Reza Che Daniels

How Data Quality Affects our Understanding of the Earnings Distribution

Springer

How Data Quality Affects our Understanding of the Earnings Distribution

Reza Che Daniels

How Data Quality Affects our Understanding of the Earnings Distribution

Springer

Reza Che Daniels
School of Economics
University of Cape Town
Rondebosch, South Africa

ISBN 978-981-19-3638-8 ISBN 978-981-19-3639-5 (eBook)
https://doi.org/10.1007/978-981-19-3639-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

How data quality affects our understanding of the earnings distribution.

by Reza Che Daniels

In household surveys, estimating parameters of the earnings distribution is frequently complicated by multiple sources of survey error, often leading to claims of "poor data quality". However, it is not always the case that multiple sources of survey error leads to poor data quality, and knowing the difference between "good data" and "bad data" is important for both research and policy-making purposes. Far too often in my experience, claims of "poor data quality" are based on ignorance at best, intellectual laziness at worst, providing too easy an escape for dealing with the mostly manageable (statistically) repercussions of data production optimisation decisions that must balance cost, timeliness and accuracy.

The 'flip-side' to this is the researcher that makes too brazen a claim about point estimates of paramaters from data that is not intended to measure the kind of outcomes reported on. For individual income data, or employee income data that we shall call "earnings", estimates of poverty and inequality are often estimated with statistically unsound methodologies that leave more doubt about those estimates than inspire confidence. Due to the politically sensitive nature of constructs such as poverty and inequality, it is the responsibility of the research community to provide sound guidance in this respect.

This book is concerned with developing a statistically sound methodology for estimating parameters of the income distribution in household surveys that often contain multiple sources of survey error—some that are observable and some that are unobservable. The country of interest is South Africa, and we focus on a period in South Africa's history just after the transition to democracy in 1994 when the geopolitical boundaries of the country had stabilized, but the best way to measure income was still being debated not only in South Africa (SA), but internationally.

In 1996 the International Expert Group on Household Income Statistics (that became known as the Canberra Group), was formed in response to widespread

recognition of data quality concerns in household income distribution statistics. They published the 2001 Canberra Group Handbook, which became a key reference manual for national statistics agencies. The whole theme of household income statistics was given a greater spotlight by the newly formed (at the time) Millenium Development Goals, the first of which was to "eradicate extreme poverty and hunger" by 2015. Researchers and statisticians interested in creating baseline estimates of poverty around the world turned to the 2001 Canberra Group Handbook for guidance.

We discuss how Statistics South Africa (SSA) deviated from, but later came closer to, key recommendations in the Canberra Manual. Along the way, however, there were important limitations in the data that need to be addressed by researchers interested in analysing this period in SA's history. Two periods are of interest in this book: (1) the immediate post-1994 period during which the geopolitical boundaries of South Africa were recast; and (2) 1997–2003, which is after both the transition to democracy in 1994 and the 1996 national census. The 1996 census was the first time that Statistics South Africa could enumerate the now contiguous geopolitical entity of the democratic country, after reincorporation of the former Apartheid-era "Bantustans" (which were supposed to be independent tribal homelands set up by the Apartheid government and excluded from the definition of South Africa in the *Bantu Homeland Citizenship Act* of 1970). This afforded SSA the opportunity to create a new sampling frame from which to draw more representative samples of the SA population for household surveys.

Between 1997–2003, SSA continued to run the October Household Survey (OHS), which was started in 1993 to obtain general information (including income) from a representative sample of individuals. However, prior to the 1996 Census the OHS had an outdated sampling frame. By isolating the period 1997–2003, we are able to zoom in on a period of important changes to the way income questions were asked. In 2000, SSA discontinued the OHS and commenced the Labour Force Survey (LFS)—a biannual survey that continued until 2008, when it was replaced with the Quarterly Labour Force Survey (QLFS). We restrict the analysis of the LFS to 2000–2003 only, because it captures a period during which questionnaire design changes to the income question stablised.

Over the course of the book we develop a framework that researchers can use to investigate data quality, and then apply this framework to SSA household surveys in the post-Apartheid era. Chapter 1 introduces the book and provides some background to debates in the survey methodology and econometrics communities concerning income statistics and the data quality concerns that must be addressed in order to generate estimates of poverty headcounts and inequality indices.

Chapter 2 develops a framework for investigating microdata quality that is a guide for researchers working with any public-use dataset, often with poor information about the survey quality control process, about how to identify different components of survey error. It is largely based on integrating the total survey error paradigm with data quality metrics that shed light on the possible factors that influence point estimation of key parameters of interest. The framework is then utilised to investigate the evolution of data quality in SSA's labour market surveys.

Chapter 3 isolates questionnaire design and item nonresponse for the employee income question in two South African labour market surveys: the October Household Survey (OHS, 1997–1999) and the Labour Force Survey (LFS, 2000–2003). This time period isolates a period of changing questionnaire design for the income question. Between 1997 and 2000, the employee income question gradually included new response options for the respondent to state that they don't know or refuse to answer the question, which turn out to be important distinctions. We use sequential logistic response models to evaluate how improvements to the income question improved the capacity to understand the nonresponse and bounded response options. We then evaluate the empirical stability of predictors of response type between 1997–2003.

Chapter 4 is concerned with conducting univariate multiple imputation for employee income with nonresponse and bounded responses. A variable with this mixture of data types is called coarse data. Because the income question consists of two parts—an initial, exact income question and a bounded income follow-up question—the resulting statistical distribution of employee income is both continuous and discrete. An analysis of the interrelationship between the exact income and bounded income variables released in the public-use data also reveals a non-trivial degree of processing error for certain survey years between 1997–2003. We identify two forms of processing error that have to be dealt with before multiple imputation can be performed. We then conduct multiple imputation using four differently specified models to test the sensitivity of imputed draws of income to mis-specification in the imputation algorithm. We also evaluate the point estimates of quantiles and moments of the multiply imputed income distributions as the number of imputations increase.

Chapter 5 draws on the lessons learnt in the preceding chapters to identify how data quality will always influence our understanding of the income distribution. We focus on what can be 'fixed', what cannot be, and what might matter for different sorts of analyses. We also generalise the findings in the book so that the methods enumerated can be applied to any household survey concerned with measuring income, anywhere in the world. Chapters three and four *taken in combination* are key to this.

