and stimuli are not displayed on the same screen. Such differences in displays have been cited as a source of construct-irrelevant variance, which could provide an advantage for paper over computer administration and larger displays over smaller displays (Bridgeman, Lennon, & Jackenthal, 2001; Chaparro, Shaikh, & Baker, 2005).

Comparability across Digital Devices

Devices in Schools

Numerous digital devices are employed in large-scale educational assessment. In K-8, tablets have become the preferred device; laptops are still preferred by high school educators (Pearson, 2015 and Deloitte, 2016). Table 8.1 illustrates students' device preference and usage in schools across grades as reported in two different studies. In the first study, tablet use was reported to be

Table 8.1 Device preference by school grade

	“Which of the following devices do you regularly use at school?” (Pearson, 2015)				“If you had to pick only one device at school, which device would it be?” (Deloitte, 2016)	
Grade	Tablet (%)	Laptop/ Chromebook (%)	Smartphone (%)	Hybrid “2 in 1” (%)	Tablet (%)	Laptop/ Chromebook (%)
K-2	78	66	53	10	53	15
3–5					36	26
6–8	69	71	66	8	30	29
9–12	49	76	82	9	25	37

highest in elementary schools, with 78% of elementary students versus 49% of high school students indicating that they regularly used a tablet; laptop and Chromebook usage was reported to be highest in high school (Pearson, 2015). A second survey of student preferences shows a similar pattern, with a stronger preference for tablets in earlier grades and a moderate preference for laptops and Chromebooks in high school (Deloitte, 2016). Tablets and laptops each come with a variety of screen sizes and operating features, not to mention the increased popularity of Chromebooks and the entry of mobile devices for instructional assessment (Deloitte, 2016). Mobile devices were identified as the number one workplace trend in the Society for Industrial Organizational Psychology’s top-ten trends in 2015 (SIOP, 2015). In preemployment testing, mobile devices are nearly synonymous with unproctored internet-based tests; their increased popularity is traced to the desire to assess talent anytime and anyplace as well as growth in mobile device ownership (Arthur, Keiser, & Doverspike, 2018).

Bring Your Own Device (BYOD) paradigms have been cited as the next biggest trend in education, but unlike preemployment testing, educational assessment has prescribed minimum requirements (e.g., screen size, operating systems, security features). A typical BYOD implementation may require students to register devices with a school to gain access to software and content. BYOD also seems ideally suited for learning assessments, which require more frequent interactions. Alternatively, they are rarely used for situations that necessitate making norm-referenced interpretations, as is the case with many summative assessments. If BYOD efforts gain even more popularity and acceptance in education, there will be increased pressure to relax existing requirements on technology, which generally prohibit small screens or mobile devices. As is evident from this discussion, technological advances increasingly will challenge concepts and assumptions of standardization in assessment. However, the result of such conflicts cannot be easily dismissed when assessments seek to support claims of score comparability. Greater flexibility in digital delivery, input, and interactions creates differences in user experiences and performance, which could increase disparities among students (Sager, 2011).

Device Comparability

Device studies represent a special case of comparability studies. There have been fewer than a dozen large quantitative studies comparing performance across tablets and computers. Overall, results generally show few consistent and significant performance differences in students’ total scores across content areas or grades examined. However, there are exceptions to the results of such studies, which primarily have been conducted on state assessments that are either untimed or very generously timed and are considered to be associated with less-motivated test takers than assessments with high stakes (Dadey et al., 2018; Davis, Kong, McBridge, & Morrison, 2016; Steedle, McBride, Johnson, & Keng, 2016).

Two major assessment consortia in the K-12 environment appear to have constrained claims of comparability to different delivery modes and not to other differences in delivery device, input device, or other conditions. The Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) allow different devices for delivery (e.g., tablets, laptops, desktops, Chromebooks), different input methods (e.g., touch screen, external keyboard, on-screen keyboard), and different technology-based features not suitable for paper testing. SBAC (2017) appears to link paper test scores to the scale used for online testing, but they do not specifically explain their methodology or equating approach (e.g., whether each form is separately linked or a standard mode adjustment is employed). In addition, no claims are made about the comparability across different devices. Neither consortium appears to have examined the impact of timing or speed across devices or modes, possibly because assessments are untimed or generously timed. PARCC notes that strict comparability across modes isn't a goal, but that score interchangeability across devices within online testing is desired (Way et al., 2016).

Research on PARCC assessments provides some of the best insights into device comparability in an operational testing program. Overall results revealed consistent evidence of comparability between testing on tablets and non-tablet devices, with similar item response theory (IRT) difficulty across devices. That being said, a number of math items on the high school assessment were flagged for device effects (Steedle et al., 2016). Results in Ohio, which included 14% of all PARCC test takers, demonstrated that students taking the exams on tablets performed significantly worse than students testing on computer; results from this state were excluded from the final analyses, leading to speculation about the robustness of comparability claims (Herold, 2016).

There is less operational research examining differences across devices, but results from mode studies likely would generalize to digital devices: lower scores would be expected for tests administered on devices with smaller screens because smaller screens require more scrolling than larger screens. This issue related to display variation could be exacerbated with a speeded test, because additional scrolling requires greater memorization as less content is displayed. Hence, research on variations across digital devices often may differ when tests are unspeeded versus speeded. Some research also suggests that this may be more pronounced with more complex and difficult items and lower performing test takers (Camara & Tang, 2017; Steedle et al., 2016). However, there is a need to disentangle performance differences across devices in speeded conditions to determine if such differences are driven primarily by item difficulty, device familiarity, item complexity (e.g., length and complexity of stimulus), or a combination of factors.

Different devices also incorporate different response options or input methods, both of which challenge comparability. Previous research suggested small differences in preference and performance favoring larger desktop keyboards over laptop keyboards (Powers & Potenza, 1996). Touchscreen keyboards and inputs now are common for tablets and mobile devices. Input precision is compromised when using a fingertip to move or select objects that are close together (Way et al., 2016), but it requires less time than more traditional input methods (e.g., mouse, keyboard; Kong, Davis, McBride, & Morrison, 2017). Mixed results have been found in this area, with some research suggesting minimal differences between on-screen and external keyboards for selected-response items but reduced length of student-produced responses on constructed-response tasks. Research also has shown somewhat less accuracy and greater fatigue with on-screen keyboards (Davis & Strain-Seymour, 2013). Chaparro, Phan, and Jardina (2013) reported that test takers typed significantly faster with an external keyboard than an on-screen keyboard but also had more errors.

Input methods also may result in different experiences, which introduces construct-irrelevant variance. For example, it is easier, or at least more familiar, for test takers to simply draw on paper rather than to use computer tools to construct a similar representation (Morelli, Mahan, & Illingworth, 2014; Sax, Lau, & Lawrence, 2011).

Some data on item latency should be available in large-scale studies of digital devices enabling comparisons, for example, of the time from when an item fully appears on screen to the time a student responds and selects the next item. However, it is generally not possible in a group setting to tell what factors affect the timing (e.g., is it scrolling, trying to use scratch paper, the method of indicating one's response to a question?).

In addition to studies conducted by the consortia, several states also have conducted comparisons of student scores across devices using operational data. Generally, such studies have been conducted on intact groups without random assignment and have not incorporated any measure of prior ability or compared groups in terms of background, demographics, familiarity, or experience using technology.

Evidence to support claims of comparability across technological devices has been identified as a critical element in the peer review of state assessment systems by the U.S. Department of Education (2015). Critical element 4.6 requires "documented adequate evidence of comparability of the meaning and interpretation of the assessment results." Dadey et al. (2018) note that documentation has been further broken down into two distinct categories:

1. Documentation that test administration hardware and software delivered across different devices are standardized across unaccommodated administrations.
2. Documentation of score comparability either through a standard comparability study (e.g., random equivalent groups or common students across different devices) or research showing that variations across devices do not alter claims regarding the interpretation of test scores.

Timing Comparability across Testing Modes and Devices

Our goal in the previous sections was to provide an introduction to some of the considerations associated with the use of technology in testing and the extent to which scores across testing modes and devices can be considered comparable when differences in digital assessment features across modes and devices exist. We now turn our attention to providing a more focused discussion of one specific aspect of digital assessment that may impact testing outcomes and the ability to compare scores across modes and devices: timing.

Mead and Drasgow (1993) conducted a meta-analysis of studies comparing computer-based and paper-and-pencil versions of 123 timed power tests and 36 speeded tests. After correcting for measurement error across 159 correlations, they reported an estimated cross-mode correlation of 0.97 for power tests and 0.72 for speeded tests, concluding that mode affected speeded tests, probably due to the additional time required to read text from screens. Similar results have been reported in other studies of speeded tests, but there have been exceptions which report no differences between speeded and power tests (Lesson, 2009).

Response Time Research

As discussed earlier, all quantitative studies of devices have been conducted in untimed or generously timed conditions, and studies reporting item response time, latency, or rapid guessing behaviors primarily have used these factors as a covariate in examining performance differences. Therefore, the interaction of these timing factors by item type, item difficulty, device familiarity, scrolling or content display, and input options is rare. Testing variations that result from different devices clearly impact the experience of test takers in terms of the presentation, display, input, and processing of information. As noted earlier, some variations will increase cognitive load and require greater recall, while other variations require different fine motor skills; all of these factors can interact with test timing. Because so much research on devices has been conducted on untimed assessments and because score differences have been the primary outcome of interest, timed testing programs cannot rely on these results as evidence of score comparability.

Response time—the difference in time (seconds) between when an item is presented and when it is responded to by a test taker—is an important outcome for studies associated with timed tests. Rapid guessing occurs when a test taker responds to an item so quickly that she/he could not have had adequate time to have read and considered the item (Schnipke & Scrams, 1997). In high-stakes testing, rapid guessing is considered a reliable indicator of speededness for computer-based tests, but for low-stakes or untimed tests it is often associated with lack of effort or motivation (see Chapter 11).

Kong et al. (2017) used a random-equivalent groups design to examine the results from 964 high school students completing a low-stakes assessment on either a tablet or a computer. Response time effort (RTE) measures the percent of items where students exhibited solution behavior and was used to measure student engagement; because the test allowed 80 minutes for 59 items, it was considered generously timed and not speeded. Overall, no significant or practical differences in RTE between devices for any of six different item types were found. That being said, there was a decrease in RTE for sections administered at the end of the test, which may be attributed to fatigue, and there was a gender effect in that males were twice as likely to be excluded from RTE analyses because of much lower engagement levels. Students testing on tablets did require approximately 1 minute longer for each section—about 3–4 seconds per item—than students testing on computer. Effect size differences favoring computers over tablets ranged from 0.29 for hot spots and 0.25 for multiple-choice items to 0.08 for drag-and-drop items. The authors concluded that "it appears that the reduced precision resulting from using the finger as the input device rather than a mouse may have created a small degree of challenge for working with on-screen objects" (2017, p. 22).

Davis et al. (2016) examined response time differences for a mix of item types administered on tablets and computers across reading, science, and math content areas and found that students testing on tablets consistently used more time to respond to items. Effect sizes were calculated based on data reported in the study and were 0.14, 0.22, and 0.18 across reading, science, and math, respectively. Ling (2016) found no main effect differences on scores or response times of eighth grade students on iPads versus computers across multiple-choice or constructed-response items.

Measuring response time is not always straightforward. ACT found that response time may be measured and reported differently across different platforms and interfaces, which can result in small but systematic differences that impact calculation of latency and response time. For example, response time can be captured in different ways, such as (a) timers (captured using the count-down clock from the client device), (b) server time (which requires all servers to be continuously synchronized to get accurate response latency), or (c) client time (captured by the client device as a recorded timestamp). Capturing examinee responses using countdown clocks (timers) on the client device avoids the pauses caused by servers, proctors, and internet connectivity and the need for network synchronization (R. Zhu, personal communication, March 21, 2018). Assessment programs seeking to understand item response latency and speededness need to understand how response times are captured across devices to determine if they are comparable and accurate.

There are numerous ways to evaluate the comparability of test scores across testing modes and devices; these range from comparisons of mean differences and correlations to IRT-based approaches. When mode or device effects are found to be significant, alternative scoring tables are usually generated to put scores from different conditions on the same scale (Way et al., 2016).

The next section reviews: (1) analyses and approaches employed by ACT to examine the comparability of scores for large-scale testing programs administered across different modes and devices, and (2) statistical adjustments that may be used to mitigate differences. Analyses are conducted at both the total test score level and the item level using classical statistics (e.g., *p*-values) and also include some IRT approaches (e.g., comparison of item parameters, differential item functioning [DIF]).

Methods Used in ACT Analyses of Testing Time

When mode comparisons are made in large-scale assessment programs, the typical scenario is that the assessment program was established in one mode or on one device and a second mode or device then was added at a later time. In such instances, there are many issues that need to be addressed. The first issue is ensuring that the construct being assessed has not changed. The user experience likely will be different across modes and some devices, but research should ensure that the construct is not altered.

The ACT has been a paper-based test since its inception in 1959, with the current battery being introduced in 1989 (ACT, 2017). Online administration of the ACT for states and districts participating in school-based testing was desired, and preliminary studies for rendering (how the items look on a computer screen) and timing were conducted. The timing study included about 3,000 examinees from 58 different schools and included multiple timings for each of the four tests in the ACT battery. The selected time limits were implemented in a large-scale mode comparability study involving more than 5,500 students from 80 high schools.

Random Equivalent Groups versus Common Students

Two basic data collection designs are used in examining timing research. The single-group design tests the same students under multiple conditions, ideally using counterbalancing and creating an environment where students are equally motivated under all conditions; the equivalent-groups design establishes multiple groups, one for each condition, that are comparable across all characteristics likely to interact with timing. In practice, it is probable that neither method will work ideally. Random assignment within classroom over a large number of classrooms is likely to come close to randomly equivalent groups, but at times that is impractical in operational settings; covariates therefore are used to try to account for any differences between intact groups (see Maxwell, O'Callaghan, & Delaney, 1993).

ACT typically has adopted a randomly equivalent groups design for studying mode differences. A participating test site sends in a roster of students, and examinees are assigned to a mode. This information is communicated back to the test centers, which then ensure that students test in the assigned mode. Motivation was a concern for some studies, in that examinees who saw no value in an assessment might perform equally poorly across mode whereas examinees who were motivated would do their best; this would introduce potential mode differences. For this reason, some of the studies involving the ACT resulted in operational (college-reportable) scores for the participants (see, e.g., Li, Yi, & Harris, 2017, for additional details). In his synthesis of 81 studies dealing with comparability across modes, Kingston (2008) states that the study design that was most informative was the one in which students were randomly assigned to different modes and had similar motivation to do their best. The random assignment tends to result in groups that are equivalent in ability, which simplifies many of the comparisons across modes.

There are logistical difficulties associated with using a random groups design, such as not being able to test intact classrooms with the same mode. However, the utility of the results offsets the difficulties in situations where the results have important implications, such as adjusting scores from different modes, devices, or testing conditions so that they can be reported and used interchangeably. Using a common student design—in which the same examinees are administered a test in each mode—has some theoretical advantages, but this approach often is problematic in practice. The same student cannot be administered the same test form twice, so multiple forms per mode are needed. In addition, motivation typically is lower on a student's second administration. Intact classrooms could be tested by mode, but counterbalancing would be better done at the individual level.

ACT has supplemented its random groups studies with additional studies; some consisted of assigning mode by classroom or school, and some were post hoc analyses of existing data collected as part of an operational administration or study in which either examinees or sites determined the testing mode. For example, in one study, a participating school had a 1:1 initiative for Chromebooks for students in the grade that was testing. After administration, that school was matched to a number of other schools in terms of available demographic characteristics that have been correlated with test performance (e.g., per pupil expenditure and the percent of students on free and reduced lunch), and the performance of students who tested on Chromebooks was compared to the performance of students testing on other devices.

Occasionally, timing studies have not been conducted at all when introducing a new mode of testing. In one program where sufficient testing time was allowed for all examinees to complete the assessment on paper, additional time was simply added to the new online assessment to provide ample time for scrolling or lack of examinee familiarity with responding on a computer. Monitoring of the examinee experience over time was done through observation and analysis of latency data. In the case of a software upgrade for WorkKeys (an assessment of career readiness used by schools, colleges, and vocational training programs), ACT staff were administered the assessments with and without the upgrade and expert judgment was used to determine that timing would not be affected (this was confirmed by subsequent monitoring; Liu, Zhu, & Gao, 2016).

Score-Level Analyses

Data analysis occurs at both the item and the test level. As is sometimes true with context position effects, some items may become easier and some items may become more difficult, but the overall impact on test scores used for decision-making could be negligible. For mode studies for WorkKeys, the ACT, and ACT Aspire, ACT typically looks at the distributions of raw test information, such as total testing time and raw scores. If the data are collected using either a random groups or single groups design, the distributions should differ only to the extent that sampling error and measurement error are factors. Often benchmark data are used. For example, if the form of a WorkKeys or ACT test being used in a special study examining different devices was previously used as an operational test form online before being retired, that form may have been seen by, say, 8,000 examinees. Two samples of 2,000 examinees each could be randomly drawn from that 8,000, and the distribution of total test time used can be compared across the two samples. The two samples from the special study across mode (or across devices) then are compared to the results from the two samples from the same mode (or same device). If the differences across mode are similar to the differences across samples within mode, it supports a negligible difference at the overall test score level. Note that this is supportive—but not conclusive—evidence. The essential point is that having a baseline of sample differences within mode can be helpful in trying to interpret observed differences.

Other score-level analyses compare the means and standard deviations of time used, examining (1) both reliability and conditional standard error of measurement of test scores, and (2) combinations of data, such as latency and number of omits and test score together. Examining both summary statistics and graphical illustrations is optimal. Test characteristic curves, distributions of theta scores, and bivariate plots of raw score by theta score and reported scale score also have been used in ACT studies of timing and mode/device comparability. It is possible to observe large differences at some places on the score scale that appear to balance out over the full sample because of how the sample is distributed. For example, there might be large differences in one direction at the high end and equally large differences in the other direction at the low end. Perhaps it is the case that high-scoring examinees in math perform much faster on one device because these examinees tend to be more familiar with that device due to the high-level math apps available for it, whereas mid-level examinees may do better on a different device due to their familiarity with it. If the sample used in the mode study has similar numbers of high- and mid-level examinees, the group-level statistics may suggest that the time needed by device is the same because the overall means of time used are the same. Looking for comparability requires more than just a cursory look at mean latency values.

Test dimensionality, raw-to-scale-score conversions, and survey results reporting on whether examinees felt they had sufficient time also have been examined at the test level. Generalizability analyses were conducted for some studies, particularly for WorkKeys and the ACT across modes. Kong et al. (2017) ran analysis of variance (ANOVA) in their study looking at response time across computers and tablets to examine mean response times by device by ethnicity. Additional analyses used to examine mode comparability—not specific to timing factors—included the Kolmogorov-Smirnov (KS) test of equivalency to look for statistically significant mode effects for all raw and scale scores, scale score correlations and effective weights, and exploratory factor analysis. In addition, irregularity reports, phone logs, test booklets presented side-by-side with online renderings of the same items, and notes from staff who observed onsite administrations have been reviewed as part of comparability analyses.

Item-Level Analyses

Item-level analyses also are important in determining if particular item characteristics tend to be more or less associated with differential timing across modes or devices. These factors could include position in test, length of passages for reading item sets, complexity of graphical material, item response required (production vs. selection), item content, and so on. For example, Kong et al. (2017) conducted *t*-tests looking for differences in timing by item types across computers and tablets. They found some slight timing differences, with some of the item types taking longer on the tablet than the computer.

Individual item latencies, classical item statistics, IRT statistics, and item characteristic curves, both empirical and generated through IRT, as well as item option analyses and omit/not reached rates should be compared in mode studies. Finally, DIF analyses often are conducted across modes and/or devices for individual items.

Resolving Differences via "Equating" Methodology

If differences in timing are found across modes and/or across devices, the next logical step is determining how to deal with them. Some large-scale testing programs (e.g., PARCC, SBAC, certification tests) allow multiple devices, and it is not practical to maintain separate time limits for all possible devices with which an examinee could test. It might be practical to have separate time limits for some conditions, such as paper versus online, but accounting for differences across operating systems, screen size, keyboard versus touch responding, and so on is probably not realistic. However, variations between different devices may be larger than those between paper and online test versions given different types of navigation, screen size, rendering options, etc. This often is one reason for having certain supported features, such as requiring a minimum screen size, or allowing testing on desktops and laptops but not tablets.

The best way to address different timing needs is not always obvious. For situations where time is not a critical component of standardization or achievement, it may be simplest to go to untimed assessments, with each examinee using whatever time is necessary for the particular configuration on which he or she is testing. When speed is not a factor in testing, slight differences will logically have less impact on performance. Another option is to consider differential test timing when linking scores across modes. A third option that may be less attractive to an operational testing program is to indicate under what conditions the score was produced—in essence, having two (or more) different reported score scales, such as one for online and one for paper. However, programs that have implemented this solution (e.g., TOEFL) have emphasized that scores cannot be used interchangeably.

Perhaps the best option is to keep different modes and devices in mind when developing assessments, trying to create a test experience that does not seem to advantage or disadvantage certain modes/devices over others. Depending on what constructs one is assessing this may not always be possible, but to the extent that it *is* possible, it is advantageous from the standpoint of comparability across modes/devices and also may reduce some irrelevant noise in the assessment.

Adjusting scores for reporting—for example, by establishing different time limits and/or adjusting scores through equating methodology—can also control for differences between modes. Table 8.2 shows raw score results and adjusted scale score results for an identical ACT form administered in two modes. The reported scale scores are treated as comparable. Whatever the selected approach is for a particular program, perhaps the most important step is to continually monitor the results over time. To ensure that comparability is maintained, issues such as examinee familiarity with devices, improved test interfaces, practice tests that mimic the operational experience, release of new devices and operating systems, use of new item types, and other factors may continue to have an impact on the timing for an assessment subsequent to the initial research. ACT detected such differential timing changes in special studies related to the ACT test battery. In the initial timing study, results suggested that the paper time limits should be extended by an additional 5 minutes for the ACT Reading and Science tests. However, it also was noted that the sample was relatively small—and likely unmotivated—and that many of the students had not gone through the online tutorial prior to taking the assessment. In a subsequent study with a more motivated sample, there had been some minor improvements to the interface and students reported they had made use of the tutorial and a sample test prior to testing. These later results indicated that the additional 5 minutes were not needed, and this timing change was subsequently reversed. Ongoing monitoring, including review of item and test latency as well as item and overall test performance, continues to be conducted to be sure that the results are still supportive of the original timing conditions.

Table 8.2 Raw and scale score means and SDs for an identical test form in two modes

			Raw Score		Scale Score	
Mode	Test	*N*	Mean	SD	Mean	SD
Online	English	1,092	43.62	14.10	19.79	6.06
	Mathematics	1,092	30.02	11.76	20.65	5.18
	Reading	1,092	23.28	7.59	20.93	6.09
	Science	1,092	20.73	7.43	20.82	5.06
Paper	English	1,056	41.26	14.43	19.79	6.03
	Mathematics	1,056	29.74	11.78	20.58	5.16
	Reading	1,056	22.00	7.57	20.91	6.07
	Science	1,056	20.72	7.20	20.80	4.96

Table 8.2 provides data combined from Li et al. (2017, Tables 14 and 21).

A Final Consideration: Examinee Experience

In addition to possible impacts on the construct a test is measuring and on the reported scores, the examinee experience also may differ across modes and devices and timing may be an important component of how an examinee perceives the experience. Regardless of whether the examinee's perception is accurate, some examinees believe that they would perform better under certain conditions, such as testing on paper over testing on a tablet.

Kingston (2008) found that when the question of preference was asked after students were administered an online assessment, the majority of students preferred testing on the computer to paper. Davis et al. (2016) reported in their survey of high school students that, while paper was popular, there was a suggestion of preference for the devices the examinee was familiar with and that exposure to tablets in the study increased students' interest in testing on tablets. Younger test takers also are more likely to prefer tablets, but tablets present some unique challenges for assessments with extended writing or scrolling (Ling, 2016). It is not possible to tease out what role, if any, timing-related concerns played in overall examinee preferences. As examinees become more familiar with a device, as more schools teach and administer classroom tests on devices, and as students become more familiar (and hence probably faster) with different types of items and methods of responding, testing preferences will no doubt evolve.

Discussion

Technology has enabled advances in assessment that may allow for greater personalization in task design, test scheduling, choice of tools and features, test delivery, response type, scoring, and the types of information reported. Such personalization is viewed as an advantage for diagnostic assessments, which are conducted frequently and in real time because it enables research on the efficacy of various instructional techniques and solutions work for different types of learners (West, 2011). However, personalization of summative assessments that are employed for different purposes (e.g., admissions, placement, and accountability) introduces variations which challenge assumptions about score comparability.

Variations across examinee experience, timing, input type, response device, screen size, interface, resolution, item type, or other test specifications can introduce device-engineered construct-irrelevant variance in standardized assessments (Arthur et al., 2018). Variations in essential features of the assessment may remove standardization and largely weaken comparisons across test takers. Because schools often invest in different devices and technologies, the K-12 environment increasingly demands that vendors deliver assessments across a wide range of devices. Not only can this practice threaten the standardization process and weaken claims of score comparability, but different testing conditions and technology may risk procedural fairness across students and schools. In K-12 tests, where most students are likely to have experience with the devices used regularly in their classroom and may not have access to those particular devices outside of school, familiarity with the testing device is an especially important consideration. Adequate time should be afforded students to prepare for the test by completing a practice test and responding to test items during classroom time with the approved device. However, for higher-stakes tests with time limits, such as licensure tests, differential timing issues such as item refresh speed across configurations or screen size and the impact of scrolling on response time are of special concern. As assessment conditions change, ongoing monitoring of timing issues needs to be undertaken, particularly in high-stakes situations with multiple modes or devices available.

For situations in which individuals can select among a range of devices, one concern is that some individuals may choose a device that does not optimize their performance. When we look at comparability across paper and online modes and across digital devices, there are likely to be some subgroup differences, though perhaps not the subgroups we typically are used to considering. For example, an affluent school or a poor school with a grant may both have a 1:1 tablet or laptop ratio for students which other schools may not have. Those students who use a particular device every day may perform better when testing on that device than other students who do not use that device on a daily basis. Testing on a familiar platform may decrease the time needed to read and respond to test items for those students, as they are used to where keys are, the screen size, or the method of responding (e.g., touch screen, keyboard, stylus).

Comparability across devices, technology, and conditions is most likely to result when it is engineered into the assessment design process (such as with evidence-centered design). However, there are limits to technological changes which can be anticipated and the flexibility that can be accommodated within open architecture and open source systems (Huff, Steinberg, & Matts, 2010). Test providers try to balance allowing access to online assessments with ensuring a comparable experience for all testers. Requiring all students to purchase or all schools to supply a consistent device configuration of hardware and software is not practical, particularly if the students and schools already have made a commitment to different configurations for classroom learning. However, having no hardware or software requirements allows students to have different experiences, some of which have timing implications, such as screen size and the amount of scrolling required to read a stimulus associated with an item. Test providers can develop or adopt applications that can run on different devices but that simulate a unified test experience once the assessment starts, such as presenting items with a fixed display size to standardize the amount of text visible across all devices. In addition, ensuring that there is widely available access to practice items and directions regarding how the test delivery system works for all approved devices, such as illustrating how to navigate through the test with practice items, may at least provide opportunities for students to become familiar with the device configurations on which they subsequently will be testing. This will help to ensure that their time during the operational assessment is not spent trying to figure out, for example, how to flag an item for review. Test providers also need to conduct the necessary research studies before approving different devices, operating systems, or configurations; this research is especially important when tests are used for high-stakes purposes such as admissions, licensing, and certification. For many high-stakes tests which have time limits (e.g., licensure tests), test timing is an important issue, and timing considerations—such as differences in item refresh speed across configurations or differences in screen size and therefore the required amount of scrolling—are of special concern and should be evaluated as part of any comparability studies.

Markets and consumers may expect assessments to be accessible across any and all devices and include the latest features and tools. However, it is the test developers' responsibility to ensure that research has been conducted to minimize construct-irrelevant differences and ensure that scores are comparable, valid, and fair across test takers. When multiple devices, tools, and features are permitted in an assessment program, the tacit assumption is that rigorous research has been conducted to ensure that such differences do not impact test performance. If that threshold cannot be met, it is better to proceed more slowly and more cautiously when scores are used for important decisions.

Technology changes at a rapid pace, with frequent releases and updates for operating systems, tools, applications, and response opportunities emerging throughout the typical life span of a major assessment program. As technology changes, there is increased demand for assessments to adopt and apply innovations that may improve the assessment process in some ways and yet disrupt both score comparability and longitudinal trend data. Research on device differences has largely been based on simple research designs and performance differences. Though differences in item latency and other metrics have been found across devices, the research has focused on tests that are unspeeded or generously timed. In this chapter, we have identified a number of challenges associated with attempting to extend research results to tests that are speeded or have more aggressive time requirements, and we caution against making assumptions of score comparability when the empirical data are so limited.

References

ACT. (2017). *ACT assessment technical manual.* Iowa City, IA: ACT.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American National Standards Institute (ANSI). (2015, September). *Standardization of units of measurement*. Retrieved from https://blog.ansi.org/2015/09/standardization-of-units-of-measurement/#gref

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Arthur, W., Keiser, N. L., & Doverspike, D. (2018). An information-processing-based conceptual framework of the effects of unproctored internet-based testing devices on scores on employment-related assessments and tests. *Human Performance*, *30*, 1–22.

Bridgeman, B. (2020). Relationship between testing time and testing outcomes. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 59–72). Abingdon: Routledge.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution and display rate on computer-based test performance* (ETS-RR-01-23). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*(3), 191–205.

Camara, W. J., & Tang, W. (2017, June). *Innovative items: When technology enhanced items cannot be used on paper forms*. Paper presented at the National Conference on Student Assessment, San Antonio, TX.

Chaparro, B. S., Phan, M. H., & Jardina, J. R. (2013). Usability and performance tablet keyboards: Microsoft surface vs Apple iPad. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*, 1328–1332.

Chaparro, B.S., Shaikh, A. D., & Baker, J. R. (2005). Reading online text with a poor layout: Is performance worse? *Usability News*, *7*(1), 1–4.

Dadey, N., Lyons, S., & DePascale, C. (2018). The comparability of scores from different devices: A literature review and synthesis with recommendations for practice. *Applied Measurement in Education*, *31*(1), 30–50.

Davis, L. L., Kong, X., McBridge, Y., & Morrison, K. (2016). Device comparability of tables and computers for assessment purposes. *Applied Measurement in Education*, *30*(1), 16–26.

Davis, L. L., & Strain-Seymour, E. (2013). *Keyboard interactions for tablet assessments. Pearson Education*. Retrieved on March 1, 2018 from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/schools/CRPAF34829_CA-Flyer_Keyboard_final_web.pdf

Deloitte. (2016). *2016 Digital education survey*. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-tmt-digital-education-survey.pdf

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology in testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: American Council on Education/Praeger.

Herold, B. (2016, July). Common core test takers experience digital device danger. Government and Technology. Retrieved from http://www.govtech.com/education/Common-Core-Test-Takers-Experience-Digital-Device-Danger.html

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large scale assessments. *Applied Measurement in Education*, *23*, 310–324.

Jurich, D. P. (2020). A history of test speededness: Tracing the evolution of theory and practice. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 1–18). Abingdon: Routledge.

Kajeet. (2018, September). *What devices are used in the K-12 classroom*? Retrieved from https://www.kajeet.net/extracurricular/what-devices-are-used-in-the-k-12-classroom

Kingston, N. M. (2008). *Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis*. Retrieved from http://www.tandfonline.com/doi/full/10.1080/08957340802558326?scroll=top&needAccess=true

Kong, X., Davis, L. L., McBride, Y., & Morrison, K. (2017). Response time differences between computers and tablets. *Applied Measurement in Education*, *31*(1), 17–29.

Lesson, H. V. (2009). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, *6*(1), 1–24.

Li, D., Yi, Q., & Harris, D. J. (2017). Evidence for paper and online ACT comparability: Spring 2014 and 2015 mode comparability studies. ACT Research Report Series 2017-1. Iowa City, IA: ACT. Retrieved on January 17, 2018 from http://www.act.org/content/dam/act/unsecured/documents/R1616-paper-and-online-testing-2017-04.pdf

Ling, G. (2016). Does it matter whether one takes a test on an iPad or a desktop computer? *International Journal of Testing*, *16*, 352–377.

Liu, C., Zhu, R., & Gao, X. (2016). *WorkKeys 2.0 timing study*. Iowa City, IA: ACT.

Lohman, D.F. (1989). Individual differences in errors and latencies on cognitive tasks. *Learning and Individual Differences*, *1*(2), 179–202.

Luecht, R. M. (2016). Computer-based test delivery models, data and operational implementation issues. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 179–205). New York, NY: Routledge.

Maxwell, S. E., O'Callaghan, M. F., & Delaney, H. D. (1993). Analysis of covariance. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 63–104). New York, NY: Marcel Dekker.

Mead, A. D., & Drasgow, G. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449–458.

Morelli, N. A., Mahan, R. P., & Illingworth, A. J. (2014). Establishing the measurement equivalence of online selection assessments delivered on mobile versus nonmobile devices. *International Journal of Selection and Assessment*, *22*(2), 124–138.

Pearson. (2015). *Student mobile device survey 2015*. Retrieved from https://www.pearsoned.com/wp-content/uploads/2015-Pearson-Student-Mobile-Device-Survey-Grades-4-12.pdf

Powers, D. E., & Potenza, M. T. (1996). *Comparability of testing using laptop and desktop computers* (ETS Report No. RR-96-15). Princeton, NJ: Educational Testing Service.

Sager, G. (2011, October) *BYOD-Worst idea of the 21st century*. Retrieved from http://stager.tv/blog/?p_2397

Sanchez, C. A., & Goolsbee, J. Z. (2010). Character size and reading to remember from small displays. *Computers & Education*, *55*(3), 1056–1062.

Sax, C., Lau, H., & Lawrence, E. (2011, February). *Liquid keyboard: An ergonomic adaptive QWERTY keyboard for touch-screens and surfaces*. In/CDS 2011. The Fifth International Conference on Digital Society (pp. 117–122).

SBAC. (2017). *Smarter balanced guide to technical readiness*. Retrieved on February 9 from https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-tmt-digital-education-survey.pdf

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-stage mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Society for Industrial-Organizational Psychology (SIOP). (2015). Top ten workplace trends: 2015. Retrieved from https://www.siop.org/Research-Publications/Items-of-Interest/ArtMID/19366/ArticleID/1725/New-Year-New-Workplace-SIOP-Announces-Top-10-Workplace-Trends-for-2015

Steedle, J., McBride, M., Johnson, M., & Keng, L. (2016). *Spring 2015 digital devices comparability research study*. Washington, DC: PARCC.

Strain-Seymour, E., Way, W. D., & Dolan, R. P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. New York, NY: Pearson Education.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. New York, NY: Teachers College Bureau of Publications.

U.S. Department of Education. (2015). *Peer review of state assessment systems, non-regulatory guidance for states*. September 25, 2015. Washington, DC: USED. Retrieved from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf

van der Linden, W. J. (2005). *Statistics for social and behavioral sciences. Linear models for optimal test design*. New York, NY: Springer Science + Business Media.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olsen, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*(2), 219–238.

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes and devices. In F. Drasgow (Ed.), *Technology in testing: Measurement issues* (260–284). New York, NY: Routledge.

West, D. (2011). *Using technology to personalize learning and assess students in real-time*. Washington, DC: Center for Technology Innovation, Brookings Institute.

Winter, P. C. (2010). *Evaluating the comparability of scores from achievement test variations*. Washington, DC: Council of Chief State School Officers (CCSSO).

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150–164). Abingdon: Routledge.

9

Using Response Time for Measuring Cognitive Ability Illustrated with Medical Diagnostic Reasoning Tasks

Patrick Kyllonen and Rick Thomas

Psychometrics and cognitive psychology are concerned with measuring human cognition. But at least since Cronbach's (1957) American Psychological Association address, it has been acknowledged that, though they are two associated disciplines, they have been distinct in their history, methods, and focus. Cronbach argued for "a true federation of the disciplines"; we believe that the analysis and modeling of response time on cognitive tests has served as a bridge between disciplines and has enabled such a true federation. In this chapter, we justify that belief by reviewing the role of response time measurement in cognitive psychology and recent attempts to provide a psychometric foundation for modeling response time. We illustrate some of these ideas in the context of a review of a cognitive architecture for medical diagnostic reasoning, which incorporates much of what we know about the cognitive processes associated with medical diagnostic reasoning. We also suggest a program of research to expand our understanding of diagnostic reasoning using cognitive psychological and psychometric methods designed to achieve a process-level understanding of reasoning task performance.

Foundations: Response Time in Testing and Cognitive Modeling

Response Time and Error Rates

For over 50 years, the analysis of response time has been a key methodology for understanding cognitive processes. For example, response time analyses have been the basis for identifying memory priming (Meyer & Schvaneveldt, 1971), mental rotation (Shepard & Metzler, 1971), the fan effect in memory retrieval (Anderson, 1974), and the verbal-articulatory loop in working memory (Baddeley & Hitch, 1974); differentiating automatic or Type 1 processing from controlled or Type 2 processing (Kahneman, 2011; Schneider & Shiffrin, 1977); and many other cognitive phenomena (Anderson, 2014; Neisser, 1967/2014). Response time results also have fed directly into the creation of cognitive architectures: programming environments enabling simulations of the human information processing system that embody the principles identified through the kinds of experiments referenced above. Cognitive architectures are essentially theories of cognition; some examples are EPIC (Executive-Process/Interactive Control;

Meyer & Kieras, 1997), ACT-R (Adaptive Control of Thought—Rational; Anderson, 2007), 4CAPS (Cortical Capacity-Constrained Concurrent Activation-based Production System; Just & Varma, 2007), and Soar (State, operator, and result; Newell, 1994). Later in this chapter, we present HyGene (Hypothesis Generation; Thomas, Dougherty, Sprenger, & Harbison, 2008), which is a cognitive architecture focused on medical diagnostic reasoning.

Response time has not been the only variable studied in cognitive psychology experiments. Error rates, for example, have been the predominant focus of study in problem-solving, decision-making, category learning, and memory experiments. As in testing, there is a distinction between speed and power in cognitive psychology (Carroll, 1993). For some cognitive tasks such as working memory tasks or decision-making tasks, most of the performance variability is in error rates; this performance variability can be due either to individual differences or to task manipulations. For example, working memory or cognitive load refers to the amount a respondent is required to remember to succeed at the task. High-load tasks may ask the respondent to remember quite a bit (e.g., a busy air traffic control tower), whereas low-load tasks (e.g., a quiet air traffic control tower) put less burden on the respondent. Performance variability is also studied with respect to individual differences such that a high-working-memory-capacity individual is one who can cope with high-load tasks more effectively than can a low-capacity individual. In psychometrics, this distinction between variability due to task manipulations versus variability due to individual differences is expressed as the distinction between item difficulty and person ability parameters.

In other cognitive tasks (e.g., memory retrieval, lexical decision), most of the variability is in response time. In much of the classical work in cognitive psychology, response time has been a featured variable, and error rates were intentionally kept low by reducing task difficulty to increase the interpretability of response time (Diependaele, Brysbaert, & Neri, 2012). When error rates are very low, performance variation due to error rates is low and most of the performance variation is reflected in response time. When error rates are high due to items being more difficult, interpretation of response time is more problematic. Some portion of response time will reflect processing time, some portion will reflect how long an individual persists before abandoning solution attempts, and some portion will be an unknown mixture of the two. For example, fast guesses and endless ruminations represent opposite ends of the persistence continuum.

In some research, error rates have been treated as interchangeable with response time, with experimental manipulations affecting each variable in qualitatively the same way through the same mechanisms. In memory research, for example, different items from memory can have different levels of activation due to how recently or frequently the item has been accessed. Activation level may be treated as affecting both the rate of processing and the probability of processing (Lebiere, Anderson, & Reder, 1994).

Speed-Accuracy Tradeoff

The phenomenon of the *speed-accuracy tradeoff* is well established and is critical in understanding the place of time limits in assessment; respondents can make fewer errors on a task by taking more time (Wickelgren, 1977). Approaches for addressing the speed-accuracy tradeoff tend to be oriented toward achieving a uniform tradeoff between participants, through various methods (Heitz, 2014; Wickelgren, 1977). One method is to give instructions, such as "work as quickly as possible without making errors." That has proven not to be very effective because respondents interpret such instructions differently. Another strategy is to impose a fairly strict deadline and then to model performance using information about whether a correct response is made within that deadline (Wright & Dennis, 1999), within a time band (minimum and maximum time window), or after a response signal (respond only when given a signal to do so).

Imposing deadlines in this way limits variability due to choice processes concerning how much time to spend on an item and thus minimizes speed-accuracy tradeoff variance. In some cases, this has been done by manipulating stimulus presentation speed (e.g., Deary & Stough, 1996; Schneider & Shiffrin, 1977; Yellott, 1971), but the response itself is not time limited. Deadline manipulations have also been used to differentiate strategy use in situations such as simple addition tasks; in these situations, the use of direct memory retrieval (which is done quickly) is compared to the use of a procedural strategy (which takes more time; Campbell & Austin, 2002).

An alternative to deadlines is to partition (or bin) responses into fixed response time intervals (e.g., shorter than 1 second, between 1 and 1.5 seconds, between 1.5 and 2 seconds, etc.) after data are collected and then to model accuracy as a function of time taken. This takes advantage of respondents' speed-accuracy tradeoff fluctuations that occur naturally within the testing session. This method was experimented with by Lohman (1990) and Evans and Wright (1993). More recently, Goldhammer, Steinwascher, Kroehne, and Naumann (2017) developed an approach capitalizing on the naturally occurring variability within persons, modeling the speed-accuracy tradeoff across deadline conditions and the conditional accuracy function within conditions using a general modeling framework (the generalized mixed modeling approach [GLMM]).

Another strategy is to manipulate incentives for responding quickly versus more carefully by varying payoffs for fast-correct, fast-incorrect, slow-correct, or slow-incorrect responding. This can be done through the scoring rule. For example, Maris and van der Maas (2012) developed a scoring rule in which maximum points were obtained with fast-correct answers and maximum points were taken away for fast-incorrect answers. At the deadline, zero points are gained or lost. Between onset and the deadline, there is a steadily diminishing difference between points gained (with a correct response) and points lost (with an incorrect response). Van Rijn and Ali (2018) extended the framework to allow varying deadlines for items based on item discrimination.

It should be noted that although speed-accuracy tradeoff has long been accepted as an established phenomenon, it has its limits. Speed-accuracy tradeoff applies nicely to *speeded* tasks, that is, those tasks for which everyone in the target population could provide the correct response if they took the time to do so. For so-called *power* tasks, ones in which people may vary in the probability of getting the correct answer if given unlimited time, speed-accuracy tradeoff might not always be a useful model. Consider the task of remembering one's high school classmates. Williams and Hollan (1981) showed that even after 9 hours (!), respondents continued to recall additional names, but they also increasingly made errors (fabrications), recalling the names of people who were not in their high school class. Speed-accuracy tradeoff by itself does not capture this phenomenon.

Cognitive Processes

Cognitive psychology historically has been concerned with the processes underlying responses on cognitive tasks. But traditional abilities testing and individual differences research also have been concerned with processes. For example, Carroll (1993), similar to Thurstone (1938), interprets the results of factor analyses using process language to describe the common and distinctive features of tests (e.g., "inductive tasks are those that require subjects to inspect a class of stimulus materials [nearly always with more than one instance] and infer [induce, educe] a common characteristic underlying these materials—a concept, a class membership, a rule, a process, a trend, or a casual relation, for example," p. 238). But the methodology used in cognitive psychology and traditional abilities research differs. In abilities research, processes are inferred by subjectively identifying similarities and dissimilarities of task demands from different cognitive tests based on factor analysis results (as illustrated by the quote from Carroll). There also may be a grain-sized difference in that abilities testing involves the analysis of sum scores over sets of items from many different tests administered to the same group of test takers. In contrast, the prototypical cognitive psychology experiment manipulates features of a particular test to enable inferences about the mental processes invoked by test variants or design conditions. For example, Shepard and Metzler (1971) found that the time it took to recognize that two perspective drawings represented the same three-dimensional shape increased linearly with the angular difference between them, based on response times to many different shape pairs varying in their angular disparity. From this finding, it was inferred that participants mentally rotated the two figures into congruence.

Beginning in the 1970s, there was interest in incorporating cognitive psychology tasks and methods into abilities measurement following Neisser's (1967/2014) "rallying cry for the cognitive revolution" (borrowing from Hyman's introduction to the classic edition). Cronbach's "true federation between disciplines" therefore was realized in several ways. A *cognitive correlates* approach involved administering tasks developed in cognitive psychology laboratories to test takers, treating various task parameters and other outcome measures as test scores that could be analyzed in the same way traditional mental abilities measures were analyzed (Fairbank, Tirre, & Anderson, 1991; Hunt, Lunneborg, & Lewis, 1975). These task parameters and other measures were then correlated with traditional ability measures such as vocabulary and reading comprehension tests of verbal ability. This cognitive correlates approach was designed to shed light on what it meant to be high on a traditional abilities factor from the standpoint of an information processing perspective; that is, to borrow Hunt et al.'s (1975) title, "what does it mean to be high verbal?" A variation of this method involved simply administering a battery of cognitive tasks developed in the experimental laboratory to a group of test takers to identify the factor structure of such tasks or the parameters from such tasks. The idea was that individual differences methods could serve as a "crucible of theory construction" (Underwood, 1975) by supporting or failing to support a process interpretation of performance. For example, Underwood, Boruch, and Malmi (1978) administered a battery of memory tests to participants to determine whether traditional distinctions made in the experimental study of memory would result in comparable individual differences factors (mostly they did not).

A *cognitive components* (or *componential analysis)* approach was another method designed to gain a process interpretation of a traditional ability test score. Cognitive components used the method of partial tasks. This method starts with a traditional ability item type, such as a verbal analogy (A:B as C:?), from which partial tasks are produced. As an example of a partial task, participants would be shown the first part of the item (A:B) and allowed time to study that pair (i.e., encode A and B and determine a rule that lined them). After this, the full item was presented (A:B as C:?) , and the time to complete the full item was recorded. The time it took to complete the full task without seeing the first part minus the time it took to solve the full item after being allowed to study the first part provided an estimate of the time it took to process the stages other than those associated with the first part (i.e., encoding C, mapping the rule to C to infer D, and responding). Creating various partial tasks with more and less information being revealed before the full task enabled the isolation of various specific processing components, such as encoding or mapping (Sternberg, 1977).

The unification of cognitive psychology and individual differences psychology has led to several changes in testing that are still being felt. A wide variety of cognitive testing item types has been introduced and used operationally in personnel selection contexts, particularly in the military (Baddeley, 1968; Irvine, 2014; Kyllonen & Christal, 1990). Similarly, a wide variety of item types drawn from cognitive psychology investigations has been introduced into testing (Scalise, 2009), including more complex games and simulations (Mislevy et al., 2014) and complex item types used in the Program for International Student Assessment (PISA), such as collaborative problem-solving (Organisation for Economic Cooperation and Development [OECD], 2017). A resurgence of interest in response time and efforts to consider how it can be incorporated in cognitive abilities testing (De Boeck & Jeon, 2019; Kyllonen & Zu, 2016) has led to several special issues on the topic, including the *British Journal of Mathematical and Statistical Psychology* (Molenaar & Visser, 2017) and the *Journal of Intelligence* (Wilhelm, 2016). Perhaps most importantly, cognitive psychology has inspired the promise that a process-level understanding of performance on ability and achievement tests will create opportunities for improved diagnosis and strategies for improving the skills and knowledge the test is designed to measure.

Cognitive Psychology and Item-Response Theory

In addition to the unification of cognitive psychology and testing there is an evolving link between cognitive psychology and psychometrics, or more specifically, item response theory (IRT). That unification centers around three categories: (a) hierarchical IRT models of response time and ability, (b) the use of IRT to model response choice based on the diffusion model from cognitive psychology, and (c) linear logistic test models (LLTM). More recently, cognitive diagnostic models (CDM) that model response stages within cognitive tasks have also been used. The next sections will describe these three categories in greater detail.