It is my hope that this monograph provides guidance to researchers and data scientists about how to both frame and deal with data quality in microdata, specifically when analysing income and the constructs of poverty and inequality that are so important to policymakers and to measuring socio-economic progress. The methods are reproducible and I'm sure others will improve upon them. This is welcomed. As a research community, let us do what we do, well.

Cape Town, South Africa Reza Che Daniels

Acknowledgements

First and foremost, I would like to thank my family for their constant support.

I would also like to thank my colleagues Murray Leibbrandt and Martin Wittenberg for their valuable advice in the creation of this monograph. Steve Heeringa at the University of Michigan played a seminal role in stimulating my intellectual curiosity in this field, and I am also grateful for his insights in the development of this book.

I cannot stress enough how much being part of the incredible teams at the Southern Africa Labour and Development Research Unit (SALDRU) and DataFirst—both at the University of Cape Town—assisted me by providing a rich intellectual ecosystem for the ideas in this book to be birthed, trialled and tested over multiple years and survey instruments. SALDRU also funded the open access costs of this book, for which I am eternally grateful.

Contents

About the Author

Reza Che Daniels is Associate Professor in the School of Economics, University of Cape Town. He was one of the Principal Investigators of the National Income Dynamics Study (NIDS), South Africa's first nationally representative longitudinal household survey. He is also one of the Principal Investigators of the NIDS-Coronavirus Rapid Mobile Survey (NIDS-CRAM), which uses a sub-sample of the NIDS to monitor the impact of COVID-19 in South Africa.

List of Figures

List of Tables

Chapter 1
Introduction

This book is concerned with the measurement and quality of employee income from household survey (micro) data. The empirical applications are based on South African household surveys compiled by the national statistics agency (Statistics South Africa). Despite this specificity, the insights are generalisable to any household survey concerned with measuring income.

Data quality is a central theme in any data compilation effort. However, it is often very difficult to diagnose where exactly in the data production process data quality falters. Data quality is a concern for both macro- and microdata. For macro-economic data, the International Monetary Fund (IMF) presides over the process of ensuring standards are developed for the production of national economic statistics associated with the System of National Accounts (see IMF, 2003 for the latest such framework). For household survey data, there are data quality frameworks for surveys themselves (see Statistics Canada, 2003, 2009 and Statistics South Africa, 2006a; 2006b), and for specific themes like income.

In all household survey data, several forms of error are present in different magnitudes, including coverage error, sampling error, nonresponse error, adjustment error, processing error, measurement error and validity. These components of error form part of the total survey error paradigm (Groves et al., 2004), and can often be exacerbated by poor data quality management within statistical organisations. To understand data quality therefore requires some understanding of the practises inside statistical organisations with respect to data quality control. Examples of such data quality control elements from Statistics Canada (2003) and Statistics South Africa (2006a) include relevance, timeliness, accessibility, interpretability, coherence, integrity, methodological soundness and accuracy.

For income data measured in household surveys, the Canberra Group's (2001; 2011) recommendations on household income statistics is the main reference. The Canberra Group was a group of national statistics and other data compilation agencies from over fifteen countries, plus representatives from many international agencies, whose main objective was to "... enhance national household income statistics by

R. C. Daniels, *How Data Quality Affects our Understanding of the Earnings Distribution*,
https://doi.org/10.1007/978-981-19-3639-5_1

developing standards on conceptual and practical issues related to the production of income distribution statistics" (Canberra Group, 2001, xi). The global level of importance accorded to this task was noteworthy, for it coincided with the adoption of the Millennium Development Goals, the first of which was to halve absolute poverty, defined as all those living below US$1.00 per day in constant purchasing power parity (PPP) adjusted terms, between 1990 and 2015.[1]

The income distribution has been a central preoccupation of economists since the inception of the discipline due to its positive correlation with individual and societal welfare. An important formalisation of the work on income distributions was made by Vilfredo Pareto in the nineteenth century, who found after analyses of empirical income data on several European countries that the probability distribution of income was right-skewed (Kirman, 2008). More detailed analyses of income distributions since then led to the realisation that several possible statistical distributions have valid application to income over different ranges of the variable (see (Cowell, 2000) for discussion).

As long as people have analysed income distributions there have been debates about the data utilised for this purpose. Income is measured both in the national accounts and with household survey data. However, the methodologies used to collect and aggregate this data renders income measured in the national accounts to be quite a different construct to income measured in household surveys (Havinga et al., 2010). This book is concerned with income measured in household surveys only.

1.1 The Income Construct in Household Surveys

Generally, when income distribution is discussed, the debate concerns the distribution of *total* income. But total income is comprised of many components. The Canberra Group (2001, 18) distinguish the following types of income that together sum to total income:

- Employee income, plus
- Income from self-employment, plus
- Income from rentals, plus
- Property income, plus
- Current transfers received.

This book is primarily concerned with employee income. Employee income is considered to be a form of cash income that is easily and accurately measured relative to property income and cash transfers (Canberra Group 2000, 13). However, the employee income question in household surveys is complicated by a feature that is designed to increase the probability that a respondent answers the question. That is, a second, bounded income bracket question is presented to respondents as a follow-up to the exact income question in the event that they refuse to answer or state that

[1] See http://www.un.org/millenniumgoals/poverty.shtml.

they don't know. This leads to an income variable with a continuous distribution for exact income responses and a discrete, grouped continuous distribution for bounded income bracket responses.

Respondents can also refuse to answer the follow-up question, or once again state that they don't know their income or that of the proxy respondent on whose behalf they are reporting. Consequently, there is also nonresponse to the employee income question. How researchers treat the many issues that confront them with income data in public-use household surveys can often be very different, leading to different estimates of parameters of the income distribution from the same dataset.

The advantage of having a follow-up income question with a lower level of information disclosure is that it reduces the social sensitivity of the question, but can also aid respondent recall. Consequently, some form of follow-up question that bounds the range of income is often also asked in household surveys for other components income, including income from self-employment, rentals, property and transfers. Therefore, while the emphasis in this book is on employee income, the insights are generalisable methodologically to any component of income that is measured in a similar way.