Hierarchical IRT Models

The hierarchical IRT model (van der Linden, 2007) actually reflects the combination of two models. One is a regular IRT model for responses, in which a response is correct (or not) due to ability level (theta) and item parameters, particularly item difficulty, item discrimination, and guessing (1, 2, and 3 parameter logistic model, respectively). Any of these models can be used for the hierarchical IRT model. Jointly, a second model is fit to the response time data in which log response time is modeled as a function of the examinee's speed and the same 1, 2, or 3 item parameters (or, it could be different parameters). At the second level of the hierarchical model, ability and speed are correlated. This model can be used to estimate ability and speed. In general, it is typically found that ability is more precisely estimated by using speed as collateral information (van der Linden, Klein Entink, & Fox, 2010) as long as there is some correlation between ability and speed. It should be pointed out that the speed and ability parameters from the hierarchical model are not the same as response time and accuracy. Speed reflects both time spent on an item and item characteristics, whereas response time only reflects time spent on an item. It is generally expected that test takers will spend more time on harder (i.e., more *time intense*, to use van der Linden's (2007) terminology) items.

The hierarchical model accommodates speed-accuracy tradeoff but assumes that a person has a fixed ability and a fixed speed that is constant throughout the test. This is referred to as the *stationarity assumption* (van der Linden, 2007); it is as if a person chooses a position on his or her own personal speed-accuracy tradeoff curve and sticks with that position throughout the duration of the test. Ability and speed can be correlated, but they are the same throughout the test. Minor deviations from the stationarity assumption can be detected with residual analyses.

The hierarchical model also assumes conditional independence between responses and response times given one's ability and one's speed; in other words, the observed correlations between time and accuracy are due only to the correlation between latent speed and latent ability. This is similar to the conditional (local) independence assumption in IRT: the correlation between *responses* on any two items is due only to both items reflecting the same ability factor;

likewise, the correlation between *response times* on any two items is due only to both items reflecting the same speed factor. Van der Linden and Glas (2010) discuss tests of the conditional independence assumption.

There are a number of reasons to question the assumption of a fixed ability and a fixed speed throughout the test, or the assumption of conditional independence between time and accuracy; these two assumptions go hand in hand. In cognitive psychology, for example, fast guesses (Yellott, 1971) and the phenomenon of post-error slowing have been demonstrated repeatedly (Ruitenberg, Abrahamse, de Kleine, & Verwey, 2014). There are various explanations for post-error slowing, such as participants adjusting their speed-accuracy tradeoff to emphasize accuracy on subsequent trials, or participants getting distracted by the error causing attention to be diverted from the cognitive task on subsequent trials. Other phenomena include learning and warm-up effects (in which examinees start slowly then speed up), strategy shifts in the middle of the test, fatigue, and end-of-test speedup as the test deadline approaches. All of these potential causes can lead to a situation in which time and ability might not be fixed across a test, but instead they could change based on events during the test.

Another reason to question the assumption of a fixed ability and a fixed speed is the finding that even when controlling for item difficulty and item time intensity it is still often found that test takers take longer on items they get wrong. More difficult items typically take longer to respond to, but the time intensity parameter in the hierarchical model accounts for that—item difficulty and item time intensity tend to be highly related (Goldhammer, Naumann, & Grieff, 2015; Klein Entink, 2009). Goldhammer et al. (2015) found that after controlling for item effects (difficulty and time intensity), longer response times were associated with getting the item wrong. This seemed to be particularly true for relatively easy items. Several studies (Bolsinova, de Boeck, & Tijmstra, 2017; Liao, 2018; Partchev & de Boeck, 2012) also found that responding slower on easy items was associated with getting the item wrong but that responding slower on hard items was associated with getting the item right (tests included a credentialing exam, PISA 2012 mathematics, and progressive matrices). Goldhammer et al. (2015) invoke a rumination explanation (i.e., cognitive processing that does not get closer to a solution), but they do not suggest why this would happen particularly on easy items nor does it explain the findings for hard items.

Part of the problem for explaining these findings may be that speed estimated in the hierarchical model is often incorrectly thought of as a processing speed parameter. The speed parameter includes both productive processing time and unproductive rumination time, and it also is governed by a termination decision for when to quit the item if the answer is not found. The *speed* parameter in these models therefore might better be described as a *tempo* parameter: the average time one chooses to spend working on items (conditioned on item characteristics). Similarly, the time intensity parameter cannot be interpreted as the processing time needed to solve the item, as it is measured as the time people tend to spend on the item whether or not they get the item correct.

Bolsinova et al. (2017) and Liao (2018) suggested ways to model the dependence between speed and ability; one example is to use response time residuals to improve the model for response accuracy. Another approach suggested by Molenaar (2018) was to use a finite mixture modeling approach to identify two response classes (fast responding and slow responding) which vary item to item. Partchev and de Boeck (2012) present a similar idea based on a categorical IRT approach called IRTree. Bolsinova and Molenaar (2018) identify several approaches for exploring conditional dependencies between time and accuracy and suggest that they could be caused by differential strategy use over different items or by occasional fast guessing. All of these studies demonstrate that a model that assumes that one's speed and ability change within the test fits the data better than does a model that assumes that speed and ability are fixed throughout the test. A useful feature of these models is that they can be used to perform process analysis within cognitive tasks. For example, one could explore item features that might elicit slowing down (to become more accurate), as in Type 2 processing, or speeding up when the answer is known.

Another implication of this kind of modeling is that it potentially reveals how time limits on tests might be set. For example, for primarily—although not necessarily exclusively—practical reasons, time limits are currently set at the test level. However, per-item time limits could be set and there are questions about how to do so and what the impact would be. Liao (2018) points out, for example, that research has suggested that low-ability examinees tend to benefit more from extended time limits, but that dependency modeling enables more refined statements about for whom and for what items relaxed time limits might make the most difference.

IRT, the Diffusion Model, and the Scoring Rule Model

A second link between cognitive psychology and psychometrics centers around use of the diffusion model. The diffusion model is widely used in cognitive psychology to model the time taken to choose a response (referred to as *evidence accumulation*); it is particularly used for binary decisions in memory and perception tasks. The model consists of both decision-time and non-decision-time (stimulus encoding, motor response) parameters. The model specifies a starting point, which typically would be neutral (i.e., halfway between the two choice responses, such as true vs. false). The starting point can be altered, however, to favor one response over another; this can be done, for example, by priming or by biasing expectations through differential payoffs (e.g., the correct answer is more likely to appear in the first rather than the second position). The model also specifies boundary separation, which is the difference in the information required between making one versus the other response. This is related to speed-accuracy tradeoff; tight boundaries favor speed, wide boundaries favor accuracy. The model specifies drift rate, which is the average rate of evidence accumulation within a trial (i.e., *speed of responding*). Although this model was designed to capture cognitive psychology experimental data, in which starting point and boundaries are routinely experimentally manipulated, Tuerlinckx and de Boeck (2005) showed that the diffusion model could be expressed as a two-parameter logistic IRT model by separating the drift rate into person (ability) and item (difficulty) components. Van der Maas, Molenaar, Maris, Kievit, and Borsboom (2011) extended this model and discussed the larger significance of linking IRT to cognitive psychology models; they also showed that the diffusion model could be used for the more complex tasks used in educational and psychological testing.

De Boeck and Jeon (2019) point out that the scoring rule models described in the Speed-Accuracy tradeoff section (Maris & van der Maas, 2012; van Rijn & Ali, 2018) are both methods one can apply to combine accuracy and time information from a test and models for behavior on a test. That is, they can be seen as competing with the diffusion model as explanations for task behavior. In fact, van Rijn and Ali (2018) compared their scoring rule approach with a diffusion model and with a hierarchical model (van der Linden, 2007) in fitting data from an eighth-grade mathematics and a college-level spelling test. They found that, compared to the hierarchical model, the scoring rule approach produced higher score reliabilities and higher external correlations, at least with the mathematics test. They point out that there is not yet software to estimate the diffusion model for large *N* tasks, such as is typically found in standardized testing situations.

Linear Logistic Test Model and Cognitive Diagnostic Model for Process Modeling

A third research area linking cognitive psychology and psychometrics relates to the use of the LLTM developed by Fischer (1973). The basic idea is to model item response as a function not of the difficulty of a particular item but of the difficulty of each of the cognitive operations or information processing steps involved in the item. Technically, the LLTM is a Rasch model with linear constraints imposed on the item parameters (i.e., the item difficulties). These constraints appear in the form of a design—or Q—matrix, an *m* attributes × *k* items matrix that reflects which of *m* attributes are invoked by which of *k* items. This method can be understood as a form of componential analysis in that item success is a function of the attributes (components) involved. It is a generalization of Sternberg's (1977) componential analysis, which also uses a Q matrix to represent the difference between partial and full tasks. The differences are that (a) Sternberg's approach uses least squares regression to estimate attribute parameters, whereas Fischer's LLTM is an IRT approach and therefore is estimated using maximum likelihood (ML) (or Bayesian) approaches; and (b) Sternberg modeled item response times, whereas as an IRT model the LLTM is typically used to model item responses. However, in his dissertation, Klein Entink (2009; Chapter 3) combined LLTM modeling of responses with loglinear modeling of response times on a figural matrix reasoning task. He showed the value of this approach in estimating component difficulties and time requirements (i.e., time intensities) for matrix solution steps (e.g., performing unique addition, subtraction, and identity operations). This approach enabled Klein Entink (2009) to quantify the increase in difficulty caused by the inclusion of a particular rule (e.g., the presence of the identity rule resulted in a less difficult item, the unique addition rule, a substantially more difficult item) and also the increase in expected time (time intensity) to complete an item based on the inclusion of particular rules (e.g., items take 57 seconds on average; the presence of the identity rule resulted in 12 seconds less time on average). Some attributes had relatively greater effects on time intensity and others on difficulty, but for the most part (1) the factors that increased difficulty were the same as the ones that increased time intensity, and (2) time intensity and difficulty correlated around 0.68. Klein Entink (2009) points out that although the design matrix for ability and time typically would be the same, it does not have to be, and also suggests that there might be theoretical reasons why some attributes might be expected to affect difficulty and different attributes to affect time. Recently, Zhan, Jiao, and Liao (2018) suggested an approach for using CDM in a joint modeling framework to model both responses and response times in a way similar to how Klein Entink combined LLTM and response time modeling in his joint model.

The LLTM (and CDM) framework is the basis for one approach to automatic item generation (AIG) in which item families (items with the same values on each of the steps) rather than items per se are calibrated (Cho, De Boeck, Embretson, & Rabe-Hesketh, 2014; Geerlings, Glas, & van der Linden, 2011; Geerlings, van der Linden, & Glas, 2013; Sinharay & Johnson, 2008). This allows one to design items and to model the difficulty of items by modeling the effects of particular attributes on item difficulty and item time. An argument is that this approach provides a sounder theoretical basis for item development in that it relies on an identification of the specific components or attributes of the construct being measured (Embretson & Yang, 2006). This, therefore, has the potential for enabling a deeper, process-level understanding of test scores.

In the previous pages, we have summarized what has been learned in cognitive psychology about the meaning and importance of response time in completing cognitive tasks. We have also examined how recent psychometric models build (or fail to build) on these lessons from cognitive psychology when they include response time in modeling test taker proficiency. In the next section, we provide an extended example intended to demonstrate how the lessons learned from cognitive psychology might be used to build a complex assessment of cognitive skills; the example in this case focuses on diagnostic reasoning in medicine. It is useful to note that the research we next outline was conducted without regard to response time issues per se. However, the activity of preparing this chapter allowed us to think more specifically about how response time does or could play a role not only in understanding diagnostic reasoning in medicine but also in thinking ahead to how we might develop assessments of medical diagnostic reasoning in the future.

Application: Diagnostic Reasoning in Medicine

Diagnostic reasoning may be defined as the process of arriving at a logical conclusion or diagnosis to explain a set of facts or data. Examples can be found in a mechanic's thinking about an automobile failure, an auditor's reasoning about an organization's health based on financial records, and a physician's reasoning about a patient based on his or her history and symptoms. A critical component of many diagnostic reasoning tasks involves hypothesis generation: the process by which decision makers generate a set of potential explanations from memory to explain observed data. Hypothesis generation arguably is the most important component of diagnostic reasoning because it determines the set of hypotheses the decision maker ultimately considers, which in turn can affect diagnostic accuracy, confidence, and the choice to search for additional information (Barrows, Norman, Neufeld, & Feightner, 1982). Thus, the decisions one makes based on the outcome of the hypothesis generation process can have important consequences for one's own health (as is the case when people must decide whether to seek medical care) and the lives of others (as is the case for diagnostic reasoning by physicians; Elstein, Shulman, Sprafka, & Allal, 1978).

Hypothesis generation processes have important implications for diagnostic accuracy (Barrows et al., 1982; Elstein et al., 1978; Thomas, Dougherty, & Buttaccio, 2014). First, identifying the correct causal interpretation of a pattern of data depends on the early stages of the hypothesis generation process. If decision makers fail to generate the correct hypothesis in the initial hypothesis set, they are unlikely to arrive at the accurate diagnosis (Asare & Wright, 2003; Barrows et al., 1982; Pelaccia et al., 2014; Thomas et al., 2008; Weber, Boeckenholt, Hilton, & Wallace, 1993). Second, overconfidence in a selected diagnosis appears to be due primarily to generating too few alternative hypotheses (Bailey, Daily, & Phillips, 2011; Dougherty & Hunter, 2003a, 2003b; Sprenger et al., 2011; Tidwell, Dougherty, Chrabaszcz, Buttaccio, & Thomas, 2016). Third, there is a strong tendency to search for information suggested by hypotheses under consideration (Bhattacharjee & Machuga, 2004)—a notion referred to as *hypothesis-guided search* (Dougherty, Thomas, & Lange, 2010; Lange et al., 2014; Thomas et al., 2008; Thomas et al., 2014). Hypothesis-guided search explains how people perceive the usefulness of the information and has been a valuable construct for elucidating the conditions under which people engage in confirmatory versus diagnostic search (Dougherty et al., 2010; Illingworth & Thomas, 2015; Lange et al., 2014; Thomas et al., 2014).

The Role of Cognitive Models/Architectures in Cognitive Theorizing

Cognitive architectures are programs that enable simulations of the human cognitive system. They are widely used in cognitive science to capture theories of cognition and test assumptions about the workings of the mind. Cognitive architectures have proven useful because they assist the researcher in understanding the results of complex interactions within the cognitive system for the behaviors of interest and afford researchers a way of studying the implications of their theoretical assumptions by observing the behavior of the model. HyGene (Thomas et al., 2008—short for **Hy**pothesis **Gene**ration) is a cognitive architecture developed to capture the cognitive processes underlying diagnosis—the generation of the most likely explanations for some pattern of observed data. HyGene provides a psychologically plausible account of how decision makers generate a set of candidate hypotheses, identify the best explanation of the data, evaluate the coherence of the generated hypotheses, provide probabilities for the generated hypotheses (Dougherty, Gettys, & Ogden, 1999; Thomas et al., 2008, 2014; Tversky & Koehler, 1994), and describe how those hypotheses are tested by exploiting the available information sources.

An Architecture for Hypothesis Generation—HyGene

HyGene is based on three principles: (1) *Cued recall*: information in the environment prompts the retrieval of associated cues from long-term memory; (2) *Limited capacity*: working memory capacity and task characteristics constrain the number of cues that can be actively considered by the decision maker; and (3) *Information propagation*: hypotheses maintained in working memory influence subsequent search behavior. The model assumes three main memory constructs: (1) exemplar or episodic memory (experienced cases), (2) semantic memory (book knowledge and prototypes), and (3) working memory (consciousness). Figure 9.1 provides a schematic overview of HyGene broken down into discrete algorithmic steps (described below). These steps are carried out iteratively in real time as the decision maker attempts to make sense of data observed in the environment by generating candidate hypotheses.

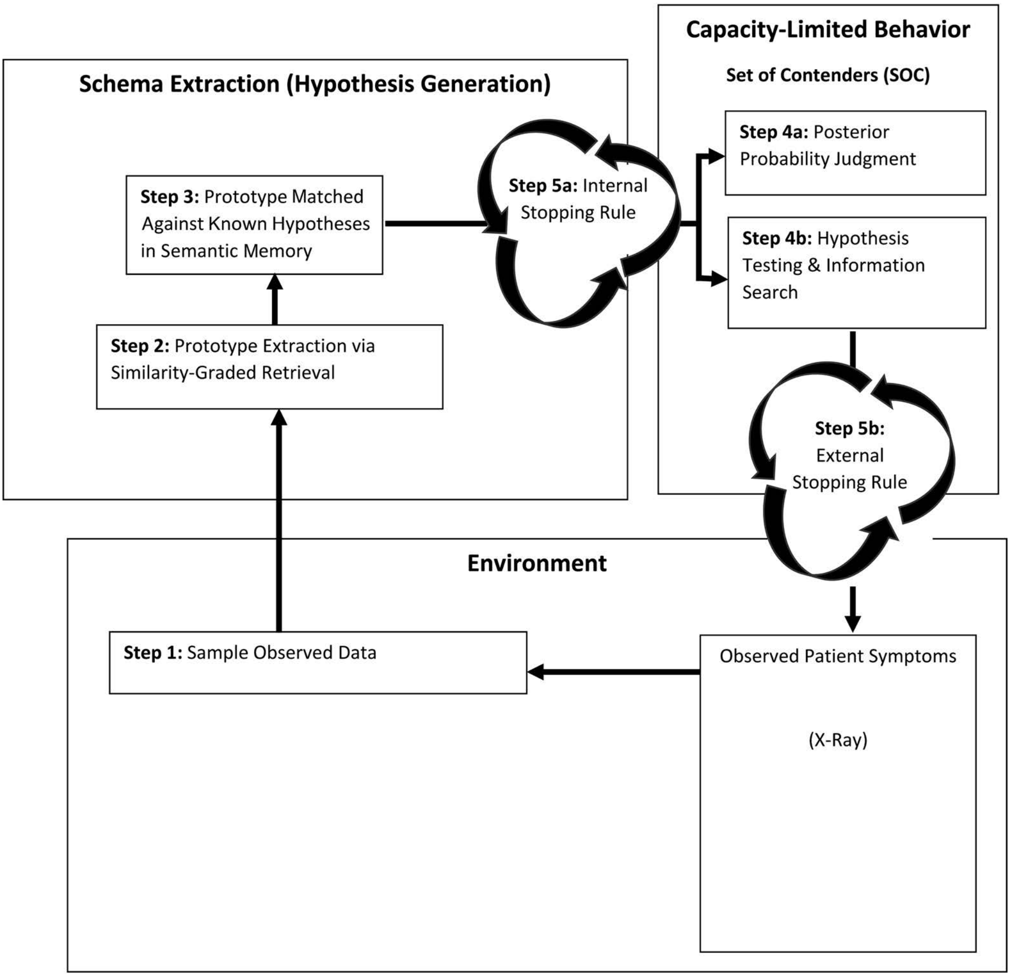

Capacity-Limited Behavior
Set of Contenders (SOC)
Schema Extraction (Hypothesis Generation)
Step 4a: Posterior Probability Judgment
Step 3: Prototype Matched Against Known Hypotheses in Semantic Memory
Step 5a: Internal Stopping Rule
Step 4b: Hypothesis Testing & Information Search
Step 2: Prototype Extraction via Similarity-Graded Retrieval
Step 5b: External Stopping Rule
Environment
Step 1: Sample Observed Data
Observed Patient Symptoms
(X-Ray)

Figure 9.1 Overview of HyGene Process Model.

Step 1 (Information Sampling): Information or data is observed in the environment (e.g., symptoms, history). In medical diagnosis, this can be from patient or caregiver inputs, physician inputs, and electronic medical records.

Step 2 (Derivation of Prototypical Representation): The semantic representations operate in concert with a set of retrieval operations that enable the decision maker to identify a set of candidate (disease) hypotheses to explain the observed data from Step 1 (patient symptoms, history). Specifically, a similarity-graded weighting algorithm uses the observed data to extract a disease prototype as well as the most likely patterns of co-occurring symptoms and test results (cf., Thomas et al., 2008).

Step 3 (Generation of Candidate Hypotheses): Semantic activation involves a disambiguation process in which the prototype (created in Step 2) is matched against "known" diseases (i.e., disease representations) in semantic memory. Importantly, the generation process allows the model to define a well-specified hypothesis space, which is necessary for deriving posterior probability distributions over hypotheses (Step 4a).

Step 4a (Probability Estimation): The activations of the hypotheses derived in Step 3 serve as input into a comparison process. The hypothesis evaluation mechanism is sensitive both to the degree of similarity between a hypothesis and the observed data and to the prior probability (i.e., experienced frequency of diseases in the practice of medicine, which can be subject to sampling and memory biases) of the hypotheses.

Step 4b (Hypothesis Testing and Information Search): Hypotheses serve as the basis for hypothesis testing and information search via a process called *Hypothesis-Guided Search* (Lange et al., 2014). That is, the usefulness of a particular medical test depends on the potential of its possible outcomes to change the physician's posterior beliefs about the disease hypotheses under consideration. Hypotheses themselves can be used to make sense of available data in the environment as it comes online over time. That is, the beliefs about particular disease hypotheses are iteratively reevaluated with the accumulation of additional information (e.g., Dx test results).

Step 5 (Search Termination): Any intelligent system must know when to stop both its internal search (memory retrieval) for hypotheses (Step 5a) as well as when to terminate external search and the collection of additional data (Step 5b), when further information or test outcomes are unlikely to be fruitful (Dougherty & Harbison, 2007; Dougherty, Harbison, & Davelaar, 2014; Harbison, Dougherty, Davelaar, & Fayyad, 2009; Harbison, Hussey, Dougherty, & Davelaar, 2012; Dougherty et al., 2010; Illingworth & Thomas, 2015; Thomas et al., 2014). Essentially, research indicates that search is likely to terminate when the estimated utility of decisions based on continued search fail to exceed the estimated utility of decisions based on the current information state by a sufficient margin, where sufficient is a threshold parameter that can capture individual differences or even the influence of time stress on stopping behavior (Dougherty & Harbison, 2007; Harbison et al., 2012).

Time Pressure Effects on Hypothesis Generation and Probability Judgment

HyGene assumes that hypothesis generation is constrained by the amount of time afforded to a decision maker to generate hypotheses. Greater time pressure results in the model generating fewer hypotheses into working memory or the Set of Contenders (SOC). Fewer hypotheses considered (i.e., fewer in the SOC) leads to two effects that are well established in the decision-making literature: (a) each hypothesis in the SOC will be judged to be more probable than it actually is; and (b) subadditivity (the sum of the probability judgments for all the hypotheses considered) will be greater than it would have been with more hypotheses. Subadditivity is a phenomenon in which the probability of a combined group (e.g., dying from cancer) is judged to be lower than the sum of the probabilities of elements of the combined group (e.g., dying from lung cancer, breast cancer, colon cancer, or all other cancers). These HyGene findings are consistent with the empirical results of Dougherty & Hunter (2003b), who found that time pressure led to greater subadditivity in a task requiring participants to judge the probability that a person came from a particular state within a region.

Time Pressure Effects on Hypothesis Testing

Like probability judgment, hypothesis testing behavior depends only on the hypotheses maintained in the SOC—the principle of hypothesis-guided search (Thomas et al., 2008; Dougherty et al., 2010). If only one hypothesis is maintained in working memory, hypothesis-guided search tends to follow an associative search or *confirmation search strategy* that can lead to preferences for nondiagnostic or positive tests: tests that do not always discriminate between differential disease hypotheses because their results may have similar, even identical, likelihoods across relevant disease hypotheses.

The simultaneous consideration of multiple hypotheses facilitates access to test diagnosticity signals from memory in HyGene. Thus, if more than one hypothesis is actively maintained in the SOC, the HyGene model can use a *diagnostic search strategy* and select tests for which the likelihoods of possible results are likely to differ across disease hypotheses. Because it takes time to generate multiple hypotheses into the SOC, HyGene predicts that lower time pressure will tend to increase the preference for diagnostic tests as a direct consequence of the increased likelihood that the model will consider more than one hypothesis (Mynatt, Doherty, & Dragan, 1993; Lange et al., 2014). HyGene makes specific predictions for when subjective and objective diagnosticity of tests will diverge via the particular hypotheses being considered. Studies using physicians have invoked the notion of hypothesis-guided search to explain how an initially incorrect diagnosis guides information selection, thereby making it less likely to encounter data to cue the generation of the correct diagnosis (Pelaccia et al., 2014).

Timing Phenomena Effects on Diagnosis: Data Acquisition Dynamics and Sequence Effects

When engaged in diagnostic hypothesis generation, data (symptoms and medical test results) are often acquired serially (one after the other). Although people should be insensitive to data (symptom) sequence, research indicates that people's decision-making is often sensitive to the order of information. The dynamic working memory buffer we use in the HyGene model translates the activations of data in working memory into weights governing the contribution of each piece of data to the hypothesis generation process. This integrated model produces two strong predictions concerning *temporal biases* in hypothesis generation that have been tested empirically.

The first prediction is a type of recency effect—symptoms presented later will contribute more to hypothesis generation and diagnosis than symptoms presented early. Lange, Thomas, and Davelaar (2012a) demonstrated such a recency effect empirically in hypothesis generation, where the preferred diagnosis (hypothesis) of participants was most consistent with the data that appeared later in the presentation sequence.

The second prediction is that the speed of symptom presentation attenuates the recency effect. As symptom acquisition speeds up, the dynamic HyGene model predicts a shift from the recency profile to a primacy profile. The model predicts this profile shift from recency to primacy because symptoms presented at the end of the sequence never accumulate enough activation to overcome the inhibition imposed from the earlier symptoms to enter working memory. Thus, at very fast rates of data acquisition, the HyGene dynamic working memory buffer predicts a primacy effect—symptoms presented early will contribute more to the hypothesis generation process than symptoms presented later. This prediction was confirmed in a diagnosis task: the recency bias attenuates (Lange, Thomas, Buttaccio, Illingworth, & Davelaar, 2013; Lange, Thomas, & Davelaar, 2012a, 2012b; Lange, Davelaar, & Thomas, 2013) and is even reversed (Lange, Davelaar, et al., 2013) under faster rates of symptom presentation. Recent findings suggest that the recency bias in hypothesis generation may reverse due to increases in the complexity of the diagnosis task (Lange et al., 2012b). Buttaccio, Lange, Hahn, and Thomas (2014) have recently extended the idea of hypothesis-guided search to the deployment of visual attention and search in which possible target representations are generated from long-term memory based on external cues (c.f., Lange, Thomas, Buttaccio, & Davelaar, 2012; Lange, Buttaccio, Davelaar, & Thomas, 2014). This theory and model directly apply to domains like radiology, where the radiologist uses patient information and symptoms to generate hypotheses for possible abnormalities that may appear in an x-ray (Hartzell & Thomas, 2017).

Buttaccio, Lange, Thomas, and Dougherty (2017) investigated the effects of time pressure on visual search by manipulating participant expectations for target characteristics via the cues presented before search. Their *Experiment 1A* indicated that search was less efficient (i.e., slower and less direct to the target) when the participants had less time to generate potential target characteristics after the cue was presented. Time pressure affected the efficiency of visual search presumably because not all relevant hypotheses (likely target characteristics) could be generated in the amount of time provided to aid the deployment of visual attention and search.

Summary of HyGene Findings

In summary, a critical component of medicine involves diagnostic hypothesis generation. Unfortunately, both laypeople and professionals tend to exhibit impoverished hypothesis generation and only consider a small subset of the relevant hypotheses. Impoverished hypothesis generation affects downstream behavior, including diagnosis, probability and confidence judgments, diagnostic test selection, and even the deployment of visual attention and search. Although people, physicians, and HyGene tend to generate the most likely hypothesis first, if the correct hypothesis is not generated early, it is unlikely to be generated at all. Importantly, task characteristics like time pressure and data timing (sequence and presentation rate) can influence diagnostic behaviors because they influence hypothesis generation. The influence of time pressure is particularly large when the diagnosis is complex; when multiple hypotheses are relevant to the observed data, people tend to exhibit the most bias (e.g., overconfidence and inefficient search) due to impoverished hypothesis generation. Because timing phenomena arise from a complex system with many interactive components, computational models like HyGene could likely play an important role in elucidating the implications of timing phenomena in diagnostic reasoning.

Other Diagnostic Reasoning Tasks and Abduction

As mentioned previously, diagnostic reasoning is not confined to medicine; it can also describe what mechanics do with cars that do not work, what auditors do when trying to determine the health of an organization based on financial records, what police investigators do when trying to understand a crime scene, and what scientists do in developing theories to explain data. From the standpoint of the philosophy of reasoning, these situations are ones that invoke abduction (i.e., “explanatory reasoning”; Douven, 2017), inference to the best explanation (Hobbs, Stickel, Appelt, & Martin, 1993), the “generation of hypotheses in order to find potential explanations of puzzling phenomena" (Holland, Holyoak, Nisbett, & Thagard, 1984, p. 136), or "a form of inference that takes us from descriptions of data patterns, or better, phenomena, to one or more plausible explanations of those phenomena" (Haig, 2018, p. 115). From the statistics literature they can be thought of as reverse causal inference ("causes of effects") situations (Gelman & Imbens, 2013); that is, diagnostic reasoning involves determining the most likely causes of specific effects (symptoms). Gelman and Imbens—focusing on the realm of statistics—suggest that such problems can be transformed into forward causal inference problems by hypothesis generation and model checking (e.g., determining what information might be missing and evaluating such information), similar to the stages in HyGene.

Abductive reasoning has been cited as characterizing medical diagnostic reasoning (Soldati, Smargiassi, Mariani, & Inchingolo, 2017). The framework of Soldati et al., similar to HyGene, provides a common framework for abductive reasoning tasks. It involves encoding data (e.g., symptoms), searching for explanations (in recency, prototype representation derivation, and generation of hypotheses), and then evaluating the hypothesis (hypothesis testing); these steps lead to a conclusion of whether to terminate or continue the search. Typically a distinction is made between the hypothesis generation and hypothesis evaluation stages of abduction ("making judgments of the best of competing explanations," [Haig, 2008, p. 1020]).

In the world of standardized testing, there are a few examples of abductive reasoning. On the hypothesis generation side, there are measures of creativity. For example, in Frederiksen's (1959) Formulating Hypothesis test (see also, Carlson & Ward, 1988; Frederiksen & Ward, 1978; Kogan, 2017), examinees are given a short period of time (typically about 2 to 3 minutes) to generate possible answers to open-ended prompts and then to indicate which of them are the most likely. Two example items from the test are the following:

> In Alcadia, a small country in Central America, the rate of death from infectious diseases declined steadily from 1900 to 1980. What factors might account for the decrease?
>
> The Port Byardia fleet had a mackerel catch that was relatively constant year to year during the 1970s, except for a sharp drop in 1974. Think of hypotheses (possible explanations) to account for the finding.

Other idea generation approaches are found in creativity measures such as Lubart and colleagues' *Evaluation du potentiel créatif* (Evaluation of the Creative Potential, EPoC) battery (Lubart, Besançon, & Barbot, 2011) and Educational Testing Service's (ETS) kit of cognitive reference test battery (Ekstrom, French, Harman, & Dermen, 1976), which includes tests—such as *Combining Objects* and *Substitute Uses* (from the *Flexibility of Use* scale)—that involve solving problems by using objects in unusual ways.

On the hypothesis evaluation side, consider the *Logical Reasoning* item type from a previous version of the Graduate Record Examination (GRE) (Educational Testing Service, 1995). This test includes questions in which examinees are given a surprising phenomenon (e.g., freezers use less electricity when packed with food than when half packed) and are asked which hypothesis, if true, would contribute to the explanation of the phenomenon (e.g., a volume of air requires more energy to be maintained at a low temperature than the same volume of food).

A Research Program to Study the Time Course of Diagnostic Reasoning

Tasks such as those just discussed or methods such as those used to identify hypothesis generation and evaluation could be employed in a research program to study the time course of diagnostic reasoning (including medical diagnostic reasoning). A program of research on assessing the skills involved in medical diagnostic reasoning could be conducted using a combination of cognitive correlates (covariate tasks measuring hypothesis generation and evaluation, as outlined in the previous section) and cognitive components approaches (i.e., partial task approach and the subtraction method). The LLTM or CDM could be used to analyze the skills involved in various full and partial tasks, both the responses and response times. Examinees could be given a series of medical diagnostic reasoning tasks from symptom presentation to decision and confidence judgments, and conditions of the tasks could be manipulated experimentally to enable isolation of the various stages of problemsolving. For example, for problem-solving Steps 1–5, the following could be manipulated:

1. **Information Sampling:** Manipulate number/type of symptoms and other inputs presented to examinee.
2. **Derivation of Prototypical Representation:** Manipulate the provision of prototypical representations (and their features) through a partial task approach.
3. **Generation of Candidate Hypotheses:** Manipulate prior knowledge (training of hypotheses), or just provide the hypotheses (partial task approach).

4a. **Probability Estimation:** Manipulate the difficulty of this stage by the difference between the hypothesis-data similarities (for an easy trial there is one hypothesis close to the data and the others are dissimilar; for a hard trial the similarities are close together).

4b. **Hypothesis Testing and Information Search:** Manipulate the information value of hypothesis tests and the variance in information value among candidate hypotheses.

5. **Search Termination:** Manipulate either deadlines or incentives.

Given a set of responses to whole and partial tasks constructed along the suggested lines, a cognitive diagnostic model (or other LLTM-type model) could be fit to the response data. Tasks would vary in their requirements for the various cognitive operations outlined in this chapter. A Q-matrix (a design matrix where the rows correspond to the test items and the columns to different cognitive operations) would represent task requirements. Cell entries, binary or continuous, would indicate the extent to which an item elicits a particular cognitive operation. A hierarchical response-response time model along the lines of Klein Entink (2009) or cognitive diagnostic model (Zhan, Jiao, and Liao, 2018) could be fit to the data to estimate both skills (for the various processes) and speed (for those processes). Although this would be an ambitious undertaking, the tools and methods for executing it are currently available.

Conclusions

The purpose of this chapter was to review the use of response time in cognitive psychology and in cognitive abilities research and to explore how response time may play a role in one class of cognitive tasks: medical diagnostic reasoning tasks. We also outlined a program of research designed to further understand the processes of diagnostic reasoning using response time and cognitive psychological methods.

Several concepts that may be useful to incorporate in future research on diagnostic reasoning emerged from the response time and HyGene reviews. First, there is a distinction between speed and power tasks; diagnostic reasoning is primarily a power task. This does not mean that response time analysis is fruitless, but it does mean that response time can be more challenging to interpret than it would be for a speeded task. The HyGene analysis suggested that a primary limitation on performance was that few hypotheses were generated, which seemed due both to working memory capacity limitations and to time limits. Second, the speed-accuracy tradeoff often—but not always—is a useful characterization of task performance. Within HyGene, speed-accuracy tradeoff has been manipulated through strict deadlines. With limited time, problemsolving (i.e., reaching a correct diagnosis) suffers due to fewer hypotheses being considered. Considering fewer hypotheses results in distorted probability and confidence judgments, poorer selection of diagnostic tests, and inappropriate search for additional information. We know little about the degree to which time pressure might affect individuals differentially, such as whether there are emotional aspects to performance deterioration (e.g., Caviola, Carey, Mammarella, & Szucs, 2017).

The review identified two research strategies that can be used to gain insight into diagnostic reasoning. A cognitive correlates approach examines correlations between performance on basic cognitive tasks and performance on the target task—in this case, medical diagnostic reasoning—as a way to understand the cognitive components of the target task. A cognitive components approach uses partial tasks along with the whole task to understand the processes and time dynamics associated with the target task. We suggested here a number of ways that the whole task of medical diagnostic reasoning could be broken down into partial tasks, and we suggested several psychometric modeling approaches for analyzing performance on the whole and partial tasks.

In recent years, significant advances have been made in modeling response time as well as response accuracy using psychometric methods that integrate them; this enables process analyses of complex task performances. Comparable advances have been made in understanding diagnostic reasoning—and medical diagnostic reasoning—and in incorporating this knowledge in cognitive architectures such as HyGene. It would be productive to integrate these two lines of research to enable further explication of the processes associated with medical diagnostic reasoning generally.

References

Anderson, J. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474. doi:10.1016/0010-0285(74)90021-8.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R. (2014). *Cognitive psychology and its implications* (8th ed.). New York, NY: Worth Publishers.

Asare, S. K., & Wright, A. M. (2003). A note on the interdependence between hypothesis generation and information search in conducting analytical procedures. *Contemporary Accounting Research*, *20*, 235–51.

Baddeley, A. D. (1968). A 3 min reasoning test based on grammatical transformation. *Psychonomic Science*, *10*(10), 341–342. http://dx.doi.org/10.3758/BF03331551

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (pp. 47–89). Cambridge, England: Academic Press.

Bailey, C. D., Daily, C. M., & Phillips, T. J., (2011). Auditors' levels of dispositional need for closure and effects on hypothesis generation and confidence. *Behavioral Research in Accounting*, *23*, 27–50.

Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general central practice. *Clinical and Investigative Medicine*, *5*, 49–55.

Bhattacharjee, S., & Machuga, S. (2004). The impact of generating initial hypothesis sets of different sizes on the quality of the initial set, and the resulting time efficiency and final judgment accuracy. *International Journal of Auditing*, *8*, 49–65.

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology, 9*, 1525. https://doi.org/10.3389/fpsyg.2018.01525

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126–1148.

Buttaccio, D. R., Lange, N. D., Hahn, S., & Thomas, R. P. (2014). Explicit awareness supports conditional visual search in the retrieval guidance paradigm. *Acta Psychologica*, *145*, 44–53.

Buttaccio, D. R., Lange, N. D., Thomas, R. P., & Dougherty, M. R. (2017). Does constraining memory maintenance reduce visual search efficiency? *The Quarterly Journal of Experimental Psychology*, *71(3),* 605–621. doi: 10.1080/17470218.2016.1270340.

Campbell, J. I. D., & Austin, S. (2002). Effects of response time deadlines on adults' strategy choices for simple addition. *Memory & Cognition*, *30*(6), 988–994.

Carlson, S. B., & Ward, W. C. (1988). *A new look at formulating hypotheses items*. ETS Research Report Series. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00268.x

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Caviola, S., Carey, E., Mammarella, I. C., & Szucs, D. (2017). Stress, time pressure, strategy selection and math anxiety in mathematics: A review of the literature. *Frontiers in Psychology*, *8*, 1488. https://doi.org/10.3389/fpsyg.2017.01488

Cho, S. J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika*, *79*(1), 84–104.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 1–11. https://doi.org/10.3389/fpsyg.2019.00102

Deary, I. J., & Stough, C. (1996). Intelligence and inspection time: Achievements, prospects, and problems. *American Psychologist*, *51*(6), 599–608. http://dx.doi.org/10.1037/0003-066X.51.6.599

Diependaele, K., Brysbaert, M., & Neri, P. (2012). How noisy is lexical decision? *Frontiers in Psychology*, *3*(348), 1–9. doi: 10.3389/fpsyg.2012.00348

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). Minerva-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209.

Dougherty, M. R., & Harbison, J. (2007). Motivated to retrieve: How often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(6), 1108.

Dougherty, M. R., Harbison, J. I., & Davelaar, E. (2014). Optional stopping in the search of memory. *Current Directions in Psychological Science*, *23*(5), 332–337. doi: 10.1177/0963721414540170.

Dougherty, M. R. P., & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, *113*, 263–282.

Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, *31*, 968–982.

Dougherty, M. R., Thomas, R. P., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 52, pp. 300–342). Cambridge, Massachusetts: Elsevier Academic Press.

Douven, I. (2017). Abduction. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2017 ed.). Stanford, CA: Metaphysics Research Lab. Retrieved from https://plato.stanford.edu/archives/sum2017/entries/abduction/.

Educational Testing Service. (1995). *GRE, the official guide: Practicing to take the general test, big book*. Princeton, NJ: Author.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service. Retrieved on 15 April 2018 from https://www.ets.org/Media/Research/pdf/Kit_of_Factor-Referenced_Cognitive_Tests.pdf

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Embretson, S., & Yang, X. (2006). Automatic item generation and cognitive psychology. *Handbook of Statistics*, *26*, 747–768.

Evans, J. S. B. T., & Wright, D. E. (1993). *The properties of fixed-time tests: A simulation study*. Technical Report 3-1993, Army Personnel Research Establishment. Plymouth, UK: Human Assessment Laboratory, University of Plymouth.

Fairbank, B. A., Tirre, W. C., & Anderson, N. S. (1991). Measures of thirty cognitive tasks: Analysis of reliabilities, intercorrelations, and correlations with aptitude battery scores. In P. L. Dann, S. H. Irvine, & J. M. Collis (Eds.), *Advances in computer-based human assessment*. Alphen aan den Rijn, Netherlands: Kluwer.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Frederiksen, N. (1959). *Development of the Test "Formulating Hypotheses": A Progress Report*. Princeton, NJ: Educational Testing Service.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem solving. *Applied Psychological Measurement*, *2*(1), 1–24.

Geerlings, H., Glas, C. A. W., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*, 337–359.

Geerlings, H., van der Linden, W. J., & Glas, C. A. W. (2013). Optimal test design with rule-based item generation. *Applied Psychological Measurement*, *37*(2), 140–161.

Gelman, A., & Imbens, G. (2013). Why ask why? Forward causal inference and reverse causal questions. NBER Working Paper No. 19614. doi: 10.3386/w19614

Goldhammer, F., Naumann, J., & Grieff, S. (2015). More is not always better: The relation between item response and item response time in Raven's Matrices. *Journal of Intelligence*, *3*(1), 21–40.

Goldhammer, F., Steinwascher, M. A., Kroehne, U., & Naumann, J. (2017). Modeling individual response time effects between and within experimental speed conditions. A GLMM approach for speeded tests. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 238–256.

Haig, B. D. (2008). Scientific method, abduction, and clinical reasoning. *Journal of Clinical Psychology*, *64*(9), 1013–18.

Haig, B. D. (2018). *Method matters in psychology: Essays in applied philosophy of science*. Switzerland: Springer Nature.

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). The lawfulness of decisions to terminate memory search. *Cognition*, *111*, 397–402. doi: 10.1016/j.cognition.2009.03.002.

Harbison, J. I., Hussey, E. K., Dougherty, M. R., & Davelaar, E. (2012). Self-terminated versus experimenter-terminated memory search. In N. Miyake, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 34th annual meeting of the cognitive science society* (pp. 426–431). Austin, TX: Cognitive Science Society.

Hartzell, C., & Thomas, R. P. (2017, September). Expectations influence visual search performance. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 61, No. 1, pp. 1529–1530). Los Angeles, CA: SAGE Publications.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, *150*. doi: 10.3389/fnins.2014.00150.

Hobbs, J. R., Stickel, M., Appelt, D., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, *63*(1–2), 69–142. AD-A259 608.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1984). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Hunt, E., Lunneborg, C., & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology*, *7*(2), 194–227. http://dx.doi.org/10.1016/0010-0285(75)90010-9

Illingworth, D. A., & Thomas, R. P. (2015). Price as information incidental search costs affect decisions to terminate information search and valuations of information sources. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 59, No. 1, pp. 225–229). Los Angeles, CA: SAGE Publications.

Irvine, S. (2014). *Computerised test generation for cross-national military recruitment: A handbook*. Amsterdam: IOS Press. ISBN 978-1-61499-362-9.

Just, M. A., & Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, and Behavioral Neuroscience*, *7*(3), 153–191.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Macmillan.

Klein Entink, R. (2009). *Statistical models for responses and response times* (Unpublished doctoral dissertation). University of Twente, Enschede, the Netherlands. Retrieved from http://www.kleinentink.eu/download/ThesisKE.pdf

Kogan, N. (2017). Research on cognitive, personality, and social psychology: II. In R. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological, and policy contributions of ETS* (pp. 413–452). Springer International Publishing. doi: 10.1007/978-3-319-58689-2.

Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence*, *14*(4), 389–433.

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4). https://doi.org/10.3390/jintelligence4040014

Lange, N. D., Buttaccio, D. R., Davelaar, E. J., & Thomas, R. P. (2014). Using the memory activation capture (MAC) procedure to investigate the temporal dynamics of hypothesis generation. *Memory & Cognition*, *42*(2), 264–274.

Lange, N. D., Buttaccio, D. R., Sprenger, A. M., Harbison, I., Dougherty, M. R., & Thomas, R. P. (2014). A memory-theoretic account of hypothesis generation and judgment and decision making. In A. Feeney & V. Thompson (Eds.), *Reasoning as memory* (pp. 71-92). Hove, UK: Psychology Press.

Lange, N. D., Davelaar, E. J., & Thomas, R. P. (2013). Data acquisition dynamics and hypothesis generation. *Cognitive Systems Research*, *24*, 9–17.

Lange, N. D., Thomas, R. P., Buttaccio, D. R., & Davelaar, E. J. (2012). Catching a glimpse of working memory: Top-down capture as a tool for measuring the content of the mind. *Attention, Perception & Psychophysics*, *74*, 1562–1567.

Lange, N. D., Thomas, R. P., Buttaccio, D. R., Illingworth, D. A., & Davelaar, E. J. (2013). Working memory dynamics bias the generation of beliefs: The influence of data presentation rate on hypothesis generation. *Psychonomic Bulletin & Review*, *20*, 171–176.

Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012a). Temporal dynamics of hypothesis generation: Influences of data serial order, data consistency, and elicitation timing. *Frontiers in Psychology: Cognitive Science*, *3*, 215.

Lange, N. D., Thomas, R. P., & Davelaar E. J. (2012b). Data acquisition dynamics and hypothesis generation. In N. Rußwinkel, U. Drewitz, J. Dzaack H. van Rijn, & F. Ritter (Eds.), *Proceedings of the 11th International Conference on Cognitive Modelling*, Universitaetsverlag der TU Berlin.

Lebiere, C., Anderson, J. R., & Reder, L. M. (1994). Error modeling in the ACT-R production system. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 555–559), Hillsdale, NJ: Erlbaum.

Liao, D. (2018). *Modeling the speed-accuracy-difficulty interaction in joint modeling of responses and response time* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Lohman, D. (1990). Estimating individual differences in information processing using speed-accuracy models. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, methodology: The Minnesota symposium on learning and individual differences* (pp. 119–163). New York, NY: Psychology Press.

Lubart, T. I., Besançon, M., & Barbot, B. (2011). *Evaluation du Potentiel Créatif (EPoC)*. Paris: Editions Hogrefe France.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633.

Meyer, D. E., & Kieras, D. E., (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*, *1044*(1), 3–65.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., … John, M. (2014). *Psychometric considerations in game-based assessment*. Princeton, NJ: GlassLab, Educational Testing Service.

Molenaar, D. (2018). *Response mixture modeling: Accounting for heterogeneity in item characteristics across response times*. Paper presented at the International Meeting of the Psychometric Society, New York, NY.

Molenaar, D., & Visser, I. (Ed.). (2017). Cognitive and psychometric modelling of responses and response times [special issue]. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 185–186.

Mynatt, C. R., Doherty, M. E., & Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives. *The Quarterly Journal of Experimental Psychology*, *46*(4), 759–778.

Neisser, U. (1967/2014) *Cognitive psychology, classic edition*. New York, NY: Psychology Press (Taylor & Francis; first published by Meredith Publishing Company).

Newell, A. (1994). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press. ISBN 0-674-92101-1.

Organisation for Economic Cooperation and Development (OECD). (2017). "What is collaborative problem solving?" In *PISA 2015 Results (Volume V): Collaborative Problem Solving*. Paris: OECD Publishing. https://doi.org/10.1787/9789264285521-7-en

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32.

Pelaccia, T., Tardif, J., Triby, E., Ammirati, C., Bertrand, C., Dory, V., & Charlin, B. (2014). How and when do expert emergency physicians generate and evaluate diagnostic hypotheses? A qualitative study using head-mounted video cued-recall interviews. *Annals of Emergency Medicine*, *64*(6), 575–585.

Ruitenberg, M. F. L., Abrahamse, E. L., de Kleine, E., & Verwey, W. B. (2014). Post-error slowing in sequential action: An aging study. *Frontiers in Psychology*, *5*, 119. doi: 10.3389/fpsyg.2014.00119.

Scalise, K. (2009). *Computer-based assessment: "Intermediate constraint" questions and tasks for technology platforms*. Retrieved 8/2/2018 from http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. http://dx.doi.org/10.1037/0033-295X.84.1.1

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *17*(1), 701–703.

Sinharay, S., & Johnson, M. S. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing*, *8*(3), 209–236.