The overall quality of household surveys also has an important bearing on the accuracy of individual income statistics. In South Africa (SA), nationally representative sample surveys have only been compiled by Statistics South Africa (SSA) since the early 1990s. Before 1994, the geopolitical borders of SA included the Bantustans, considered separate by the Apartheid Government to the state of SA. Consequently, in the national statistics community in the mid 1990s, more emphasis was placed on creating new sampling frames for the democratic SA than refining questionnaire design for constructs like employee income. This necessary trade-off in the data production process led to poorer quality income data initially that gradually improved as other operational aspects of the household surveys themselves improved.

1.2 Objectives and Chapter Typology

The main objectives of this book are:

- To develop a framework for investigating microdata quality and apply this framework to South African labour market household surveys that include a question on employee income.
- To investigate the relationship between questionnaire design for employee income and the respondents who choose to answer the question in different ways (including bounded income bracket responses, refusals and don't know responses).
- To formulate practicable solutions for researchers concerned with generating a derived employee income variable from public-use income variables with varying degrees of coarseness, using multiple imputation for this purpose.

Chapter Two is directed at understanding the universe of errors that can arise in household surveys and linking these to data quality protocols inside statistical

organisations. It identifies specific data quality metrics for each component of survey error that can arise. It then applies this framework to South African labour market household surveys. The chapter provides a general taxonomy for investigating data quality that can be useful to researchers whose aim it is to understand the relationship between survey error and data quality in public-use datasets. In order to demonstrate this, the individual income variable is reviewed for the employed population of South Africa, evaluated over multiple survey instruments and time periods, ranging from 1995–2007.

Chapter Three in this book isolates the design of the employee income question in household surveys and the propensity of respondents to provide a particular response type. Employee income is typically measured in a way that allows respondents to provide either an exact income value or an interval into which it falls. It is then the user's responsibility to generate a variable that combines these two response types appropriately. However, missing data is also present when respondents refuse to answer the question or state they don't know. Understanding the different subsets of respondents sheds light on the trade-offs of questionnaire design for employee income, and provides valuable insight into the response process that can inform single and multiple imputation exercises.

The final substantive chapter then goes on to investigate public-use employee income data with a mixture of continuously distributed income observations, grouped-continuous observations and item nonresponse. This mixture of data types is called coarse data in the literature, and has important implications for imputing plausible values for such data. In SSA's household surveys, we also find that there is a non-trivial degree of processing error in the two income variables released in the public-use dataset that must be treated appropriately before multiple imputation exercises can commence. We then conduct multiple imputation and discuss several aspects of the imputation algorithm – from the estimation method, to constraints on the bounds of the plausible draws, to the specification of prediction equations – all of which have a bearing on the reliability of imputed draws. Once these concerns have been addressed, multiple univariate imputations of employment income from coarse data can be obtained in a manner that allows researchers to account for the greater uncertainty inherent in that data. This then allows for the reliable estimation of univariate parameters of the income distribution.

The time-frame for the analysis spans the mid 1990s to the latter part of the 2000s for Chapter Two. However, for Chapters Three and Four, the time-frame is restricted to 1997–2003. This is because this period was associated with important changes in the way household surveys were conducted in Statistics South Africa. Between 1995–1999, the October Household Survey was a repeated cross-sectional survey that collected labour market data as well as more general household information. From 2000 onwards, this survey was split into the Labour Force Survey (LFS) and the General Household Survey. Only the LFS is analysed in this book. This allows us to identify the role of questionnaire design in improving the quality of income data.

The LFS was designed as a rotating panel survey whose explicit purpose was to obtain accurate estimates of employment and unemployment. In Chapters Three

and Four, only the September Waves of the Labour Force Survey are analysed in conjunction with the OHS 1997–1999. Because the LFS is a *rotating panel* household survey (see Cantwell, 2008 for a definition), a proportion of the respondents change with each Wave of the survey, ensuring that it is representative of the South African population at the time of going to field. Therefore, it is possible to analyse the cross-sectional OHSs in combination with individual waves of the rotating LFS panel.

The final chapter in this book concludes the discussion. Since each chapter contributes original insight into different aspects of data production and use, the Conclusion stresses the need to factor all of the issues discussed in this book into an overall set of guidelines for estimating parameters of the income distribution. The discussions in chapters three and four, in combination, provide particularly powerful insights about how to ultimately derive reliable points estimates about poverty and inequality.

References

Canberra Group. (2001). *Expert group on household income statistics: Final report and recommendations*. Ottawa: The Canberra Group

Canberra Group. (2011). *Canberra group handbook on household income statistics* (2nd ed.). Geneva: United Nations Economic Commission for Europe.

Cantwell, P. J. (2008). Rotating Panel Design. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods.* Thousand Oaks: Sage Publications.

Cowell, F. A. (2000). Measurement of inequality. In A. B. Atkinson & F. Bourguignon (Eds.), *Handbook of income distribution* (Vol. 1). New York: Elsevier.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley Press.

Havinga, I., Kamanou, G., & Vu, V. (2010). A note on the (mis)use of national accounts for estimation of household final consumption expenditures for poverty measures. In S. Anand, P. Segal, & J. E. Stiglitz (Eds.), *Debates on the measurement of global poverty.* London: Oxford University Press.

International Monetary Fund (IMF). (2003). Data quality assessment framework—Generic framework. IMF, Washington D.C.: Mimeo.

Kirman, A. (2008). Pareto, Vilfredo (1848–1923). In S. N. Durlauf & L. E. Blume (Eds.), *The new palgrave dictionary of economics* (2nd ed.). Palgrave Macmillan.

SSA. (2006a). *Draft data quality framework 001: South African statistical quality assessment framework*. Pretoria: SSA.

SSA. (2006b). *Data quality policy 001: Policy on informing users of data quality*. Pretoria: SSA.

Statistics Canada. (2003). *Statistics Canada quality guidelines* (4th ed.). Ottawa: Statistics Canada.

Statistics Canada. (2009). *Statistics Canada quality guidelines* (5th ed.). Ottawa: Statistics Canada.

Chapter 2
A Framework for Investigating Microdata Quality, with Application to South African Labour Market Household Surveys

2.1 Introduction

This chapter identifies a framework for investigating microdata quality that is particularly useful to researchers working with public-use micro datasets where limited information about the data quality protocols of the survey organisation are present. It then utilises this framework to investigate South African labour market household surveys from the mid 1990s to 2007. In order to develop the framework, we rely on the total survey error (TSE) framework to articulate the forms of statistical imprecision that exist in any public-use dataset. The magnitudes of statistical imprecision are largely dependent on the efficacy of the survey organisation's data quality control protocols, which are, in turn, affected by human resource and budget constraints.