Soldati, G., Smargiassi, A., Mariani, A. A., & Inchingolo, R. (2017). Novel aspects in diagnostic approach to respiratory patients: Is it time for a new semiotics? *Multidisciplinary Respiratory Medicine*, *12*, *15*. doi: 10.1186/s40248-017-0098-z.

Sprenger, A. M., Dougherty, M. R., Atkins, S. M., Franco-Watkins, A. M., Thomas, R. P., Lange, N., & Abbs, B. (2011). Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers in Cognitive Science*, *2*(129), 1–15. doi: 10.3389/fpsyg.2011.00129.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review*, *84*(4), 353–378.

Thomas, R. P., Dougherty, M. R., & Buttaccio, D. (2014). Memory constraints on hypothesis generation and decision making. *Current Directions in Psychological Science*, *23*, 264–270.

Thomas, R. P., Dougherty, M. R., Sprenger, A., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*, 155–185.

Thurstone, L. (1938). Primary mental abilities. *Psychometric Monographs*, *1*, ix + 121.

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. S., Buttaccio, D., & Thomas, R. P. (2016). Sources of confidence in judgment and decision making. In J. Dunlosky's (Ed.), *Handbook on metacognition* (pp. 109-125). Oxford, England: Oxford University Press.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70,* 629–650. doi:10.1007/s11336-000-0810-3.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567. http://dx.doi.org/10.1037/0033-295X.101.4.547

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist*, *30*, 128–134.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. (1978). Composition of episodic memory. *Journal of Experimental Psychology: General*, *107*(4), 393–419.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York, NY: Springer.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response time as collateral information. *Applied Psychological Measurement*, *34*, 327–347.

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356.

van Rijn, P., & Ali, U. S. (2018). A generalized speed–accuracy response model for dichotomous items. *Psychometrika*, *83*(1), 109–131.

Weber, E. U., Boeckenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1151–1164.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85.

Wilhelm, O. (2016). Special issue: "Mental speed and response times in cognitive tasks." Journal of Intelligence (ISSN 2079-3200). http://www.mdpi.com/journal/jintelligence/special_issues/mentalspeed

Williams, M. D., & Hollan, J. D. (1981). The process of retrieval from very long-term memory. *Cognitive Science*, *5*, 87–119.

Wright, D. E., & Dennis, I. (1999). Exploiting the speed-accuracy tradeoff. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 231–248). Washington, DC: American Psychological Association.

Yellott, J. I. (1971). Correction for fast guessing and the speed-accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology*, *8*(2), 159–199. http://dx.doi.org/10.1016/0022-2496(71)90011-3

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modeling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 262–286. DOI: 10.1111/bmsp.12114

10

Response Times in Cognitive Tests

Interpretation and Importance

Paul De Boeck and Frank Rijmen

Time in the context of cognitive tests is not a simple notion. It refers to the time the test taker is using to respond to the items of the test or to take the whole test, and it refers to the time the test taker is given for individual items, sections, or for the whole test. These are the *response time(s)* and the *time limit(s)*, respectively. The most relevant question regarding response time is how it must be interpreted, in other words, what the value is of response time for cognitive assessment. The interpretation depends on whether the time limit was communicated and whether it is used in scoring or not. The most relevant question regarding time limits is how strict or lenient they should be. Both these questions are highly complex. Still, an answer to these questions is important for the main purpose of cognitive tests. For example, the interpretation of response times will inform our understanding of cognitive processes; and, likewise, the stringency of a given time limit will affect the validity of the inferences we make from exam scores. An excellent overview of issues, models, and findings is given by Kyllonen and Zu (2016).

In this chapter, we will first discuss the interpretation of response times in the traditional type of cognitive ability tests starting from the notions of power and speed and based on empirical findings. This is followed by brief discussions of measurement invariance issues, response time decompositions and related accommodations for respondents with special needs, and response times in next-generation assessments. We close the chapter with practical conclusions.

Power and Speed

A first and evident possible interpretation of response times is that they are a measurement of speed. Speed is a rather ambiguous concept; it can be working speed, cognitive speed, or something else. It is unclear whether speed is a separate ability or a style, how much it is affected by motivation and attention, whether it is not also an indicator of the ability one intends to measure, and whether it is not simply the result of a chosen speed-accuracy balance. Historically, a rather sharp distinction was made between speed and power (e.g., Gulliksen, 1950). In the context of this distinction, speed is measured by how many items one can solve in a given amount of time, or, equivalently, how much time it would take to answer a given set of items. To avoid a possible confounding with power, the items need to be easy, so that the time spent on an item is the time needed to respond correctly rather than the time spent attempting to answer an item by those examinees unable to solve the item even given unlimited time. One can think of a number of items completed within a given amount of time or of the time needed to finish a given number of items as analogous to miles per hour for a car. Difficulties on the road that slow our car down prevent an accurate measure of its speed.

While difficulty should not play a role for the measurement of speed, time restrictions should not play a role for the measurement of power. Power is typically measured in comparison with the difficulty level a test taker can manage. A score on a test with a substantial number of difficult items is useful in this regard. Difficulty is relative, which is evident, for example, in the context of item response theory, wherein the probability of success on a given item depends on both the item's difficulty level and a given examinee's ability level, which are also both expressed on a shared scale.

In practice, a test never comes with unlimited time, so that, although the conceptual distinction between speed and power is clear, the distinction is not so clear in practice. An additional problem for the two notions is the assumption of maximum performance, which implies that the test taker performs at a maximum level: as fast as possible when speed is measured and with all available power if power is measured. It is uncertain when and if the assumption of maximum performance is met in practical testing situations, although one can easily imagine reasons why it may not be. People may have a habitual pace of cognitive work, driven by getting things done or by the enjoyment of the cognitive processes they employ; or, people may differ in their level of persistence, some never giving up and others abandoning difficult tasks as soon as they arise. One other important reason, and possibly one of the factors that affect pace, is the balance between speed and accuracy. It is well known that speed can be at the cost of accuracy and that a higher accuracy level can be at the cost of speed (Heitz, 2014). Even when a test taker is given unlimited time, she/he may still choose to work fast.

Given that the maximum performance assumptions are not and cannot be met, it makes sense that the distinction between power and speed is no longer considered as important as it has been in the past. Instead, the terms "speed" and "ability" are used, realizing that they correspond with "effective speed" and "effective ability" (van der Linden, 2009), where the modification "effective" refers to the actual level of speed and ability as co-determined by factors other than maximum performance, such as the speed-accuracy tradeoff. Note that often the term "level" is also used rather than "ability" to replace the older notion of "power" (e.g., Carroll, 1993; Davidson & Carroll, 1945). It is further commonly assumed that ability measurement is not really distorted when time limits are sufficiently lenient. Of course, deciding the degree of leniency may be of considerable practical importance given the need for testing efficiency in practice.

Understanding Response Time

In this section, we will focus on response time and only to some extent on time limits. The approach we will take is a bottom-up approach instead of a top-down approach. We are in the first place interested in understanding response time data and how they are related to accuracy data (correct vs. incorrect responses). For practical considerations and recommendations these will be the grounds we rely on as opposed to a more top-down approach based on the constructs of speed and ability. Rather than relying on definitions of constructs to interpret results, we try to understand the data and data structure before interpreting the results and formulating recommendations.

We will begin by summarizing research findings regarding response times for test takers presented with cognitive ability tests for which a measure of response time per item is available. Although response times for simple cognitive tasks may be relevant for the study of speed (Carroll, 1993), we will not discuss those findings here. The following paragraphs summarize findings related to complex cognitive tests. See De Boeck and Jeon (2019) for an additional review of this literature.

1. Based on latent variable modeling of response accuracies and response times, a speed dimension and an ability dimension can be differentiated, as in the hierarchical model for speed and ability (van der Linden, 2007), with loadings of (log) response times on the speed dimension (i.e., negative loadings on speed and positive loadings on slowness) and loadings of binary response accuracies on the ability dimension. Although we follow the terminology as used by van der Linden (2007), we do not refer to the corresponding constructs of speed and ability. The ability latent variable is a response accuracy latent variable and the speed latent variable is a response time latent variable. The fact that two different latent variables are needed is not a surprising finding, but it contradicts the layperson belief that more able (higher accuracy) persons are faster.
2. The correlation between ability and speed varies with the test. (See van der Linden, 2009, for examples). If the speed-accuracy balance were the sole basis for the correlation, the correlation would be negative because focusing on accuracy would make one work slower; yet, the correlation can be positive, (almost) zero, or negative. From a purely pragmatic perspective, *explaining* a correlation may seem unnecessary because any correlation—regardless of its sign—will improve the measurement of ability, which is the main purpose of cognitive tests. Unfortunately, it is neither clear how the correlations should be interpreted nor what the measured ability is, other than that it is the effective ability, possibly affected by various confounding influences (e.g., speed-accuracy tradeoff, motivation).
3. The correlation between item difficulty and item time intensity (the item-level parameter for response time) is positive. (See again van der Linden, 2009, for examples). The more difficult items of a test take more time. The robust replication of this finding contrasts with the variation of the correlation between ability and speed. A simple explanation is that an item that requires more cognitive work creates more opportunities for mistakes—any one of which is sufficient for an incorrect result. (The reverse is not necessarily true. A higher failure rate does not imply that more work was required. For example, a difficult knowledge question (e.g., "what is the capital of Albania?") may not require any work at all if one either realizes one's lack of knowledge or happens to know the answer.) The positive relationship between item response time and item level of accuracy also applies in adaptive testing when the test taker is presented with items that more or less match the ability level (Shi, 2017).
4. The latent variable model with two dimensions, one for response times (speed) and one for response accuracy (ability), is violated by item-wise dependencies between accuracy and response time (Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova & Maris, 2016; Bolsinova, Tijmstra, & Molenaar, 2017; Bolsinova & Tijmstra, 2016; De Boeck, Chen, & Davison, 2017; Meng, Tao, & Chang, 2015; Partchev & De Boeck, 2012). These dependencies are direct effects of response time on response accuracy (or vice versa) after controlling for the latent variables and item parameters and can also be captured through correlated residuals. In most studies, the item-wise dependency between response time and response accuracy is negative: slower responses tend to be less accurate. The dependency means that on average both within a person and within an item there is a relationship between response time and response accuracy that cannot be explained by the latent variables or item parameters.

5. The dependency between response time and response accuracy is positively correlated with item difficulty (Bolsinova, De Boeck et al., 2017; Bolsinova, Tijmstra et al., 2017; De Boeck et al., 2017). For more difficult items, the dependency is less negative or even positive (slower is more accurate). This is a robust finding; it generalizes across test formats (multiple choice and open format) and across conditions with and without time pressure (De Boeck et al., 2017; De Boeck & Jeon, 2019). Based on our recent findings, it also applies to adaptive tests. The relationship has also been observed when response time is used as a covariate for response accuracy as a dependent variable in the case where an ability latent variable and an item parameter are included in the model (Goldhammer et al., 2014; Naumann & Goldhammer, 2017).
6. When the dependency is modeled as depending on the person, it appears to be negatively correlated with ability. This means that the negative dependency between response time and response accuracy (slower is less accurate) is more pronounced for high-ability test takers and is less pronounced or even reversed for low-ability test takers (Bolsinova, Tijmstra et al., 2017; Goldhammer et al., 2014).
7. Finally, there are indications suggesting that the overall negative relationship between residual response time and response accuracy is in fact curvilinear even though the linear component of the relationship is negative (Chen, De Boeck, Grady, Yang, & Waldschmidt, 2018). For response times that are shorter than can be expected based on the average response time of the respondent and the item in question, accuracy increases with residual response time up to a certain point (a negative residual response time value and a roughly 0.80 proportion of correct responses); for longer response times, the accuracy decreases. Because the turning point comes at a below-zero value of the residual, the global relationship is negative. This result was obtained from relating bins of double-centered log response times (x-axis) to proportions of success per bin (y-axis) (Chen et al., 2018). Interestingly, Bolsinova and Molenaar (2018) also have shown more recently that there is curvilinearity in the dependency and that it has the very same shape as found by Chen et al. (2018). A curvilinear effect of time-on-task was also found by Naumann and Goldhammer (2017).

The explanation of these findings is not evident. Taken together, points 5 and 6 mean that for higher probabilities of success (easy items, high abilities) the dependency is negative (or less positive), whereas for lower probabilities of success (difficult items, low abilities) the dependency is less negative or even positive. Yet, while *faster is more accurate* may seem a general rule, for the two reasons we mentioned, it is not. First, it does not apply (or it applies less) to difficult items and low abilities; and second, the more precise relationship is most likely curvilinear.

Several explanations for these findings or subsets of these findings have been presented in the literature. The first explanation is based on *a dual processing theory* (Goldhammer et al., 2014; Naumann & Goldhammer, 2017). In our view, the theory does not need to imply a categorical distinction between two types of processing. Instead, there may be a range going from completely automated processing such as when readily available knowledge is retrieved (e.g., 7 × 8 = ?) to fully controlled sequential processing such as when one works through different steps (e.g., 8 + 8 = 16 + 8 = 24, etc., or 5 × 8 = 40 + 2 × 8 = 56). Across persons and items, all kinds of mixtures between these two extremes may exist. For this dual processing mixture explanation to apply, one must assume that automated processing is faster and more accurate than controlled processing and that controlled processing takes time to be successful. If this interpretation applies, then, what is captured by the hierarchical model of ability and speed is an average of automated versus controlled processing, reflected in the latent variables and item parameters, while deviations from the average explain the residual associations between response time and response accuracy. Furthermore, if it is also assumed that (primarily) automated processing applies for higher abilities and easier items, whereas (primarily) controlled processing applies for lower abilities and more difficult items, then a switch in the dependency follows. Fast responding would be more accurate for easy items and high abilities because it is inherent to automated processing, whereas controlled processing takes time to be successful so that fast responses would be less accurate. Response mixture models have been formulated in line with this explanation, with classes of responses for relatively faster and relatively slower responses (Molenaar & De Boeck, 2018; Molenaar, Oberski, Vermunt, & De Boeck, 2016).

Second, if the speed-accuracy balance varies during the test, one can expect response time to be positively associated with response accuracy. However, when it's the activated cognitive capacity (e.g., attention, concentration) that varies, one can expect response time to be negatively associated with response accuracy. This would be in line with the diffusion model notion of drift rate (accumulation capacity of information). A stronger drift rate makes for faster and more accurate responses (Ratcliff, Smith, Brown, & McKoon, 2016), so that a variation of the drift rate during the test would lead to a negative dependency between response time and accuracy. Because the dependency is on average negative, variation of processing capacity may be the explanation (De Boeck et al., 2017). For the correlation of dependency with item difficulty and ability, the principle of dominant responses being faster may be invoked: for easy items the correct response is the dominant response and therefore faster and for difficult items, incorrect responses are dominant and therefore faster. Similarly, for high-ability test takers, the correct response is dominant and thus faster, while for low-ability test takers, incorrect responses are dominant and thus faster.

Third, the previous two explanations are difficult to reconcile with a curvilinear relationship between residual response time and response accuracy. Instead we briefly discuss here a possible although speculative explanation for the curvilinear relationship, based on two assumptions.

The first assumption is that examinees respond to an item as soon as they identify an answer they believe has a good chance of being correct. The chances of a very fast response being correct are smaller than for a response after more time, depending on the item and the certainty criterion the respondent is using. The second assumption is that when no likely answer can be found, the certainty criterion to release a response is lowered, especially when time is limited, and eventually either an educated guess or a blind guess will be made if one is relatively or completely uncertain, respectively. In this way, the first assumption explains the upward section of the curvilinear relationship and the second assumption explains the downward section. Depending on the specific encounter of a respondent with an item, the solution process works out well or not so well due to factors unrelated to systematic person and item factors, which explains the deviation from expectation, given the person parameters and the item parameters. The above assumptions explain why the proportions of success are a curvilinear function of the double-centered log response times without contradicting phenomena such as rapid guessing and slow guessing. (Rapid guessing simply means that the certainty criterion to release a response is very low and slow guessing means that the criterion is lower still after unsuccessful work on an item.) This explanation also leaves room for factors such as persistence (the reluctance to lower one's certainty criterion) and speed-accuracy balance (the level of certainty criterion).

Taken together, the interpretation of these findings remains too speculative to draw conclusions about the response process and its relationship with response time. Therefore, it would be premature to propose a joint model for response times and accuracy that would allow us to find out and measure more than does the now common hierarchical model for speed and ability. Models with extra dependencies between response times and accuracy have a better goodness of fit than the hierarchical model and can even improve measurement if certain assumptions are made (Bolsinova & Tijmstra, 2018), but a good explanation for the dependency is still lacking.

For most practical measurement purposes, we believe that the hierarchical model (van der Linden, 2007), although somewhat suboptimal, is a reasonable model to work with for joint accuracy and response time data.

Measurement Invariance as a Function of Time

There are two findings from the previously discussed studies with potential measurement consequences, because they concern time and measurement invariance. First, the dependencies in question imply a lack of measurement invariance as a function of response time. The accuracy item parameters of fast responses and slow responses are different, so that the measurement of ability is to some degree confounded with the measurement of speed (De Boeck et al., 2017). Interestingly, while the item parameters are different, the two ability latent variables (one for slow responses and the other for fast responses) cannot always be differentiated or, when they are distinct, they are highly correlated (Partchev & De Boeck, 2012). Second, measurement is not invariant across strict and lenient response time limits per item (De Boeck et al., 2017) although the correlation between the two conditions is again very high—about 0.80 (Davison, Semmes, Huang, & Close, 2012). This latter correlation is not as high as the former correlation between the abilities underlying spontaneously fast and slow responses, but it was obtained from two different but parallel test versions, whereas the former was obtained from the very same test. The violation of measurement invariance under the influence of rather drastic time pressure conditions as in Davison et al. (2012) or Ren, Wang, Sun, Deng, and Schweizer (2018) gives rise to a second dimension; however, the consequences in terms of measurement correlation are only moderate.

Decomposition of Response Time and Accommodations for Special Needs

The total response time consists of several components. For example, consider a simple arithmetic item "John has five apples. Mary has twice as many apples as John. How many apples do they have together?" One can distinguish between reading time (access time), time to translate the words into an equation, solving the equation, and formulating a response. Each of those processes takes time and contributes to the total response time. There is an extant literature on the decomposition of response times for simple cognitive tasks (see Chapter 9).

In educational assessments, tasks are of a higher complexity. Nevertheless, one can, at least conceptually, distinguish between components such as access time, time to engage with and solve the item, and motor response time. As a matter of fact, this distinction is being proposed in Chapter 4 to determine whether a student should be given extra time, where extra time would be given to students that require longer access time due to a reading disability, but not to students who need more time to devote to processes that are part of the construct—in our example, the time needed to solve an equation.

Depending on the task, the relative weight and the nature (automated versus sequential) of the different components are likely to differ, giving rise to different relations between response times and ability. Research on the different processes involved in taking a test and how they relate to response time, both for the overall population as well as for individual test takers with specific learning disabilities, is not only of interest from a research perspective but also helps to provide a scientific basis for deciding who should benefit from extended testing time. A conceptual analysis to determine which knowledge and skills are involved in taking a test and which of these are intended to be reflected by the test score is a prerequisite for determining whether extended time limits are appropriate for specific cases.

Response Times in Next-Generation Assessments

In the section on "Understanding response time," we focused on traditional cognitive ability tests. However, several planned and recently launched assessments place a larger emphasis on the application of knowledge and skills in real-world scenarios. For example, the next generation of science standards (NGSS) emphasize three dimensions of science, one of which, 'Science and Engineering Practices,' identifies the need to "describe behaviors that scientists engage in as they investigate and build models and theories about the natural world and the key set of engineering practices that engineers use as they design and build models and systems" (see nextgenscience.org for an overview of the NGSS). Behaviors such as developing and using models to explain scientific phenomena take time, and students who are good at science are likely to be efficient in carrying out these behaviors as well. Time is likely to become a more important part of the construct of next-generation assessments.

Practical Conclusions

Our practical conclusion from the results discussed thus far is rather conservative. Although time matters, it may not be so crucial that we should start experimenting with alternative models and alternative time limit practices in cases where the results matter for important decisions. Experimenting is beneficial for a better understanding of time and to develop better and more accurate models, and we should certainly invest in research of that kind, but for the time being our recommendation is not to change existing practices in high-stakes conditions. Although some modeling improvements are possible (e.g., including dependencies in the models), we wonder whether such improvements lead to larger changes than implied by the *generalization discrepancy*. The generalization discrepancy is the discrepancy between two tests developed by experts with the same level of expertise and with the same measurement purpose in mind. The issue is whether the improvement obtained with a different (and hopefully better) scoring of a given test would lead to a larger discrepancy between scores than the discrepancy between two scores obtained from two different tests based on equal levels of expertise from the part of the developers and with the same purpose of measurement.

Our rather conservative conclusion implies that stronger time pressures than the currently common and rather lenient time limits are not desirable, but we certainly do not argue against adjustments in the other direction for respondents with special needs. However, these adjustments are necessarily based on an ad hoc conceptual and pragmatic analysis and can hardly be grounded in empirical research because of the often individual or small-group nature of special needs populations.

References

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*, 1126–1148.

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical & Statistical Psychology*, *69*, 62–79.

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology*, *9*, article 1525.

Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics*, *41*, 123–145.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical & Statistical Psychology*, *71*, 13–38.

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical & Statistical Psychology*, *70*, 257–279.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytical studies*. New York, NY: Cambridge University Press.

Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence*, *69*, 16–23.

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis. *Educational and Psychological Measurement*, *5*, 411–427.

Davison, M. L., Semmes, R., Huang, L., & Close, C. N. (2012). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Educational and Psychological Measurement*, *72*, 245–263.

De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, *70*, 225–237.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10, article 102.

Goldhammer, F., Naumann, J., Stelter, A., Toth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*, 608–626.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, article 150.

Kyllonen, P., & Thomas, R. (2020). Using response time for measuring cognitive ability illustrated with medical diagnostic reasoning tasks. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 122–141). Abingdon: Routledge.

Kyllonen, P.C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4), 14. https://doi.org/10.3390/jintelligence4040014

Lovett, B. J. (2020). Extended time testing accommodations for students with disabilities: Impact on score meaning and construct representation. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 47–58). Abingdon: Routledge.

Meng, X. B., Tao, J., & Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*, 1–27.

Molenaar, D., & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Pychometrika*, *83*, 279–297.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov IRT models for responses and response Times. *Multivariate Behavioral Research*, *51*, 606–626.

Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences*, *53*, 1–16.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*, 23–32.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Science*, *20*, 260–281.

Ren, X., Wang, T., Sun, S., Deng, M., & Schweizer, K. (2018). Speeded testing in the assessment of intelligence gives rise to a speed factor. *Intelligence*, *66*, 64–71.

Shi, Y. (2017). *Response time and response accuracy in computer adaptive testing*. Presentation at IACAT 2017, Toki Messe, Niigata, Japan.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272.

11

A Cessation of Measurement

Identifying Test Taker Disengagement Using Response Time

Steven L. Wise and Megan R. Kuhfeld

Achievement testing is used to measure an individual's knowledge, skills, and abilities for a variety of purposes including educational attainment, proficiency assessment, and certification/licensure testing. To measure achievement, we begin by identifying a content domain of interest, developing a set of items that collectively represents the intended domain, and administering a series of these items to a test taker. The response given to each item provides a small amount of information about the test taker's achievement level. These bits of information are then aggregated across items to calculate a test score from which an inference can be made regarding the test taker's achievement level. This familiar measurement process has been extensively used for over a century to assess achievement, and our commonly used measurement models—both classical and item response theory (IRT) methods—provide the psychometric foundations underlying this aggregation of item-response-based information into test scores.

Though this is a straightforward process, the validity of an inference made on the basis of a test score rests on a fundamental assumption that all of the item responses reflect what the test taker knows and can do regarding the domain of interest. On one level, this assumption appears obvious and easy to meet. Careful and competent development of items that measure the domain would appear to logically ensure that each of the responses to these items reflects the intended domain. But the threat to the assumption comes not from the item content, but from the test taker. The assumption requires that the item responses come from *engaged* test takers who devote effort in applying their knowledge, skills, and abilities to answer each item they receive. In other words, we assume that if we administer an item to a test taker, that test taker will *try* to use what they know and can do to correctly answer the item.

In reality, however, it sometimes happens that test takers are disengaged during a test event. We define *disengagement* as a test taker responding to an item without applying her/his knowledge, skills, and abilities to provide an answer. In multiple-choice testing, one example of this is when the test taker is unmotivated to perform well on the test and does not care to read the item and consider its challenge before answering. Another example is the test taker who quickly submits answers to the remaining unanswered items as the test's time limit is expiring (see Chapter 6). Regardless of why it happens, disengagement threatens score validity because it introduces construct-irrelevant variance[1] that tends to negatively bias test scores (Haladyna & Downing, 2004).

The impact of disengagement on test performance can be sizeable; a synthesis of a number of studies found, for example, that unmotivated test takers tend to underperform in comparison to their motivated peers by an average of 0.58 standard deviations (Wise & DeMars, 2005). The purpose of this chapter is to explore both what happens when test takers disengage and how this disengagement should be managed during scoring.

Measuring Test-Taking Engagement

Most test takers appear to exhibit adequate levels of engagement throughout their test events. How, then, do we know when disengagement has occurred? Several methods have been used to assess what has variously been called test motivation, test-taking effort, or test-taking engagement. By far the most widely used method has been to ask test takers to complete a brief self-report instrument about their degree of engagement immediately after testing (Eklöf, 2006; Sundre & Moore, 2002). Another method uses person-fit statistics to assess the degree to which a test taker's responses are consistent with the measurement model being used (Meijer, 2003). Alternatively, one might make an inference about test-taking engagement from how carefully a test taker completes some accompanying task, such as a student survey (Boe, May, & Boruch, 2002; Zamarro, Hitt, & Mendez, 2016). A common feature of each of these methods is that they provide an overall assessment of a test taker's engagement. That is, the test event is the unit of analysis.

When computer-based tests (CBTs) are used, a more fine-grained approach to measuring engagement becomes feasible and is based on the time it takes a test taker to respond to an item. This suggests that we could evaluate engagement down to the level of individual item responses, and that it would be useful to conceptualize the administration of a test not as a unitary event but as a sequence of encounters between a test taker and items.[2] This idea was first investigated by Schnipke (1995), who studied the responses given during timed, high-stakes, multiple-choice tests. She found that as time was running out, some test takers would begin to rapidly submit answers to the remaining items, apparently in hopes of getting some correct through lucky guessing. Schnipke termed such behavior *rapid guessing*, as opposed to the more typically seen *solution behavior*, and concluded that the presence of rapid guessing during a test event indicated that the test was speeded for that test taker. Schnipke described rapid guessing as instances where "the examinee responds rapidly as time expires; accuracy will be at or near chance because the examinee is not fully considering the item. The examinee may skim the items briefly for keywords, but the examinee does not completely read the item" (p. 5). Later, it was discovered that rapid guessing is also commonly present in the data from unspeeded, low-stakes tests (Wise & Kong, 2005). In these instances, rapid guessing was due not to time pressure but to a lack of test-taker motivation.

Thus, even though the antecedents differ, rapid guessing sometimes occurs with both low-stakes and high-stakes tests. In high-stakes contexts, rapid guessing represents a tactical choice by a test taker who is trying to maximize his or her score,[3] whereas in low-stakes contexts, rapid guessing indicates instances in which a test taker was unconcerned about doing well on the item. In both instances, however, rapid guessing indicates that the test taker was disengaged: she/he chose not to apply her/his knowledge, skills, and abilities when answering the item (Wise, 2017).

Given both the presence of rapid guesses in test data and that they typically impart a systematic negative bias on test scores, it is natural to contemplate what to do about them. When aggregating item responses from a test event during scoring, should rapid guesses be included or excluded? To answer this question, it is helpful to consider the nature and dynamics of rapid-guessing behavior. If rapid guesses do not reflect engaged test takers' knowledge, skills, and abilities, what then do they reflect?

The Nature of Rapid-Guessing Behavior

This section provides an overview of much of what is known about rapid-guessing behavior. We will illustrate some of these findings using data from our organization's MAP® Growth™ assessment, which is a multiple-choice computerized adaptive testing (CAT) system that administers interim achievement tests to K-12 students. Because MAP Growth can be considered low stakes and unspeeded, it is reasonable to assume that rapid guessing on this assessment generally reflects unmotivated test taking.

Several general findings have emerged regarding rapid guessing. First, test takers rarely exhibit rapid-guessing behavior throughout a test event. In high-stakes settings, as described earlier, it is mostly observed toward the end of test events when time is expiring. In low-stakes settings, rapid guessing can occur throughout test events, although its prevalence tends to increase across item position.

Second, it is not uncommon for a test taker to move multiple times between solution behavior and rapid guessing. Figure 11.1 shows the pattern of responses from a MAP Growth Math test event in which there was frequent switching between the two behaviors. To provide some perspective on the rapidity of responses, for the vast majority of rapid guesses this test taker submitted a response to the item within a few seconds of its being displayed.

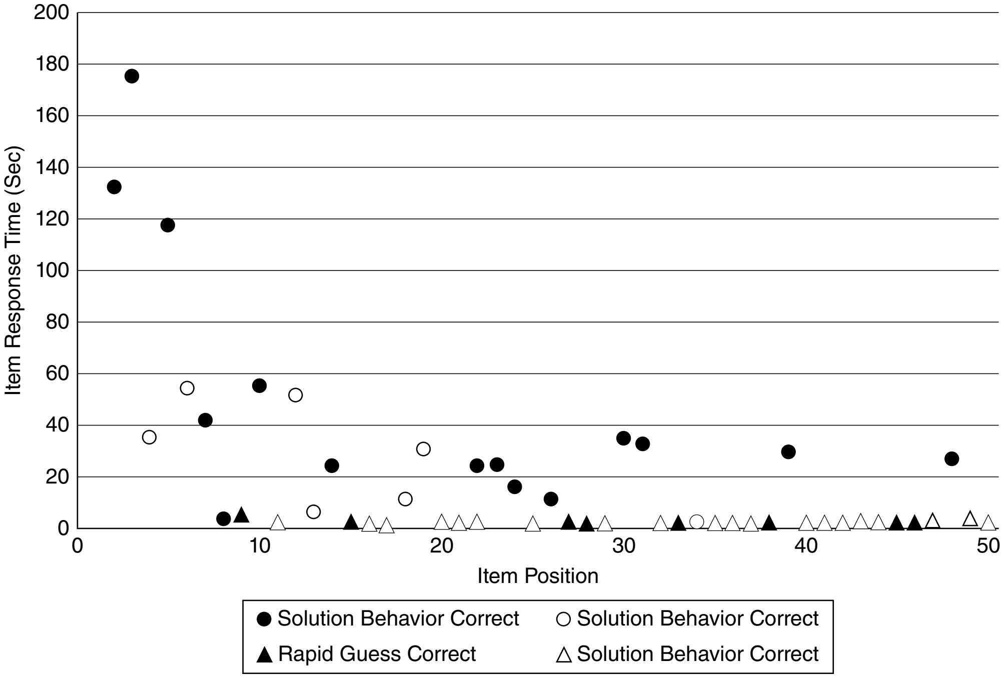


Figure 11.1 Responses from a MAP Growth Math test event in which there was frequent switching between solution behavior and rapid-guessing behavior. The rapid-guessing threshold for each response was established using the 10% normative threshold method (Wise & Ma, 2012).

Third, rapid-guessing behavior appears to correlate with several factors that provide clues as to its nature. Rapid guessing has multiple influences and is affected by characteristics of the item, the test taker, and the context in which the item is administered (Wise, Pastor, & Kong, 2009). It is more likely to occur with items that contain more reading or are perceived to be more mentally taxing (Wise et al., 2009). Males tend to rapid guess more frequently than females and, at the K-12 level, rapid guessing increases with grade (Wise, Ma, Kingsbury, & Hauser, 2010). Some studies have found rapid-guessing behavior to be unrelated to a test taker's achievement level (Wise, 2017), while others have found rapid guessing to occur more often with lower achievers (Goldhammer, Martens, Christoph, & Lüdtke, 2016; Wise & Gao, 2017; Wise et al., 2010). Regarding testing context, the likelihood of rapid guessing has been found to be related to item position, test stakes, and the time of day testing occurs (Wise et al., 2010).

The Rapid-Guessing Response Process

One of the key sources of validity evidence is the *response process* test takers use when responding to test items (AERA, APA, & NCME, 2014). Hubley and Zumbo (2017) noted that "identifying and understanding the mechanisms underlying how different respondents interact with, and respond to, test items and tasks is essential to understanding score meaning and test score validation" (p. 8). We believe the response process underlying rapid-guessing behavior to be fundamentally different from that used during solution behavior. Our evidence for this assertion has three components. First, there is typically a discontinuity between the accuracy levels of rapid guessing and solution behaviors. Second, unlike solution behaviors, rapid guesses tend to contain little, if any, psychometric information about a test taker's achievement level. Third, across items, the popularity of different response options for a multiple-choice item appears to be unrelated to option correctness/incorrectness.

Accuracy Discontinuity

In the basic conceptualization of disengaged responses and solution behavior (Wise, 2017), each behavior can be characterized by its own distribution of response times, as illustrated in Figure 11.2. In this hypothetical example, rapid guesses occur primarily during the first several

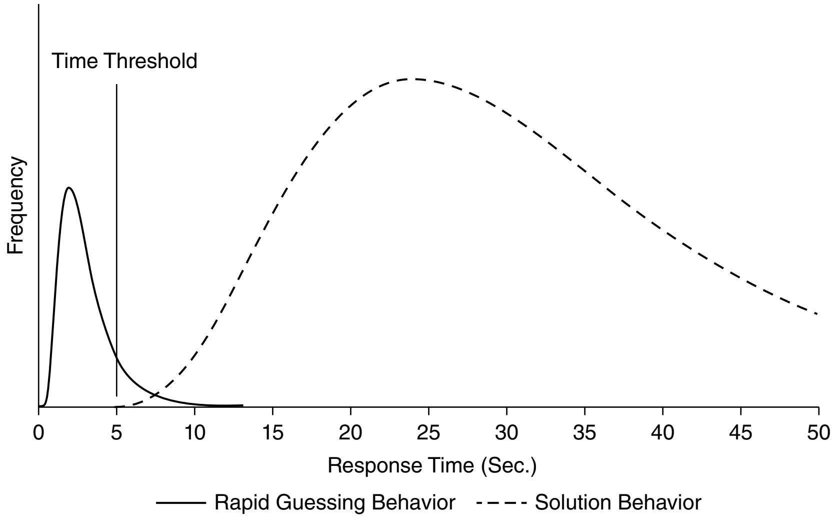


Figure 11.2 Conceptual distributions of the response times associated with responses to an item under rapid-guessing and solution behavior. The time threshold is arbitrarily chosen for discussion purposes.

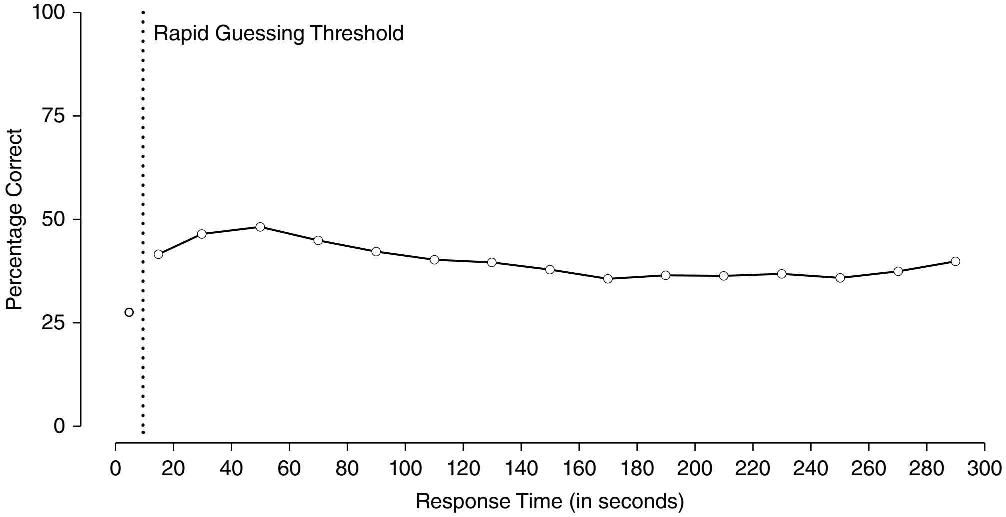


Figure 11.3 Discontinuity between rapid-guessing accuracy (to the left of the dotted line) and solution-behavior accuracy for various response time segments associated with responses to a MAP Growth Reading item.

seconds after the item is administered, and rarely after 10 seconds. In contrast, no solution behaviors appear before 5 seconds and they show a distribution that extends well beyond 50 seconds. If one chose to use a 5-second threshold[4] to operationally classify a response as a rapid guess, all responses occurring before 5 seconds would be classified as rapid guesses, those occurring between 5 and 10 seconds would be a mixture of rapid guesses and solution behaviors, and those occurring after 10 seconds would be comprised almost exclusively of solution behaviors. By comparing the accuracy of responses from these regions, we gain insights about the response processes being used in responding to the item.

Figure 11.3 summarizes the accuracy of over 200,000 responses to a MAP Growth Reading item across various response time segments. The dotted line in the graph indicates the rapid-guessing threshold, which for this item was 9.48 seconds.[5] To the left of the threshold, the accuracy of responses classified as rapid guesses is indicated. To the right of the threshold, the accuracy of solution behaviors up to the 20-second mark is indicated, followed by the accuracy of successive 20-second time segments. While the accuracy of rapid guesses was around 28%, the accuracy for the first solution behavior time segment jumped up to 42% and remained in the 35–50% range throughout the remainder of the time segments.[6] Importantly, in the solution behavior portion of the graph, adjacent time segments show similar accuracy rates and describe a generally smooth accuracy trend. Figure 11.3 shows a clear discontinuity between the relative accuracy of rapid guesses and solution behaviors, consistent with the idea that rapid guesses are determined by a different response process. Across other MAP Growth Reading items, the discontinuity varies in magnitude from minimal to quite large. As will be explained below, an item's rapid-guessing accuracy is not predominantly influenced by the accuracy of its solution behaviors.

Differential Psychometric Information

During a test, the correctness of an item response provides the psychometric information that is aggregated with that from other item responses to estimate achievement. The potency of an item response's information depends on the strength of a positive relationship between

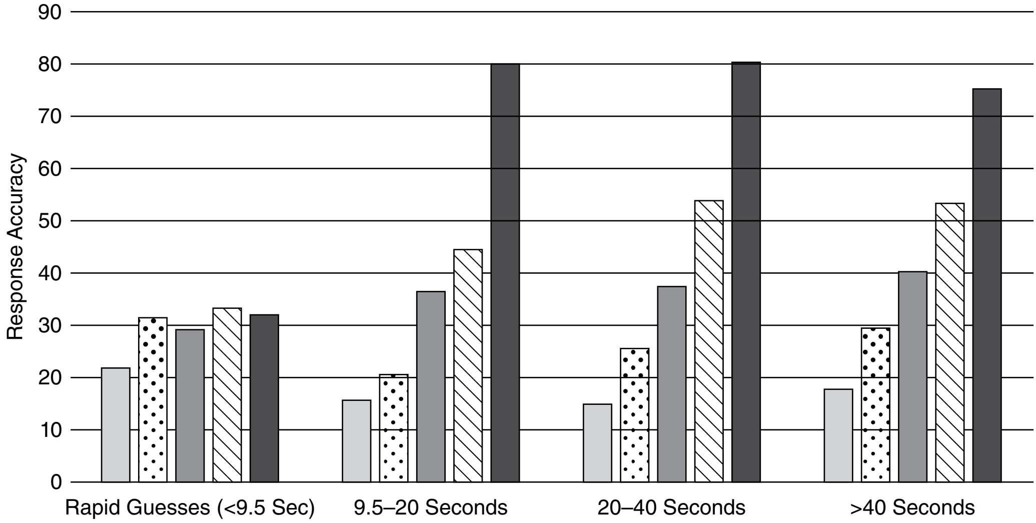


Figure 11.4 For the same Math Growth item as in Figure 11.3, response accuracy across overall test performance quintiles for rapid guesses and several solution behavior response time segments. The graphs show that rapid guesses are uninformative, unlike the solution behaviors.

response correctness and achievement level. This is a familiar concept in measurement: under classical test theory we routinely remove items that do not exhibit positive item-total correlations during item analyses, and in IRT we delete, as misfitting, items whose item characteristic curves are not monotonically increasing across achievement level.

Figure 11.4, which is based on responses to the same item shown in Figure 11.3, depicts additional information about response accuracy. The overall achievement distribution of the respondents to this item was divided into quintiles, each of which contained roughly 40,000 test takers. Figure 11.4 shows response accuracy, by quintile, for four time segments. The accuracy pattern for the initial time segment (i.e., rapid guesses) is clearly non-monotonic, with accuracy remaining relatively flat—particularly across the top four quintiles. This illustrates that the correctness/incorrectness of a particular rapid guess to this item carries little to no psychometric information about a test taker's achievement level. In contrast, the other three time segments (solution behaviors) exhibited the expected positive relationship between response accuracy and achievement level. Figure 11.4 illustrates how dramatically the accuracy pattern changed; rapid guesses exhibited an accuracy pattern clearly different from that observed from responses occurring a handful of seconds after the time threshold.

We should note that rapid guesses *do* sometimes exhibit positive relationships with achievement level. We inspected the content of several dozen MAP Growth items and identified two scenarios in which higher achievers could exhibit higher rapid-guessing accuracy. In the first scenario, some Reading items present a reading passage followed by a factual question about some aspect of the passage content. Occasionally, the question is about a topic or issue about which a test taker might already have some pre-existing knowledge. In this case, disengaged test takers who decide not to engage in solution behavior may proceed to quickly answer the question without first reading the passage. If higher achievers are more likely to have relevant pre-existing knowledge, rapid-guessing accuracy would be expected to show a positive relationship with achievement. Such rapid guesses, however, could still be considered construct irrelevant if one considers that the point of the item was to measure reading comprehension rather than factual knowledge that the test taker might possess without reading the passage.

In the second scenario, some MAP Growth items require multiple types of reasoning, with some types being more mentally taxing than others. For example, some Math items display a three-dimensional image of a geometric figure and require the test taker to identify a two-dimensional pattern that could be folded into the specified figure. Solving the problem under solution behavior requires test takers to visualize the folding of the different patterns. Test takers may perceive this item to be too mentally taxing and instead choose rapid-guessing behavior. In making the rapid guess, however, there may be one or more response options that could be easily ruled out by less mentally taxing reasoning. For example, if the geometric object is a cube, any response option showing a foldable pattern that did not contain exactly six sides could quickly be identified as incorrect. Being able to exclude one or more options in this less mentally taxing manner would decrease the effective number of choices, which in turn would increase the likelihood of a correct rapid guess. Hence, a rapid guesser might recognize this basic characteristic of cubes, while choosing not to engage in the more mentally taxing response process (visualization) required under solution behavior to identify the correct option. In this scenario, a positive relationship between rapid-guessing accuracy and achievement could be observed if the likelihood that the "six-sided" requirement would be recognized by rapid guessers were positively related to achievement level.

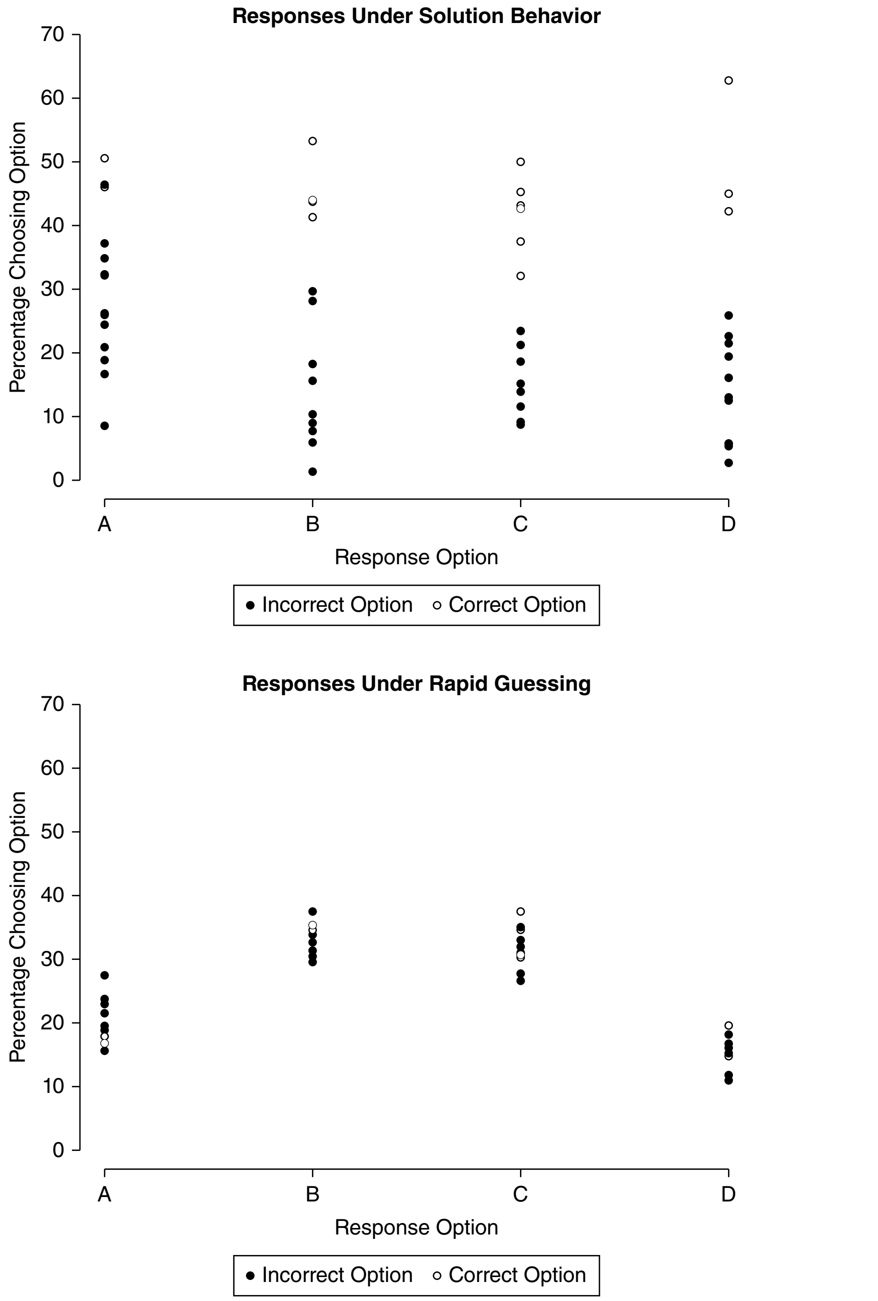
Despite these exceptions, it is generally the case that the psychometric information inherent in the correctness of rapid guesses tends to be extremely limited. At best, it is clearly deficient compared to the information provided by solution behaviors.

Coherence of Response Option Popularity

An additional characteristic of a well-functioning multiple-choice item is that the correct option is the one chosen most often, with the remaining options (i.e., distractors) being far less popular. Another way of stating this is that, across a set of items, each option position ought to be most popular when it is correct, and less popular when it is incorrect. Figure 11.5 shows the popularity of response option selection for responses from 15 MAP Growth Reading items. The popularity of options for responses classified as solution behaviors are shown in the upper display of Figure 11.5. Each option position exhibited a wide range of popularity and—almost without exception—an option was most popular when it was the correct answer to an item. In addition, on average, each option appears to be about as popular as the others. These results are consistent with what we would normally observe for a set of multiple-choice items.

In contrast, the results for responses classified as rapid guesses, shown in the lower display of Figure 11.5, are markedly different, with several new patterns being evident. First, the range of popularity across items for each response option position is much narrower. Second, the pattern of options being more popular when they are correct is no longer present. Finally, there are clear differences in the average popularity of option positions. Options B and C, for example, were markedly more popular than Options A and D. The finding that Option B was consistently selected by rapid guessers 30–40% of the time—regardless of option correctness—while Option D was selected only 10–20% of the time even when it was correct, strongly suggests that the response process underlying rapid guessing had little to do with the content of the response options.