The objective of this chapter is to provide researchers with the tools needed to assess the quality in public-use datasets, to the extent that components of survey error are identifiable. Researchers will always have imperfect information in this regard, yet in South Africa at least, this has not stopped both the academic community and policymakers from making public statements about data quality that are often ill-informed and frequently incorrect.

The choice of time-period to investigate microdata quality in South Africa (SA) coincides with a period of profound change in the country associated with the transition to democracy in 1994. Geopolitical changes included the provincial boundaries within SA and the incorporation of former Bantustans, which were previously "homelands" for Black South Africans (some of which were self-governing) created by the Apartheid government. The national statistics agency (Statistics SA) therefore had to increase the scope of their operations and develop new sampling frames. Over time, new surveys were conducted and gradually more attention was devoted to the quality of the data and sophistication of the survey instruments.

The October Household Survey (OHS) was the first household survey conducted in democratic South Africa to include a labour market component, and officially started in 1993. However, both the 1993 and 1994 versions of the survey have magnitudes of survey error that have resulted in very few researchers utilising them (see Wittenberg,

R. C. Daniels, *How Data Quality Affects our Understanding of the Earnings Distribution*,
https://doi.org/10.1007/978-981-19-3639-5_2

2006 for discussion). We therefore commence with the OHS 1995 to OHS 1999. The Labour Force Survey (LFS) replaced the OHS as the labour market survey for SA in 2000. We analyse the data from the LFS until 2007, whereafter it became the Quarterly LFS and changed in frequency and design.

In order to understand what was going on inside the national statistics agency in the mid 1990s, a qualitative interview with a retired sampling statistician (Professor David Stoker) was conducted (see Daniels and Wittenberg, 2010). Prof Stoker worked in Statistics SA (SSA) in various capacities from the late 1980s until the early 2000s, and was in a unique position to shed light on the data quality pressures facing SSA over the time period. Information from this interview is supplemented by the survey Metadata and other survey documentation released to the public by SSA in each year of the OHSs and LFSs. In narrating these issues, a valuable historical record has been created of microdata quality in South Africa during one of the most fascinating periods in the country's history.

The rest of this chapter proceeds as follows. Firstly, we discuss the importance of framing data quality debates such that they do justice to both data production (the perspective of the survey organisation) and data consumption (the perspective of the researcher). Then we consider the interaction between specific data quality elements and components of survey error. This creates the framework for investigating microdata quality. We then apply this framework to SA labour market household surveys from 1995 to 2007. Lastly, we discuss the generalisability of the framework and its scope for application to other surveys and countries.

2.2 Framing the Discourse on Data Quality

Microdata quality is an artifact of a data production process controlled by survey organisations with finite budgets. This data production process commences with the conception of a project and concludes with public release of the data. Consumers of data (researchers) become concerned with data quality in the public-use dataset when it becomes apparent that univariate, bivariate and/or multivariate distributions in the data are problematic. This means that both the production and consumption dimensions of microdata need to be considered when attempting to create a framework for investigating microdata quality.

In this section we locate the discourse of creating a framework for microdata quality at the nexus of the data production and consumption process, i.e. when considering parameters of interest on variables released in a public-use dataset. Researchers only observe the final product released by the statistical organisation, and so do not have the information to make accurate judgments about where in the data production process data quality falters. However, they can see inconsistencies in the statistical distributions of variables of interest that often hint at poor data quality. Survey organisations, on the other hand, rarely consider bivariate and multivariate relationships before publishing the data, and so often miss the insights researchers glean as users of the data.

Below we define data quality elements in the data production process. This helps clarify the context in which survey organisations operate. Then we discuss a taxonomy of statistical errors in the survey process encapsulated by the total survey error (TSE) framework. TSE has proved itself useful to survey organisations to guide an understanding of the relationship between data quality and sources of statistical error. For researchers, the TSE framework is useful as a conceptual map to think more clearly about data quality in public-use datasets.

2.2.1 Data Quality Elements in the Data Production Process

Data quality management, evaluation and reporting has become an increasingly important issue to statistical organisations and (inter)national agencies tasked with generating or compiling information for third-party users. In turn, for users of the data, understanding data quality necessitates an understanding of the processes leading up to public release. Formal recognition of the need for data quality indicators has been acknowledged in the broader statistical community for some time. Recent efforts by the economics community with respect to microdata quality has also raised the primacy of this debate (see Flinn et al., 2001).

Brackstone (1999) identifies six dimensions of data quality: relevance, accuracy, timeliness, accessibility, interpretability, and coherence. Underlying these six dimensions is the idea that the data ought to be 'fit for use'. "Fitness for use encompasses not only the statistical quality concepts of variance and bias, but also other characteristics such as relevance and timeliness that determine how effectively statistical information can be used" (StatCan, 2003, 6). These ideas have become the bases for many national statistical organisations developing data quality manuals, such as Statistics Canada (2003, 2009). Statistics South Africa (2009, 2010) define two additional dimensions of data quality, namely methodological soundness and integrity (SSA, 2010). These two additional qualities hint at resource constraints (particularly human resource constraints) that may be more binding in developing countries. However, they are not necessarily separate from Brackstone's data quality concerns and can in fact be considered to be fully nested within them.

Brackstone's (1999: 143) six themes are worth elaborating: "relevance" refers to the degree to which statistical information meets the needs of users or clients; "accuracy" refers to the degree to which the information correctly describes the phenomena it was designed to measure, and includes such concepts as mean square error; "timeliness" refers to the delay between the reference period and the date of public release, and typically involves a trade-off against accuracy; "accessibility" refers to the ease with which users can obtain the information; "interpretability" refers to the availability of the supplementary information and metadata necessary to interpret and use the data correctly; and "coherence" refers to the degree to which it can be successfully brought together with other statistical information within a broad analytical framework and over time.