Under solution behavior, option selection is driven primarily by the location of the correct option (as we would expect); under rapid guessing, option selection is driven more by option position than by correctness. This helps explain the variation in the magnitudes of the accuracy discontinuity we observed with the MAP Growth items. Whenever Option B contained the correct answer, rapid guesses were consistently correct about a third of the time; when Option D contained the correct answer, rapid guesses were correct only about a seventh of the time. Thus, because MAP Growth is an adaptive test for which the solution behavior accuracy rate



Responses Under Solution Behavior

Responses Under Rapid Guessing

Figure 11.5 Popularity of response option selection for 15 MAP Growth items under solution behavior and rapid-guessing behavior. At each option position, the symbols indicate both the percentage of the time the option was chosen for an item and whether that option was correct.

should be near 50%, the accuracy discontinuity appeared larger when Option D was correct and smaller when Option B was correct.

Collectively, the accuracy discontinuity, differential psychometric information, and option popularity patterns provide strong evidence that the response process underlying rapid guessing is very different from that used during solution behavior. Li, Banerjee, and Zumbo (2017) noted, however, that response time information "does not actually explain the cognitive processes that are involved in question-answering nor why they are used" (p. 172). In the next section, we propose an explanation regarding the cognitive processes that occur during rapid guessing and test-taking disengagement.

A Model of Rapid-Guessing Behavior

On high-stakes tests, rapid-guessing behavior tends to be clustered near the end of test events, as motivated test takers seek to maximize their scores as time is running out. On low-stakes tests—with test takers who may not be motivated to put forth the effort to attain maximal performance on an item—rapid guessing is more idiosyncratic, often appearing intermittently and throughout a test event. Can both types of disengagement, with their different antecedent conditions and patterns of occurrence, be represented under a single test-taking model? Wise (2017) proposed a model for explaining rapid-guessing behavior that involves a test taker making two choices. The first choice is whether to engage in solution behavior or rapid-guessing behavior. The second is the selection of a particular response option. If both choices are completed rapidly, disengagement can be inferred.

The First Choice

Understanding the first choice made by the test taker between solution behavior and rapid guessing involves consideration of two theoretical perspectives. The first is *dual-processing theory*, which specifies two distinct cognitive processes. The second is the *demands-capacity model of test-taking effort*, which describes how a test taker chooses between these cognitive processes.

Dual-Processing Theory

"The distinction between two types of thinking, one fast and intuitive, the other slow and deliberative, is both ancient in origin and widespread in philosophical and psychological writing" (Evans & Stanovich, 2013, p. 223). Empirically, a great deal of evidence has been found that the functioning of the brain can be characterized by two different types of cognition that have different functions (Stanovich, 2011). The first type of cognitive process, Type 1, has been described with numerous attributes, including that it is fast, non-effortful, autonomous, does not require working memory, and is relatively undemanding of cognitive capacity. The second, Type 2, can be characterized as slow, effortful, controlled, requiring working memory, being demanding of cognitive capacity, and involving analytical reasoning.

In the context of a multiple-choice test, response time could be useful to differentiate between Type 1 and Type 2 processes (Freeman & Dale, 2013; Kyllonen & Zu, 2016). Rapid-guessing behavior appears to be a manifestation of a Type 1 process, whereas solution behavior appears to be consistent with a Type 2 process. Moreover, it is important to note that under dual-processing theory Type 1 is the default process, and that "cognitive ability also is involved in the ability to effectively intervene with Type 2 reasoning and solve the problem" (Evans & Stanovich, 2013, p. 237).

The Demands-Capacity Model

One of the most interesting characteristics of rapid-guessing behavior is that engagement can change from one item to the next (as illustrated in Figure 11.1). The demands-capacity model of test-taking effort (Wise & Smith, 2011) was developed to account for this behavior. This model proposes that, when a test taker encounters a test item, two factors determine whether or not the test taker engages with the item. The first is the item's *resource demands* (*RD*), which basically means: *how much work does the item appear to require to fully answer effortfully*? RD is influenced by factors such as the amount of reading required, how mentally taxing the item appears to be, and—to a lesser extent—item difficulty (Wise et al., 2009). The second factor is the test taker's *effort capacity* (*EC*), which refers to the amount of effort the test taker is currently willing and able to give at the time the item is administered. EC is potentially influenced by numerous factors such as test stakes, time pressure, performance incentives, fatigue, how engaging the previous items were, and a variety of internal test taker factors such as achievement level, boredom, self-efficacy, conscientiousness, and competitiveness.

Both RD and EC can change during a test event, as the test taker proceeds through items. Items vary in RD, and a test taker's EC can change as she/he becomes bored, fatigued, more or less interested in the test, or feels increasing time pressure. According to the demands-capacity model, when encountering an item, the test taker compares his or her level of EC against the item's RD. If EC is higher, the test taker will engage in answering the item and exhibit solution behavior. However, if EC is less than RD, the test taker will disengage, resulting in a rapid guess. In a high-stakes testing situation in which the test taker perceives meaningful consequences associated with test performance, the stakes alone are likely to keep EC at a high enough level that solution behavior will consistently occur. In this case, the primary engagement threat is the possibility that the test taker will run short of time and resort to rapid guessing in hopes of improving their score. In contrast, in a low-stakes situation EC may become low enough that the variation in RDs can result in intermittent engagement (i.e., if RD exceeds EC for some items, but not others).

Combining ideas from dual-processing theory with the demands-capacity model, we believe that the first choice is essentially that of quickly deciding whether to apply a Type 1 or Type 2 process to respond to the item. This choice is driven by a comparison between the test taker's current level of EC and the item's RD. If EC does not exceed RD, the Type 2 process will not be applied; this will allow the default Type 1 process to be used in responding to the item.

The Second Choice

If a test taker chooses to give a rapid guess to an item, what happens next? Many researchers (the first author included) have previously described a rapid guess as a random response. This characterization stems from the commonly observed finding that the mean accuracy rates of rapid guesses closely resemble those which would be expected by random responding (Wise, 2015). Although it is true that, across items, rapid-guessing accuracy looks like random responding, when we look more closely a somewhat different picture emerges. For example, the lower display of Figure 11.5 shows that the four response option positions were clearly of unequal popularity. But if rapid guesses were truly random (i.e., all options had the same probability of being selected), each option position would have been selected equally often.

Furthermore, even if test takers intend to respond in a random fashion, could they actually do that? To ensure random responding, a test taker would need some type of randomizing device (like a coin, a die, a computer, etc.) that they would not typically have access to during a test event. It has been well established that people are incapable of simulating a randomizing device in their heads (Bar-Hillel & Wagenaar, 1991).

We believe that having committed to a rapid guess, a test taker's second choice consists of quickly making a largely impressionistic selection among the response alternatives. This guess may be influenced by pre-existing knowledge, but it is particularly vulnerable to the bias often seen when people are asked to make random choices. Attali and Bar-Hillel (2003) showed that when test takers guess the answers to multiple-choice items, they tend to show *edge aversion*, a common choice bias that occurs in many contexts when people are asked to make random choices. In a multiple-choice testing context, edge aversion appears as test takers showing greater preference for the middle of the set of options and lesser preference for the first and last option. Figure 11.5 shows a good example of this; Options B and C were markedly more popular than Options A and D. The presence of edge aversion in rapid-guessing responses underscores our assertion that rapid guesses are just that—guesses—and do not reflect the cognitive process assumed under solution behavior.

Under the two-choice disengagement model, rapid guesses represent Type 1 thinking and solution behaviors represent Type 2 thinking. Moreover, when we administer achievement tests, we tacitly expect engaged test takers to give responses that reflect their Type 2 thinking. It is tempting, then, to conclude that Type 1 thinking is antithetical to achievement testing. This conclusion, however, comes with a caveat. There is the possibility of instances in which a test taker has thoroughly learned some fact or association to the point of automaticity. If a test item asks about this, the test taker's response (probably correct) could occur very quickly and reasonably be characterized as both a Type 1 response (Evans & Stanovich, 2013) and classified as a rapid guess. This would constitute a misclassification of an engaged response as disengaged.[7] Although it is unclear how often this occurs in practice, test givers would be advised to review their items in terms of the likelihood that rapid responses represent automaticity rather than disengagement. Items that measure basic facts or comprehension would probably be most likely to receive responses reflecting automaticity.

What to Do about Rapid Guesses?

When we administer an achievement test, we assume that the test taker is engaged in the task of demonstrating what they know and can do. The use of item response time to identify rapid-guessing behavior provides us an item-by-item view into the veracity of that assumption—a view that was unavailable prior to the introduction of CBTs. This added information encourages us to consider a test administration not as a unitary event but as a series of item-person encounters during which the test taker may be engaged on some items but not on others.

We have shown evidence to support the basic conclusion that a rapid guess reveals a test taker's decision not to use her/his knowledge, skills, and abilities in answering an item. Essentially, a rapid guess represents an item-person encounter during which the test taker chose not to be measured. Such a choice could be due to either a lack of motivation or a lack of remaining testing time. Under rapid guessing, the resulting item response—regardless of whether it is correct or incorrect—reflects a construct-irrelevant response process and provides little, if any, psychometric information about the test taker's achievement level.

How, then, should we treat rapid guesses when we score a test? The traditional option is to include all item responses during scoring. The test taker receives a set of items and responds to those items, and only the correctness/incorrectness of the responses is used to calculate the test taker's score. An alternative option is to somehow take engagement into account during scoring. One way this might be done is through de-emphasizing or filtering out rapid guesses during scoring (Guo et al., 2016; Wang & Xu, 2015; Wise & DeMars, 2006). If rapid guessing were sufficiently pervasive during a test event, another approach might be to simply invalidate the test score. For example, our organization recently adopted (and then suspended) a policy of invalidating a test event if the percentage of responses classified as rapid guesses exceeds 30%.

The choice of which scoring option to adopt depends largely on two factors: the type of inference to be made about the test score and whether the test stakes are predominantly for the test taker or the test giver. Three scenarios representing different combinations of these factors will be considered as examples.

In the first scenario, test takers are trying to demonstrate some level of proficiency to gain something they want. These types of high-stakes assessments include classroom assessments, graduation exams, college entrance tests, and certification/licensure tests. Because the inference to be made is focused on whether or not proficiency has been demonstrated, the stakes are clearly higher for the test taker. Consequently, because the responsibility for test performance (and therefore engagement) lies with the test taker, the inclusion of rapid guesses during scoring could be appropriate—even while recognizing that their presence likely negatively distorted the score. That is, if a test score is negatively distorted due to low motivation or running out of time, it might be viewed that test performance is the test taker's responsibility and that the test giver should not feel compelled to correct for the distortion.

In the second scenario, the focus is on the achievement status of the individual test taker when a low-stakes test is used. An inference is to be made about what the test taker has learned and perhaps what she/he is ready to learn next. Examples of this include MAP Growth and a variety of low-stakes classroom assessments designed to provide instructional information to educators. As with most low-stakes tests, the stakes are considerable for the test giver—who seeks useful information regarding the achievement status and instructional needs of the test taker—and may be lower for the test taker. In this type of situation, because it is important that the test giver obtains the most accurate indicator of the test taker's achievement, it would be appropriate to take rapid guessing into account during scoring. Therefore, because of the potential to distort achievement estimates, rapid guesses should be excluded from scoring.

The third scenario combines elements of the first two. The test taker is asked to demonstrate achievement proficiency, but there are minimal personal consequences associated with test performance. In this case, however, the inference to be made is not about that individual's proficiency but about the rate at which proficiency was demonstrated across a group of test takers. Examples of this type of low-stakes testing include statewide accountability tests, the National Assessment of Educational Progress, and various international achievement testing programs (e.g., the Programme for International Student Assessment [PISA]). In these testing contexts, the stakes are higher for the test giver, and because the onus is on them to obtain the most valid (i.e., accurate) picture of the group's proficiency, it generally would be advisable for rapid guesses to be excluded during scoring. Including rapid guesses would likely lead to an underestimation of the number of test takers who were truly proficient—a result that test givers would generally consider undesirable.

Hence, the choice to include/exclude rapid guesses during scoring hinges primarily on whether measurement is intended to indicate as accurately as possible what the test taker knows and can do or if it represents an achievement hurdle that is considered the test taker's responsibility to clear. The pursuit of accurate scores implies that rapid guesses should be excluded; achievement hurdles represent an exception under which rapid guesses may be included.

One factor to consider when choosing a scoring option is the degree to which scores are likely to be distorted. Although it is clear that the expected amount of distortion increases with the number of rapid guesses, until the percentage of rapid guesses exceeds 10%, MAP Growth scores do not tend to be meaningfully affected (Wise & Kingsbury, 2016). It is important to note, however, that MAP Growth is a CAT. The amount of distortion expected from a rapid guess increases with the difference between the test taker's probability of passing the item under solution behavior and the probability of passing under rapid guessing. During a CAT, test takers are expected to pass items only about half the time; because of this, distortion will be relatively low. In contrast, for tests that administer easier items that test takers have a higher probability of passing, rapid guesses will have a higher distortive effect.

Wise and Kingsbury (2016) described a method for quantifying the degree to which a particular score was distorted. They estimated distortion as the difference between the usual IRT-based MAP Growth score (based on all item responses) and the corresponding effort-moderated score, in which rapid guesses were excluded (Wise & DeMars, 2006). They also plotted distortion scores against the proportion of rapid guesses, which revealed the functional relationship between the amount of rapid guessing and distortion. Measurement practitioners may find this procedure useful for investigating the impact of distortion on their particular test and for establishing guidelines regarding the amount of rapid guessing required to meaningfully distort scores.

An additional issue concerns the degree to which rapid guessing also can distort content representation. Wise (2020) found the propensity of rapid guessing to vary across item content areas, suggesting that the collective content of a test taker's set of engaged responses may differ meaningfully from the intended test blueprint. This finding indicates a previously unexplored way in which the presence of rapid guessing can threaten test score validity.

Concluding Comments

Measurement practitioners have been aware for many decades that disengaged test taking occurs in our achievement tests. Prior to CBTs being introduced, however, inferences about a test taker's engagement had to be made at the level of the individual test event. The ability of CBTs to record item response time permits a more fine-grained assessment of engagement through the identification of rapid-guessing behavior. Research using this capability has revealed that test takers generally disengage during only a portion of their items and that these disengaged responses reflect a momentary cessation of measurement. These findings underscore the essential role of test-taking engagement in our pursuit of valid scores.

Furthermore, for test events in which the proportion of rapid guesses is not too large, researchers have begun to explore the extent to which validity is improved if rapid guesses are excluded during scoring (Guo et al., 2016; Wang & Xu, 2015; Wise & DeMars, 2006). This possibility of salvaging valid scores, even when nontrivial amounts of disengagement have occurred, represents a promising area for future measurement research efforts.

Notes

[1] Construct-irrelevant variance is measurement error variance that arises from systematic error (Haladyna & Downing, 2004). Conceptually, this refers to a distortive influence on test scores that is unrelated to the construct being measured.

[2] Some CBTs allow test takers to go back and review their item responses, which can complicate the calculation of the total time a test taker spends interacting with an item. Although a CBT could collect metadata indicating whether (and for how long) item review occurred, this issue may not pose a great concern when considering rapid guessing. For example, Wise and Gao (2017) investigated disengagement on a CBT that provided item review and identified over 2,500 instances of rapid responses to the initial presentation of items. None of these rapid responses, however, were subsequently reviewed by test takers.

[3] When number-correct scoring is used, rapid guessing at the end of high-stakes tests is sensible, because unanswered items are sure to be incorrect. From an IRT scoring standpoint, however, rapid guessing is rational only if unanswered items are to be scored as incorrect. Nevertheless, some test takers may be motivated to answer all of the items within the allotted time, regardless of whether the behavior is rational or not.

[4] This is an arbitrarily chosen item threshold for this particular example. In practice, the threshold values typically would be unique to each item.

[5] The rapid-guessing threshold for this item was calculated using the 10% normative threshold method (Wise & Ma, 2012), which sets the threshold at 10% of the mean time that test takers historically spend taking the item (with a maximum threshold of 10 seconds).

[6] Because MAP Growth is an adaptive test based on the Rasch model, the solution behavior accuracy rate should be near 50% and rapid-guessing accuracy should more closely resemble that from random responding.

[7] There is an additional issue to consider. If a test taker has preknowledge of a test's items (and the associated answers), correct responses could be submitted rapidly. This behavior, which confounds the idea of rapid guessing as disengagement, represents a security threat. It might be differentiated from rapid guessing, however, by its accuracy rate. Rapid guesses are mostly incorrect, while responses based on preknowledge would be mostly correct.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*, 109–128.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, *12*, 428–454.

Boe, E. E., May, H., & Boruch, R. F. (2002). Student task persistence in the third international mathematics and science study: A major source of achievement differences at the national, classroom, and student levels (Research Report No. 2002-TIMSS1). University of Pennsylvania, Center for Research in Evaluation in Social Policy.

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66*, 643–656.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241.

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, *45*(1), 83–97.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. OECD Education Working Papers, No. 133. OECD Publishing, Paris.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*, 173–183.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27.

Harik, P., Feinberg, R. A., & Clauser, B. E. (2020). How examinees use time: Examples from a medical licensing examination. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 73–89). Abingdon: Routledge.

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. D. Zumbo & A. M. Hubley (Eds.), *Social indicators research series: Vol. 69. Understanding and investigating response processes in validation research* (pp. 1–12). New York, NY: Springer International Publishing.

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4), 14. https://doi.org/10.3390/jintelligence4040014.

Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response time as validity evidence: Has it lived up to its promise and, if not, what would it take to do so. In B. D. Zumbo & A. M. Hubley (Eds.), Social indicators research series: Vol. 69. Understanding and investigating response processes in validation research (pp. 159–177). New York, NY: Springer International Publishing.

Meijer, R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, *8*(1), 72–87.

Schnipke, D. L. (1995). *Assessing speededness in computer-based tests using item response times* (Unpublished doctoral dissertation). Johns Hopkins University, Baltimore, MD.

Stanovich, K. E. (2011). *Rationality and the reflective mind.* New York, NY: Oxford University Press.

Sundre, D. L., & Moore, D. L. (2002). The student opinion scale: A measure of examinee motivation. *Assessment Update*, *14*(1), 8–9.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*, 456–477.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*, 237–252.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretations, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61.

Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, *33*, 83–94.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*, 343–354.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*, 86–105.

Wise. S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010, May). *An investigation of the relationship between time of testing and test-taking effort.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*, 185–205.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendal (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139–153). Washington, DC: American Psychological Association.

Zamarro, G., Hitt, C., & Mendez, I. (2016). *When students don't care: Reexamining international differences in achievement and noncognitive skills*. Unpublished manuscript, Department of Education Reform, University of Arkansas.

12

Concurrent Use of Response Time and Response Accuracy for Detecting Examinees with Item Preknowledge

Seo Young Lee and James A. Wollack

Introduction

It has been well established that item response times can provide useful insight into examinees' test taking behaviors because observed response time patterns reflect the underlying nature of the problem-solving process (Lee & Chen, 2011; Luce, 1986; Schnipke & Scrams, 2002). Test takers are generally expected to spend a reasonable amount of time on each item in the exam. If response times are considerably shorter on some items, it can be indicative of a lack of time to complete the exam (i.e., speededness; see Chapter 6) or low motivation (see Chapter 11).

While much of the interest in response times has come from educational and psychological perspectives (e.g., Luce, 1986; Ranger & Kuhn, 2015), the application of response time data recently has gained traction for evaluating test security (Boughton, Smith, & Ren, 2017; Meijer & Sotaridona, 2006; Qian, Staniewska, Reckase, & Woo, 2016). With the transition of testing modes from paper and pencil to computer, attempts to memorize, steal, or share test content have become a serious threat to test security. Such security breaches can result in some candidates entering the testing session with prior knowledge of confidential test material, often referred to as *item preknowledge*. This results in an unfair advantage to those examinees and undermines the validity of score interpretations for all candidates. While traditional types of cheating such as answer copying or using unauthorized materials can be prevented and detected by thorough check-in procedures and proctoring before and during the test, these procedures are not useful for controlling item preknowledge. Statistical analysis has the potential to identify and control this approach to cheating.

Although no detailed studies exist demonstrating the responding behaviors of candidates with preknowledge, it is a commonly held belief that candidates will require less time to respond to items for which they have prior knowledge, leaving more time for them to respond to the remaining items. Hence, compared to a group of candidates without preknowledge, it is expected that candidates with preknowledge may represent another group with different response time patterns. In addition, it is reasonable to assume that candidates with preknowledge are more likely to select the correct answers; this would result in higher probabilities of getting the items correct when compared to other examinees of the same ability but without such knowledge. To the extent that examinees with preknowledge demonstrate this expected behavior, evidence of preknowledge may be present within these two sources of evidence: response times and response patterns. However, preknowledge may manifest itself to different degrees across the two sources of evidence such that using either in isolation may not be sufficient for identifying groups of respondents with preknowledge. Nevertheless, most methods to detect preknowledge have used the data with respect to only one of the two sources of evidence, and most have focused on item response accuracy.

Research on the Detection of Item Preknowledge Using Item Response Accuracy

A common approach in studies relying on item response accuracy data is to compare examinees' ability estimates based on items believed to be exposed (hereafter compromised) to those based on uncompromised items. The logical basis of this comparison is that examinees with item preknowledge would perform significantly better on compromised items than on uncompromised items, while examinees without preknowledge would show similar performance regardless of the compromised status of the items. The comparison has been evaluated by several statistics such as the likelihood ratio test (Sinharay, 2017), the score test (Sinharay, 2017), and posterior shift (Belov, 2016). In addition, there are approaches for detecting item preknowledge based on differences in IRT ability estimated using compromised and uncompromised items (Eckerly, Babcock, & Wollack, 2015).

While these studies have demonstrated that approaches attending entirely to response accuracy are effective at flagging examinees whose response patterns contain evidence of preknowledge, there may be cases in which the consideration of response times enables us to better detect item preknowledge. For example, there could be normal behaving candidates who just happened to have unusual response patterns. Additionally, the evidence of preknowledge in response patterns alone may not be sufficient when only a small number of test items are exposed to a small group of examinees.

Research on the Detection of Item Preknowledge Using Response Times

Though the majority of research in this area has used data on response accuracy to detect item preknowledge, there also have been studies that solely used response time data. In these studies, aberrance has been identified by comparing observed response times with expected response times under a certain type of response time model such as the effective response time model (Meijer & Sotaridona, 2006), the lognormal response time model (van der Linden, Scrams, & Schnipke, 1999; van der Linden, 2006), or the conditional log response time model (Toton & Maynes, 2019). Studies have demonstrated that the comparison of response times can be an effective approach for the detection of item preknowledge. For example, van der Linden and van Krimpen-Stoop (2003) showed that the residuals between the observed response times and expected response times estimated by the lognormal response time model were able to detect examinees with item preknowledge, whereas the analyses with only item responses were not effective. Meijer and Sotaridona (2006) showed that the mismatches between observed and effective response times were useful for detecting examinees with item preknowledge. Van der Linden and Guo (2008) demonstrated that the comparison of observed response times and predicted response times estimated by the lognormal response time model (Ln-RT; van der Linden, 2006) worked well to flag examinees who showed aberrant behaviors. However, in this study they also pointed out that the sole use of response times did not provide definitive evidence that such aberrant behaviors were caused by preknowledge; additional evidence obtained from response accuracy provided conclusive diagnosis of item preknowledge. This study raised awareness regarding the effectiveness of the concurrent use of response times and response accuracy for detecting item preknowledge.

Research on the Detection of Item Preknowledge Using Both Response Accuracy and Response Time

In addition to the models that use either response accuracy or response time, a model that incorporates both sources of data within a hierarchical framework also has been described in the literature (H-IRTRT; van der Linden, 2007). Given that the features of item preknowledge are likely to be reflected in both item responses and response times, a psychometric model that accommodates both data sources has the potential to provide more powerful evidence of item preknowledge. In the H-IRTRT model, response accuracy and response time are independently modeled at the first level of the hierarchical model. At the second level, the relationship between item responses and response times is modeled. Any IRT and response time models parameterized by item and person parameters are applicable to the H-IRTRT model. To account for the relationship, van der Linden introduced multivariate normal distributions between item and person parameters at the second level. Note that the joint distribution for item parameters is dependent on the models used at the first level.

Several studies have applied the H-IRTRT model to identify item preknowledge (e.g., Boughton et al., 2017; Qian et al., 2016; van der Linden & Guo, 2008). A noticeable feature of these studies is that they used a two-step approach; the first step was to analyze response time data to flag responses as aberrant and the second step was to use item response accuracy data to determine if the flagged responses were caused by item preknowledge. This two-step approach assumes that response time is a better indicator of aberrant behavior than item response accuracy, but it also considers that response times alone are not sensitive enough to confirm the cause(s) of aberrant behavior (e.g., item preknowledge). This two-step model therefore does not capitalize on the primary advantage of a single model incorporating both response time and accuracy: the potential to simultaneously evaluate both sources of aberrance. A viable way to use both item accuracy and response time simultaneously is to identify examinees with item preknowledge using a mixture model.

Mixture Models

A mixture model approach assumes that there are subpopulations in an overall population and distinguishes each subpopulation by allowing different parameter values of the same model (Rost, 1990; von Davier & Rost, 2007, 2016) or by applying different models for each subpopulation (von Davier & Yamamoto, 2006; Yamamoto, 1989). The application of mixture models in educational testing has become increasingly popular because it enables building a model reflecting different response behaviors of examinees in a single population. It is common that examinees with different response behaviors exist in the same exam group. For example, some examinees may have no problem completing the exam within the allotted time, while other examinees run out of time and rush to select answers as they approach the end of the exam. In such cases, traditional psychometric models that assume normal response behavior are not appropriate and other psychometric models (e.g., mixture models) are required to describe and differentiate such behaviors. Several studies have employed mixture models to classify irregular response behaviors such as random guessing (Meyer, 2010; Wang & Xu, 2015; Wang, Xu, Shang, & Kuncel, 2018) and rapid guessing resulting from test speededness (Schnipke & Scrams, 1997), but the use of these models in detection of item preknowledge has received little attention.

Meyer (2010) was the first to propose a mixture model that combines a mixture Rasch model (MRM; Rost, 1990) and a mixture lognormal model of response times (MRM-RT) for the purpose of differentiating examinees' rapid guessing behavior from solution behavior. The application of the MRM-RT showed that the use of mixture models that incorporate item accuracy and response time is a promising way to distinguish between aberrant and normal behavior. However, in contrast to many previous studies that have considered the relationship between these two data sources, the MRM-RT approach used by Meyer (2010) assumes that item accuracy and response time are independently homogeneous within latent groups. In addition, it accounts for the difference in observed response times by the mean difference between latent groups without specifying item- and person-specific parameters. The approach implies that the development of a mixture model that explains the relationship between item accuracy and response time may improve such a mixture model approach and help to flag aberrant response behaviors more accurately.

Another mixture model approach was proposed by Wang and Xu (2015) and later extended by Wang et al. (2018) for the purpose of distinguishing between aberrant and normal response behaviors. While a traditional mixture model assumes that each examinee exclusively belongs to one latent group, the Wang and Xu model allowed for an examinee to switch response behaviors multiple times in a test. Consequently, the group membership for examinees is evaluated for each item rather than for the entire test. This feature enables one to reduce the effect of aberrant responses on the estimation of parameters, resulting in more precise estimates. However, it does not provide comprehensive understanding of the response behavior for each examinee based on the entire test and therefore cannot be used to detect preknowledge at the level of the examinee.

The rest of this chapter will describe a new mixture model that focuses on the detection of examinees with item preknowledge. We begin by describing the model and we then provide an example of use of the model with operational examination data.

Mixture Rasch-Lognormal Response Time Model

Lee (2018) proposed a mixture model that incorporates item responses and response times in a single model by extending the H-IRTRT model into a mixture model for the detection of item preknowledge. Among IRT models and response time models that can be plugged into the H-IRTRT, Lee used the Rasch model because the most pronounced effect of preknowledge on the item responses themselves is expected to be related to item difficulty. For response times, Lee adopted the Ln-RT model. This mixture Rasch-Lognormal response time model will be referred to as the MixRL model throughout this chapter.

Level 1 Models

As introduced by Rost (1990), the mixture extension of the Rasch model (MRM) describes the probability of an examinee getting an item correct conditional on the examinee's ability and latent class membership. When the MRM is employed in the MixRL model, the item parameters for compromised items are different across latent groups, while those for uncompromised items are identical, as defined in equations (12.1) and (12.2). A constraint of this sort is necessary to solve the problem of scale indeterminacy existing in MRM. However, it does not mean that the MixRL model requires knowledge of the correct compromise status for every item. To apply the MixRL model, a set of items that is strongly believed to be uncompromised needs to be specified and fixed to be identical across latent groups. Although the percentage of uncompromised items required for effective use of the MixRL model has not been thoroughly studied, simulation studies by Lee (2018) showed that the model performed well when 20% of items were fixed as uncompromised. All remaining items are freely estimated by the MixRL model. This enables the user to minimize the risk of the misspecification of item compromise status while solving the problem of scale indeterminacy, because item parameter estimates for truly uncompromised items in the remaining items would be similar between groups, indicating no evidence of item compromise.

An item response for the MRM follows a Bernoulli distribution,

$$U_{ij} \sim f(u_{ij} \mid \theta_j, b_{ig}, g_j) \tag{12.1}$$

with success parameter

$$P\left(U_{ij}=1\middle|\theta_j, b_{ig}, g_j\right)=\frac{exp\left(\theta_j-b_{ig}\right)}{1+exp\left(\theta_j-b_{ig}\right)}(I_c)+\frac{exp\left(\theta_j-b_i\right)}{1+exp\left(\theta_j-b_i\right)}(1-I_c), \tag{12.2}$$

where $I_c=1$ for compromised items and $I_c=0$ for uncompromised items, b_{ig} is the item difficulty parameter for compromised item i in latent group g, b_i is the item difficulty parameter for uncompromised item i, and θ_j is the ability parameter for examinee j.

Analogous to the MRM, a mixture extension of the Ln-RT model for the detection of examinees with item preknowledge was developed. Again, the time intensity parameters (i.e., item-specific response time parameters) differ only for compromised items across the latent group.

$$\begin{aligned} f\left(t_{ij}\middle|\tau_j, \alpha_{ig}, \beta_{ig}, g_j\right) &= \frac{\alpha_{ig}}{t_{ij}\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left[\alpha_{ig}\left(\ln t_{ij}-\left(\beta_{ig}-\tau_j\right)\right)\right]^2\right\} I_c \\ &+ \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_{ij}-\left(\beta_i-\tau_j\right)\right)\right]^2\right\}(1-I_c), \end{aligned} \tag{12.3}$$

where t_{ij} is the observed response time for examinee j on item i, β_{ig} is the time intensity parameter for item i in latent group g, τ_j is the speed parameter for examinee j, and α_{ig} is the reciprocal of the standard deviation of the distribution of response time on item i in latent group g.

We assume conditional independence between item responses and response times following van der Linden and Glas (2010), so that the joint mixture model is given by

$$\begin{aligned} f\left(u_{ij}, t_{ij}\middle|\theta_{jg}, b_{ig}, \tau_{jg}, \alpha_{ig}, \beta_{ig}, g_j\right) &= f\left(u_{ij}\middle|\theta_{jg}, b_{ig}, g_j\right) f\left(t_{ij}\middle|\tau_{jg}, \alpha_{ig}, \beta_{ig}, g_j\right) \\ &= \sum_{g=1}^{G} \pi_g f_g(u_{ij}, t_{ij} \mid \theta_{jg}, b_{ig}, \tau_{ig}, \alpha_{ig}), \end{aligned} \tag{12.4}$$

where π_g are the mixing proportions that sum to one (i.e., $\sum_{g=1}^{G} \pi_g = 1$).

Level 2 Models

To account for the relationship between the two person parameters from the MRM and the mixture Ln-RT model (i.e., $\xi_{jg} = (\theta_{jg}, \tau_{jg})$) within each latent group, the two person parameters are assumed to follow a bivariate normal distribution

$$\xi_{j|g} \sim MVN\left(\mu_{pg}, \Sigma_{pg}\right) \tag{12.5}$$

with mean vector $\mu_{pg} = (\mu_{\theta_g}, \mu_{\tau_g})$, and covariance matrix

$$\sum_{pg} = \begin{pmatrix} \sigma^2_{\theta_g} & \sigma_{\theta_g \tau_g} \\ \sigma_{\theta_g \tau_g} & \sigma^2_{\tau_g} \end{pmatrix}. \tag{12.6}$$

The joint distribution of speed and accuracy item parameters for each of the two groups (i.e., $\psi_{ig} = (b_{ig}, \beta_{ig})$) also are assumed to follow a bivariate normal distribution

$$\psi_{ig} \sim MVN\left(\mu_{Ig}, \Sigma_{Ig}\right) \tag{12.7}$$

with mean vector $\mu_{Ig} = (\mu_{b_g}, \mu_{\beta_g})$, and covariance matrix

$$\sum_{Ig} = \begin{pmatrix} \sigma^2_{b_g} & \sigma_{b_g \beta_g} \\ \sigma_{b_g \beta_g} & \sigma^2_{\beta_g} \end{pmatrix}. \tag{12.8}$$

Potential of the MixRL for the Detection of Item Preknowledge

Throughout a series of simulation studies, Lee (2018) showed that the MixRL model was consistently successful in differentiating examinees with preknowledge from examinees without preknowledge under conditions including different proportions of compromised items and different proportions of examinees with preknowledge. In the comparison with other mixture models (i.e., mixture Rasch model, mixture Ln-RT model, and Meyer's model), the MixRL model outperformed all other mixture models for the detection of preknowledge in most simulated conditions. These results imply that the MixRL model is a promising approach for the detection of item preknowledge. However, given that real data are more complex than the simulated data in Lee's study, evaluation of the performance of the MixRL model in more realistic conditions is appropriate.

Use of MixRL with Operational Examination Data

Methods

We evaluated the performance of the MixRL model for detecting examinees with item preknowledge by fitting the model to the common credentialing dataset used throughout the *Handbook of Quantitative Methods for Detecting Cheating on Tests* (Cizek & Wollack, 2017). The dataset came from a computer-based linear credentialing exam and included 170 operational test items. This study used a subset of examinees from the original data set, comprising 393 examinees trained outside the United States. The study dataset included 21 examinees and 61 test items that were identified as being involved in some type of fraudulent testing behavior; however, it is unclear whether all flagged examinees had preknowledge and whether all flagged items were compromised. Additionally, there may be compromised items as well as examinees with preknowledge that were not previously identified by the program.

This study was designed to assess the performance of the MixRL model by manipulating two conditions: the proportion of examinees with item preknowledge and the proportion of compromised items. To examine different proportions of examinees with preknowledge, we created conditions in which 10%, 15%, and 20% of examinees had item preknowledge. The 21 flagged examinees were paired with differing numbers of randomly selected not flagged examinees (i.e., 189, 119, and 84). In these conditions, the number of test items was fixed to 170.

In an analogous manner, three proportions of compromised items (25%, 50%, and 75%) were manipulated for purposes of examining the effects of varying amounts of item compromise. Because the relative difficulty of the compromised and uncompromised items may impact the performance of the model, two separate approaches to sampling flagged items were used. For the first sample (Item set A), we produced a distribution similar to the distribution for the full sample of flagged items and sampled 24 items. The second item selection method focused on maximizing the discrepancy between flagged and not flagged examinees in terms of the proportion of correct responses and observed response times. This enabled us to use the evidence of item preknowledge existing in the dataset as much as possible. We identified 24 flagged items that showed a large discrepancy (Item set B).

Because the MixRL model requires the user to identify a set of uncompromised items to be identical between the two groups, we examined three different realistic ways to define what is meant by an uncompromised item within the mixture model: (1) all items not flagged by the credentialing exam program, (2) all items being administered for the first time, and (3) all items which were both first-time items and not flagged. Each condition was repeated 20 times. For each replication, a different set of not flagged examinees was selected randomly.

The MixRL model was estimated by Markov Chain Monte Carlo (MCMC) estimation using OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2014). Parameter estimates were obtained as posterior means from 8,000 post burn-in iterations after burning-in the first 7,000 iterations.

The detection of examinees with item preknowledge was assessed using four criteria: false positive rate, true positive rate, precision, and classification accuracy. For purposes of this study, the testing program's classifications of the compromise status of items and people were treated as true values. Therefore, the false positive rate was calculated as the proportion of examinees not flagged by the testing company who were identified by the MixRL model as belonging to the preknowledge group. The true positive rate was obtained as the proportion of examinees flagged by the testing company who were correctly classified in the preknowledge group. The precision was the proportion of examinees classified as having preknowledge that also were flagged by the testing company. The classification accuracy was the proportion of examinees classified in the same way by the testing company and the MixRL model.

Results

Table 12.1 shows the results of the detection of item preknowledge when the proportion of item preknowledge was manipulated. The detection of examinees with preknowledge was strongest

Table 12.1 The detection of item preknowledge under different proportions of preknowledge examinees

Type of Uncompromised Items	Proportion of Preknowledge Examinees	Number of Examinees in Total	False Positive Rate		True Positive Rate		Precision		Classification Accuracy	
			Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Not flagged items ($n = 109$)	10%	210	0.068	(0.032)	0.460	(0.203)	0.392	(0.148)	0.885	(0.016)
	15%	140	0.008	(0.021)	0.048	(0.147)	0.281	(0.329)	0.851	(0.007)
	20%	105	0.001	(0.004)	0.000	(0.000)	0.000	(0.000)	0.799	(0.003)
First-time items ($n = 37$)	10%	210	0.052	(0.084)	0.186	(0.276)	0.174	(0.199)	0.872	(0.053)
	15%	140	0.022	(0.034)	0.114	(0.213)	0.318	(0.345)	0.848	(0.013)
	20%	105	0.056	(0.100)	0.205	(0.267)	0.275	(0.330)	0.796	(0.078)
Not flagged first-time items ($n = 25$)	10%	210	0.084	(0.113)	0.324	(0.288)	0.239	(0.217)	0.857	(0.094)
	15%	140	0.087	(0.073)	0.431	(0.269)	0.422	(0.214)	0.840	(0.041)
	20%	105	0.081	(0.095)	0.345	(0.278)	0.409	(0.286)	0.804	(0.077)

when relatively few examinees were compromised, a finding that aligns with the simulation study results by Lee (2018) in which she showed that the true positive rates increased as the proportion of examinees with item preknowledge decreased. It also appeared that the method used for specifying uncompromised items within the MixRL model was an important variable in terms of the ability of the model to correctly classify candidates.

When all 109 not flagged items were specified within the model to be uncompromised (and therefore not allowed to vary across classes), approximately half of the examinees with preknowledge were classified correctly when the dataset included only 10% of examinees with preknowledge. However, the true positive rates decreased precipitously as the proportion of examinees with preknowledge increased. In fact, when 20% of the examinees had preknowledge, none were classified as having preknowledge.

The other three criteria (i.e., false positive rates, classification accuracy, and precision) also showed the same pattern. When item preknowledge was investigated by freeing only those items that were strongly believed to be compromised and constraining all other items to be the same across classes, the false positive rate was adversely affected.

When the investigation was performed by constraining only those items most strongly believed to be uncompromised (and allowing all other items to be separately estimated for each class), the detection of item preknowledge showed a slightly different pattern but did not show marked improvement. When equality constraints were placed on all first-time use items, true positive rates were around 0.2, regardless of the proportion of examinees with preknowledge. However, false positive rates also were considerably higher, hovering around 0.05. In the condition with constraints imposed only for those items with the strongest evidence of being secure (not flagged first-time items), detection rates were noticeably higher (between 0.324 and 0.431) but it was at the expense of even higher false positive rates. Given that the datasets included far more not flagged examinees than flagged examinees, the tradeoff of more true positives in exchange for more false positives did not appear worthwhile, as it resulted in this condition having the lowest classification accuracy for the highest level (20%) of proportion of examinees with preknowledge.

If the examinees with item preknowledge are correctly distinguished from normal examinees by the MixRL model, the item parameter estimates associated with the compromised items for the preknowledge group are expected to be lower than those for the normal group. Table 12.2 displays the comparison of item parameter estimates on average between the two groups for the items freely estimated by the MixRL model. Again, when the uncompromised items were defined as first-time items or not flagged first-time items, the other items included

Table 12.2 The comparison of item parameter estimates between normal and preknowledge groups

Type of Uncompromised Items	Proportion of Preknowledge Examinees	Item Difficulty				Time Intensity			
		Normal		Preknowledge		Normal		Preknowledge	
		Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Not flagged items	10%	−0.654	(0.201)	−1.388	(0.324)	3.851	(0.254)	3.715	(0.085)
	15%	−0.686	(0.227)	−0.954	(0.684)	2.957	(0.427)	3.515	(0.243)
	20%	−0.469	(0.328)	−0.297	(0.017)	2.866	(0.300)	3.293	(0.192)
First-time items	10%	−0.648	(0.051)	−0.517	(0.425)	4.062	(0.041)	4.002	(0.073)
	15%	−0.657	(0.176)	−0.742	(0.465)	4.010	(0.204)	3.909	(0.265)
	20%	−0.654	(0.138)	−0.709	(0.452)	3.747	(0.524)	3.676	(0.549)
Not flagged first-time items	10%	−0.642	(0.110)	−0.790	(0.506)	3.961	(0.256)	3.901	(0.219)
	15%	−0.647	(0.110)	−1.139	(0.275)	3.965	(0.225)	3.872	(0.179)
	20%	−0.596	(0.198)	−0.987	(0.562)	3.857	(0.381)	3.857	(0.275)

Table 12.3 The detection of item preknowledge under different proportions of compromised items

Type of Compromised Items	Proportion of Compromised Items	Number of Items in Total	False Positive Rate		True Positive Rate		Precision		Classification Accuracy	
			Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Item set A	25%	96	0.073	(0.095)	0.140	(0.236)	0.050	(0.059)	0.885	(0.077)
	50%	48	0.011	(0.017)	0.012	(0.043)	0.019	(0.055)	0.937	(0.014)
	75%	32	0.004	(0.004)	0.000	(0.000)	0.000	(0.000)	0.943	(0.004)
Item set B	25%	96	0.083	(0.076)	0.324	(0.316)	0.117	(0.103)	0.886	(0.057)
	50%	48	0.053	(0.051)	0.295	(0.300)	0.174	(0.116)	0.912	(0.033)
	75%	32	0.109	(0.020)	0.569	(0.066)	0.230	(0.029)	0.874	(0.018)

in the estimation were not necessarily compromised; a subset of uncompromised items must be included in the other items.

Table 12.3 shows the results of the detection of examinees with item preknowledge when the proportion of compromised items was manipulated. As expected, for Item set A, the MixRL model did not perform well in identifying examinees with preknowledge. However, with stronger evidence of preknowledge on the compromised items (Item set B), the performance of the model was considerably improved. In particular, when the majority of test items were compromised (i.e., 75%), more than 55% of the flagged examinees were classified into the preknowledge group. This result aligned with results from the simulation study Lee (2018) conducted. However, it should be noticed that the true positive rates increased at the cost of higher false positive rates, which was a known challenge for the application of mixture models in highly unequal sample sizes.

Concluding Comments and Practical Implications

In this chapter, we highlighted the usefulness of response times for understanding test taking behaviors, especially item preknowledge. The real-data-based simulation studies using the MixRL model demonstrated that the mixture model approach may be viable for the detection of item preknowledge when the data contain perceivable evidence of item preknowledge. Although not all examinees with preknowledge were detected by the model, given the fact that other studies using the same credentialing data have shown that more than half of the flagged examinees were not detected by any of the approaches at all (Wollack & Cizek, 2016), it is reasonable to state that the performance of the MixRL model for the detection of item preknowledge was comparable to other approaches. In addition, considering that a study using the H-IRTRT model (Boughton et al., 2017) flagged 95 candidates as having considerably shorter response times than expected, while only 6 were the candidates flagged by the testing program, the use of a mixture model approach may provide a better result for detecting examinees with item preknowledge.

At the same time, it is important to keep in mind that this is a real dataset, and in much the same way that the testing company may have failed to flag compromised items because the amount of evidence was not sufficiently high, it is likely that there are examinees who really did have preknowledge but who were not flagged by the company. This is made all the more likely when one considers that the evidence standard for examinees is likely much higher than the corresponding standard for items, given the serious consequences associated with misclassifying examinees.

These results suggest that in practice the determination of compromised items plays a significant role in the detection of examinees with preknowledge, and practitioners should decide the compromise status of items carefully through rigorous evaluations. This is also true in the application of the MixRL model, because our results show that the detection of examinees with preknowledge was affected by the set of items defined as uncompromised. Unfortunately, in practice it may not be straightforward for practitioners to determine which items are uncompromised. One approach, as implemented in this study, is to classify items based on how many times each item has been administered because newly administered items may be less likely to be compromised. Another feasible way is to perform a preliminary data analysis to identify items showing problematic performance. For example, the comparison of item responses and response times between various groups may provide evidence about potentially problematic items. As mentioned above, because the MixRL model does not require having an exact item compromise status for every item, we would suggest that practitioners take a conservative approach to selecting uncompromised items.

Practitioners who are considering applying the MixRL model should note that it is a model-based approach and therefore that if the model does not fit the data well, the results may be significantly impacted. To avoid this problem, model-data fit should be assessed before using the MixRL model. Unfortunately, there currently is limited research on this issue.

Once the MixRL model is applied, practitioners will still need to decide whether or not every examinee who shows features of item preknowledge (i.e., high probability of correct answers and short response times) should be assumed to have had preknowledge. As mentioned earlier, the classification of examinees is affected not only by patterns of item responses and response times but also by other factors such as the successful distinction between compromised and uncompromised items. Like other classification approaches, the risk of false positives always exists. Therefore, the final decision about whether or not individual examinees actually had item preknowledge should be informed by further investigations for each individual examinee. For these investigations, we suggest that multiple sources of evidence be collected using both observational and statistical approaches. As discussed in other studies (i.e., Boughton et al., 2017; Qian et al., 2016) based on data from real testing settings, item preknowledge is not the only reason for aberrant responses; other types of behavior such as low motivation or speededness may also contribute to this finding. In determining if an examinee's test score should be considered valid, it is always beneficial to triangulate sources of evidence.

It is undeniable that response times can be an easily obtainable and valuable resource used to investigate testing behaviors such as item preknowledge. In recent years, there have been many studies attempting to utilize response times to detect candidates' aberrant responses. However, because there is not a single gold standard approach that is applicable for all scenarios, future studies should be performed to improve existing methods and to develop new ones.

References

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83–97.

Boughton, K. A., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 177–190). New York, NY: Routledge.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Eckerly, C. A., Babcock, B., & Wollack, J. A. (2015). Preknowledge detection using a scale-purified deterministic gated IRT model. Paper presented at the annual meeting of the National Conference on Measurement in Education, Chicago, IL.

Harik, P., Feinberg, R. A., & Clauser, B. E. (2020). How examinees use time: Examples from a medical licensing examination. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 73–89). Abingdon: Routledge.

Lee, S. (2018). *A mixture model approach to detect examinees with item preknowledge* (Unpublished doctoral dissertation). University of Wisconsin-Madison, Madison, WI.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359–379.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing* (LSAC Computerized Testing Report 03-03). Newtown, PA: Law School Admission Council, Inc.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47.

Ranger, J., & Kuhn, J.-T. (2015). Modeling information accumulation in psychological tests using item response times. *Journal of Educational and Behavioral Statistics*, *40*(3), 274–306.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analysis. In C.M. Mills, M. Potenza, J.J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46–68.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2014). OpenBUGS user manual (version 3.2.3) [Computer software manual]. Retrieved from http://www.openbugs.net/Manuals/Manual.html

Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge in experimental data using conditional scaling of response times. *Frontiers in Education 4,* Article 49. https://doi.org/10.3389/feduc.2019.00049.

van der Linden, W. J. (2006). A lognormal model for response time on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 643–661). The Netherlands: Elsevier B. V.

von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 393–406). Boca Raton: Chapman and Hall/CRC.

von Davier, M., & Yamamoto, K. (2006). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 99–115). New York: Springer.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*, 456–477.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*, 469–501.

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150–164). Abingdon: Routledge.

Wollack, J. A., & Cizek, G. J. (2016). Real cheating data for researchers: A description of the common datasets from the 'Handbook of quantitative methods for detecting cheating on tests'. Paper presented at the annual meeting of the Conference on Test Security, Cedar Rapids, IA.

Yamamoto, K. Y. (1989). *HYBRID model of IRT and latent class models* (ETS Research Report No. RR-89-41). Princeton, NJ: Educational Testing Service.

Index

Note: *Italicized* page numbers refer to figures, **bold** page numbers refer to tables.