These components of data quality are resource-dependent, and for a well funded statistical organisation like Statistics Canada (which Brackstone, 1999 based his work on), the scope to invest in each of these dimensions of data quality is high. That said, Groves (2004) and Heeringa and Groves (2006) note that regardless of the size of resources available, there is always an optimisation problem when it comes to maximising data quality with a finite budget. But the size of the budget itself is not trivial. In fact, in low-income countries survey operations in national statistical offices can be severely restricted due to very small budgets (compared to their more well funded high-income country counterparts). Glewwe (2005) notes that in developing countries, these constraints imply that more careful planning is needed before a survey goes to field in activities such as drafting budgets and securing financing, developing a work plan for remaining activities, drawing a sample of households to be interviewed, writing training manuals, training field and data entry staff, preparing fieldwork and data entry plans, conducting pilot tests and launching publicity campaigns. Data quality concerns must therefore also be considered within the environment in which statistical organisations function.

2.2.2 The Total Survey Error (TSE) Framework

The TSE framework can be used as a taxonomy to understand the scope of potential error sources in a micro dataset. The determinants of data quality are principally under the control of the survey organisation, where conscious effort needs to be invested in each step of the survey process in order to manage the quality of the data obtained. When the data finally get to a stage ready for public release, certain forms of survey error may still be present in the data. It is then up to researchers to identify if, how and when any remaining sources of error will affect their analyses. But researchers do not have the necessary auxiliary information to diagnose all forms of survey error precisely. This is exacerbated when survey organisations release poor documentation with public-use datasets. Under these circumstances, researchers can often face grave doubts about whether their analytical results are indeed valid or if they are rather an outcome of an unreliable data generating process.

Components of survey error can generally be split into two forms: errors of observation and errors of nonobservation. Errors of nonobservation are those arising because measurements were not taken on part of the population, whereas observational errors are deviations of the answers of respondents from their true values (Groves, 1991, 2). In line with this, the TSE framework disaggregates the components of error into two themes: (1) measurement of the variable of interest, and (2) representation of the population of interest. Figure 2.1 presents a schematic overview of TSE.

Under the measurement theme, the possible sources of error include validity of the construct, measurement error and processing error. For the representation theme, the sources of error include coverage error, sampling error, nonresponse error, and adjustment error. Because researchers and survey organisations frame the concept

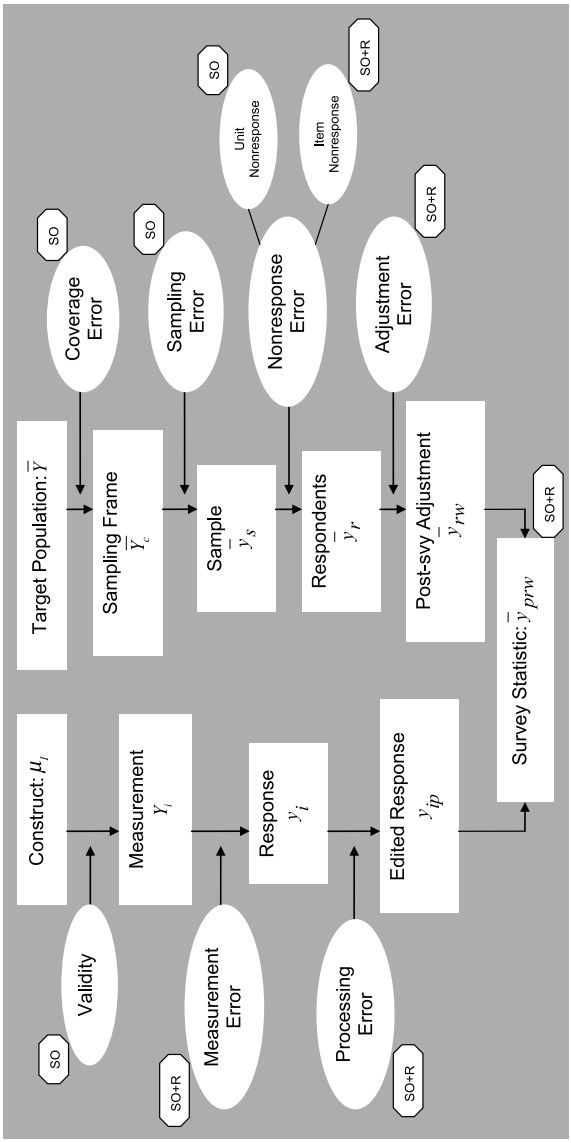

Fig. 2.1 Agency (i.e. Survey Organisation (SO), Researcher (R)) in the total survey error framework. Source Adapted from Groves, Fowler, Couper, Lepkowski, Singer & Tourangeau, 2004, 48

of data quality differently, it is helpful to consider the agency of these two groups in the TSE framework.

A few terms in the figure require explanation (taken from Groves, 2004, vi). Coverage error stems from the failure to give some person or group of persons any chance of inclusion in the survey sample. Non response error stems from the failure to collect data on all persons in the sample, while sampling error arises from differences in the survey sample compared to the population it is trying to measure. Measurement error stems from inaccuracies in responses recorded on the survey instruments, and can be attributable to four different components: (a) effects of interviewers on the respondent's answers to survey questions; (b) error due to respondent's inability to answer questions, lack of effort, or other psychological factors; (c) error due to weaknesses in the wording of survey questionnaires; and (d) error due to effects of the mode of data collection (e.g. face-to-face surveys, telephone surveys).

Non response error can be split into unit nonresponse (meaning entire sampling units refuse to participate in the survey) and item nonresponse (meaning an individual responds to some questions in the questionnaire, but not to others). End-users of the data are unable to deal with unit nonresponse, but are able to deal with item nonresponse, where single and multiple imputation methods become applicable given a plausible model about the response mechanism.

Adjustment error arises out of the need to adjust the survey for coverage error, sampling error and (unit) nonresponse error. Typically this is done by calculating weights. In South Africa, survey organisations usually combine individual weights into a single weight that is included in the public release version of the dataset. When this is the case, researchers are unable to separate out the components of the weight, and so are left without the means to investigate how each weight was calculated.

From Fig. 2.1 we can see that on the measurement side of the TSE framework, researchers have insight into processing error and certain forms of measurement error. However, it is unusual that any informed insight can be gleaned about construct validity in public use datasets—certainly insofar as understanding the sensitivity of question wording on outcomes is concerned, which would be part of the question pre-testing phase presided over by the survey organisation. Cases where researchers are able to directly engage with construct validity do exist though, especially when appraising whether a questionnaire accurately captures some externally defined construct, such as (broad or narrow) unemployment or the informal sector.

On the representation side of the TSE framework, item nonresponse and adjustment error are the two components that researchers can gain some insight into. Item nonresponse can be imputed by either the researcher or the survey organistion, but adjustment error is usually the domain of the survey organisation. However, there are circumstances when researchers are able to identify whether errors have been made in the adjustment process. In South Africa, Branson and Wittenberg (2007, 2011) and Branson (2009) have analysed the weights in Statistics SA's labour market household surveys and found several inconsistencies.

Finally, it is incumbent upon both the survey organisation and the researcher to compute final survey statistics appropriately. It is the former's responsibility to provide all the documentation, weights and survey design features (such as vari-

ables used to stratify, cluster and make finite population corrections) necessary for researchers to generate accurate point estimates from public-released data. It is then the researcher's responsibility to account for survey design features in their univariate, bivariate and multivariate analyses (for example, see Daniels and Rospabe, 2005).

2.3 The Interaction Between TSE and Data Quality

While the TSE framework provides data users with a quick schematic overview of potential error sources, the data quality controls within survey organisations provide insight into the protocols for data production that can have a direct bearing on the overall quality of public-use data. In this section we demonstrate how data quality guidelines interact with the TSE framework. We use two editions of "Statistics Canada Quality Guidelines" (2003, 2009) to inform the discussion, as well as two editions of Statistics South Africa's Statistical Quality Assessment Framework (SSA, 2009, 2010). We can summarise the relationship between the TSE framework and components of data quality as per Fig. 2.2.

From Fig. 2.2 we see that the concept of "accuracy" is what brings together both the TSE framework and the functional operational concerns of the Survey Organisation. Indeed, Statistics Canada (2003, 6) note that the very purpose of publishing quality guidelines is to inform the debate on "how to assure quality through effective and appropriate design or redesign of a statistical project or program from inception through to data evaluation, dissemination and documentation." Below we elaborate on how each component of the TSE framework interacts with components of data quality.

2.3.1 Validity of the Construct of Interest

In the TSE framework, validity is defined as the observational gap between constructs and measurements (Groves et al., 2004, 50). In other words, validity is concerned with how well the survey instrument measures the construct of interest. In statistical terms, the notion of validity acknowledges two sources of variability—one at the level of the individual respondent and another at the level of different trials of the survey (ibid, 50).

From a data quality perspective, it is very difficult to know a-priori how valid a particular construct may be over different trials of the survey. It is also very expensive to run multiple trials of a survey simply to obtain sufficient data to be able to estimate this. However, it is possible to assess how respondents' responses may vary given a different phrasing or wording of the survey questions for example. This is the idea behind pre-testing questionnaires, which can span any number of different dimensions from wording a particular question differently and testing whether respondents

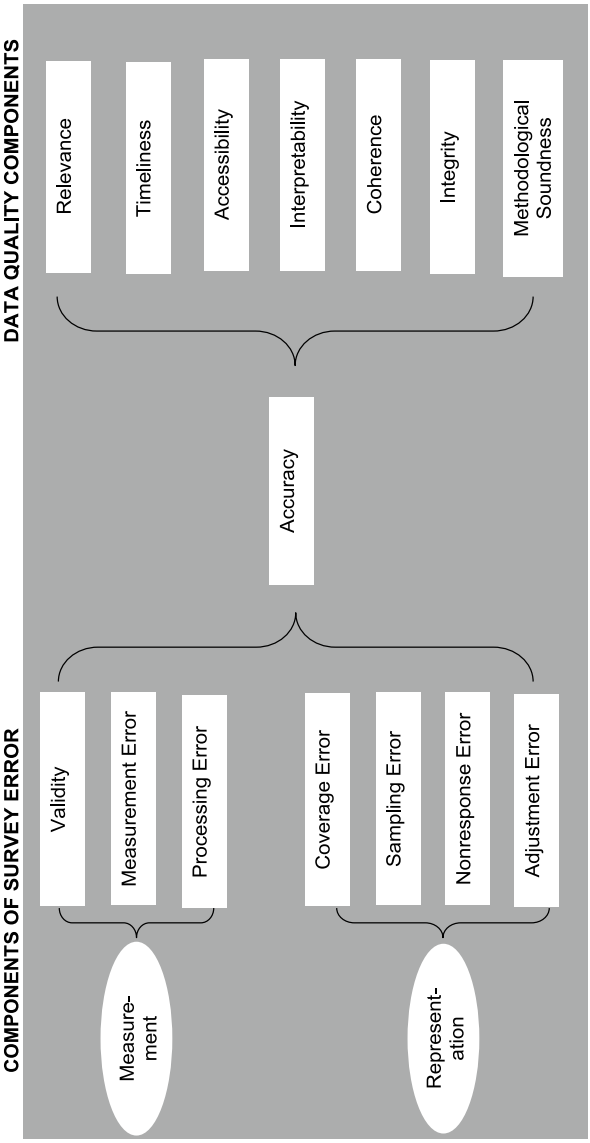

Fig. 2.2 The relationship between total survey error and components of data quality

respond differently, to translating questionnaires into different languages and conducting similar diagnostic exercises. Questionnaire design is thus partly relevant to the idea of validity. Pre-testing questionnaires can aid the understanding of both validity and measurement error.

To concretise the discussion, consider the construct validity of income. From a practical point of view, income can refer to many different sources. Thus the validity of income has to do with everything from the component of income being measured to the scope of income (i.e. whether that income is an individual or household measure). Different types of income measurements in a household survey include employee income, income from self-employment, rental income, property income and income from transfers (Canberra Group, 2001, 2011). Household surveys in South Africa that measure all of these types of income include the Income and Expenditure Surveys (SSA, 1995, 2000, 2005) and the National Income Dynamics Survey (SALDRU, 2008, 2010–2011).

The main data quality elements associated with validity are relevance. The process of transcribing the constructs of interest to the questionnaire is a very important part of any survey.

2.3.2 Measurement Error

Measurement error is defined as the observational gap between the ideal measurement and the response obtained (Groves et al., 2004, 51). The "error" component implies a departure from the true value of the measurement as applied to a sample unit and the value provided (ibid, 52).

The effects of different sources of measurement error can be very difficult (and sometimes impossible) for researchers to identify in public-use datasets. For example, Wittenberg (2004) notes that in trying to measure the occupational distribution of manufacturing sub-sector employment in South Africa using the Manufacturing Census, the Population Census and the October Household Surveys, one of several possible explanations of divergences in the point estimates could be due to fieldworker errors. The difficulty here though lies in the inability of researchers to precisely determine the potential sources of the problems, for Wittenberg (ibid) also notes that the discrepancies discovered could have been due to a range of other factors, all of which can only be speculated upon when investigating the empirical magnitudes.

On the other hand, changes in questionnaire wording are precisely identifiable by researchers given careful analysis. For example, Bhorat (1999) noted that the definition of the informal sector in the October Household Surveys 1995 was problematic. This changed in later years of the survey, but in so doing Yu (2009) made the point that it made time-series analyses of the repeated cross-sections of informal sector workers problematic. Yu (2007) notes that the manner in which broad and narrow unemployment rates were measured also changed across survey years, and that these kinds of changes to questionnaire wording impose important trade-offs.

Due to the multidimensional nature of measurement error, data quality guidelines need to be developed for each possible source of error. Groves (2004, 359) notes that when considering the interviewer as a source of measurement error, it is crucial to understand the manner in which they can affect the survey. It is also possible (and necessary) to monitor the results of interviewers as close to real time as possible. When developing indicators to assess interviewer variance in household interview surveys for example, Groves (ibid, 364–5) discusses Kish's (1965) original interviewer intraclass correlation coefficient, which is the ratio of variance between interviewers to the total variance of a measure. This is a very direct way to assess interviewer performance, and can aid the discussion of measurement error when it becomes apparent that certain interviewers behave erratically (e.g. submit completed questionnaires with identical values for many questions).

The respondent is also a source of measurement error, and the manner in which errors can be introduced by the respondent are numerous. Groves (2004, 407–408) notes that from models of the interview process and newer cognitive science perspectives, there are five stages of action relevant to survey measurement error, including: (1) how the respondent encodes (processes and stores) the information asked of him/her; (2) how the respondent comprehends the question; (3) how the respondent retrieves the information; (4) how the respondent judges the appropriate answer to provide the interviewer with; and (5) how the respondent communicates the information to the interviewer. Clearly the relationship between the interviewer and the respondent is important here, and this reiterates the need for interviewer training and possible matching of interviewers to respondents on socio-cultural grounds (such as race or language).

The importance of designing a sound questionnaire is related to the discussion above in that it has an impact not only on the influence and image of a statistical agency, but also, from a data quality perspective, on respondent behaviour, interviewer performance, collection costs and respondent relations (StatCan, 2009, 28). The principles for designing a questionnaire include that it should collect data that corresponds to the survey's Statement of Objectives while taking into account the statistical requirements of data users, administrative and data processing requirements as well as the nature and characteristics of the respondent population. Furthermore, it should flow smoothly from one question to the next, facilitate respondents' recall, facilitate the coding and capture of data, minimise the amount of editing and imputation that is required, and lead to an overall reduction in the cost and time associated with data collection and processing (ibid, 28).

There are consequently several different data quality elements involved for this source of error, including accuracy, methodological soundness, coherence and relevance. All of these must be managed effectively in order to minimise measurement error in public-use data.

2.3.3 Processing Error

Processing error is defined as the observational gap between the variable used in estimation and that provided by the respondent (Groves et al., 2004, 53). Processing error is about data collection, capture and coding. These operations use a large portion of the survey budget, requiring considerable human and physical resources as well as time (StatCan, 2009, 32). Depending on the degree of automation of these tasks, there can also be a large amount of paradata (e.g. indicators of whether or not a unit is in the sample, history of visits, mode of data collection, administrative information and cost information) generated in this process (ibid, 32).

In the evolution of SSA's household surveys, there are many instances of processing error. For example, Yu (2007) identifies inconsistencies with several variables related to earnings, such as work experience and hours worked, which have some values greater than logical upper bounds (though, alternatively, this could be a source of measurement error if the respondent or interviewer was the source of the information). Yu (2007) also identifies coding inconsistencies with race, marital status and education in several October Household Surveys (ibid). Processing error also exists in the component statistical files of the publicly-released OHS 1998, where some observations are repeated in the person file but absent in the worker file (ibid). These examples demonstrate an important feedback loop on data quality from researchers to the survey organisation. It is rare that the survey organisation will be able to pick up errors of this nature in a set of routine checks, but researchers who are concerned with very specific issues relating to the data will.

The main data quality element involved in data capture, collection and coding is accuracy (StatCan, 2009, 37). The key principle guiding data collection is to minimise the burden on the respondent while ensuring privacy and security of the information provided in all data gathering and processing operations (ibid, 32). Because these operations have a high impact on data accuracy, quality and performance measurement tools should be used to manage the collection, capture and coding processes within the survey organisation (ibid, 32).

While these principles point to explicit guidelines for data capture, collection and coding, the degree of success in minimising processing error is rarely perfect (see StatCan, 2009, 32–36). Newer forms of technology (e.g. computer assisted interviewing software) can aid the degree to which the process is minimised, but whenever there is a human element involved there is the scope for making mistakes.

2.3.4 Coverage Error

Coverage error is defined as the nonobservational gap between the target population and the sampling frame (Groves et al., 2004, 54). Coverage itself is the completeness of the information for the target population that would be derived if all of the frame units were to be surveyed (StatCan, 2009, 19). Coverage errors include missing in

scope units, included out-of-scope units, misclassified units and duplicates. Coverage errors therefore are a function of both frame undercoverage (or overcoverage) and differences in the survey estimate for those actually covered from those for which an estimate is required (ibid, 19).

Coverage error is a particularly important source of error in poorer countries or countries in transition, where the geopolitical units may be new or changing. South Africa during the mid 1990s is such an example, where the names and internal geopolitical boundaries of provinces were redefined more than once in the 1990s. Furthermore, in poorer countries national statistical agencies often have more limited budgets, and the capacity to keep sampling frames up to date is more limited (Yansaneh, 2005). There are international organisations that can assist statistical organisations in these countries with optimising resources for improved frame maintenance and sample selection, such as the United Nations Statistics Division (see "Development of National Statistical Systems", UNSD, 2011). For cost minimisation purposes, master sampling frames combined with master samples are frequently advocated for statistical organisations with limited resources (see Pettersson, 2005). These are methods that generate frames and samples to be used in many different surveys by the same statistical organisation over time.

The data quality elements that arise for coverage error pertain largely to the degree to which the sampling frame accurately captures the target population; hence, accuracy and relevance are the key elements (StatCan, 2009, 21). For survey organisations, this means that sampling frames need to be well designed and kept up to date. Certain countries have very different conventions on the type of information that can be stored by public statistical agencies. For example, in Sweden there is a population register and an updated list of names and addresses for almost all residents, whereas in the USA the population is so large that telephone numbers are often used as frames (Groves et al., 2004, 55). The specific type of coverage errors that can arise therefore also depend on the country, its population size (or number of firms in the event of enterprise surveys), and the degree to which information can be stored about individuals.

An important relationship between coverage and frames is to ensure that the survey population is reasonably consistent with the target population on the one hand, and that the frame then conforms to the survey population on the other (StatCan, 2009, 19). Coverage error can reduce the degree to which the frame and the survey populations match and can result in cost increases, loss of timeliness and a diminished accuracy of the estimates from a bias and variance point of view (ibid, 19). Consequently survey organisations need to implement procedures to minimise this discrepancy. Contemporary ways of doing this include using remote sensing and satellite imagery.

2.3.5 Sampling Error

Sampling error is defined as the nonobservational gap between the sampling frame and the realised sample (Groves et al., 2004, 57). Sampling error consists of two

components, namely sampling variance and sampling bias (Krotki, 2008). Sampling variance is the part that can be controlled by sample design factors such as sample size, clustering strategies, stratification, and estimation procedures (ibid, 2012).

Sampling is a means of selecting a subset of units from a target population for the purpose of collecting information that can be used to draw inferences about the population as a whole (StatCan, 2009, 23). The sample design encompasses all aspects of how to group units on the frame, determine the sample size, allocate the sample to the various classifications of frame units, and select the sample (ibid, 23). Sample designs are either probability-based or non-probability based, the latter being generally fast, easy and inexpensive to undertake (ibid, 23). Some of the principles for dealing with probability-based sample designs include that it should be as simple as possible within the context of a design that (1) is based on randomisation, (2) has population units that have a known positive probability of being selected, and (3) has calculable selection probabilities (ibid, 23).

When probability-based samples are designed to be used for more than one survey, i.e. when dwelling units or clusters of dwellings on the same sampling frame are reserved for use in future surveys, then that kind of sample is known as a master sample. Master samples are frequently used in developing countries for cost reduction purposes and to ensure that investments in creating probability-based designs can be utilised for more than one survey (Pettersson, 2005).

An important data quality element associated with sampling is accuracy (StatCan, 2009, 26). This means that every decision that is made about the survey needs to be thought about in relation to how well the sample represents the population. The size of the sample is also important in reducing sampling error. This point naturally extends to subsample sizes that may be necessary to obtain representivity at geographical levels smaller than the nation state (e.g. provincial and/or urban-rural representation). The variables of interest in the survey are also important. For example, to obtain provincially representative statistics on poverty requires that sufficiently large enough samples are drawn for the population groups that are most likely to live in poverty in those provinces.

The design of the sample needs to balance accuracy within the budget constraint. Multi-stage complex samples are therefore the norm when it comes to probability-based surveys, and will include careful thought about stratification, primary sampling units, clusters, weights and design effects from previous surveys that may aid sample size considerations for current surveys (StatCan, 2009, 25). If the survey is a rotating panel, then the sample needs to be designed to account for rotation, whereas if it is a periodic survey, then the sampling process can be a simpler process. Attrition in any panel survey further complicates sampling error, and needs to be carefully monitored as the panel progresses over time.

The importance of survey documentation that correctly reflects the choices that were made and the problems that were encountered then becomes key, since it records and catalogues the information needed to understand the trade-offs of decisions that affect the accuracy of the outcomes.

2.3.6 Nonresponse Error

Nonresponse error is defined as the nonobservational gap between the sample and the respondent pool (Groves et al., 2004, 58). "Nonresponse error arises when the values of statistics computed based only on respondent data differ from those based on the entire sample data" (ibid, 59). Nonresponse can be split into two components: unit nonresponse and item nonresponse. Unit nonresponse is when an entire sampling unit (e.g. individual, household or firm) does not participate in the survey because they could not be contacted or refused to participate in the survey for some reason. Item nonresponse is when a particular question in the questionnaire is not answered by the respondent, either because the respondent refused to answer the question or because the interviewer failed to ask the question.

The main data quality element involved in nonresponse error is accuracy (StatCan, 2009, 49). Nonresponse can have two effects on data: (1) it biases estimates when nonrespondents differ from respondents; and (2) it increases the variance of estimates because the sample size is reduced (ibid, 46). It is therefore important to understand what has become known as the nonresponse mechanism, i.e. the process that leads to nonresponse. For unit nonresponse, the degree of effort expended by the survey organisation on minimising non-contacts and refusals to participate in the survey is key to reducing its incidence. This has budgetary implications, so unless the survey organisation explicitly allocates resources for this process, the degree to which they understand the unit nonresponse mechanism is compromised. Depending on the survey, if no effort is invested in following up unit nonrespondents, then it is frequently addressed by reweighting the data.

The basic ideas behind nonresponse were developed by Rubin (1976, 1987), as were a set of solution methods based on imputation strategies of various forms. The key idea behind nonresponse analyses is to establish whether the process that leads to missing data can be ignored. Ignorability refers to a property that permits the survey organisation (in the case of unit nonresponse or item nonresponse) or the researcher (in the case of item nonresponse only) to *not* take explicit account of the process that leads to missing data when conducting analyses. Ignorability was first developed as a condition for missing data by Rubin (1976, 1987), and helped distinguish the conditions of missing completely at random (MCAR—what Rubin, 1976 originally called Observed at Random), missing at random (MAR), and not missing at random (NMAR).

For item nonresponse, understanding the response mechanisms amounts to determining whether the missing data are missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). Statistics Canada (ibid, 46) define these "classic" response mechanisms as follows: uniform nonresponse is an MCAR mechanisms where the response probability is completely independent of the units and the measurement process, and is constant over the entire population; nonresponse depending on an auxiliary variable is a MAR mechanism where response depends on certain auxiliary data or variables available for all units measured; and

nonresponse depending on the variable of interest is a NMAR mechanism where the response probability depends on the variable of interest.

The principles for dealing with nonresponse in a survey are related to budget, time and staff constraints, the impact on overall quality and the risk of nonresponse bias (ibid, 46). It is also dependent on the mode of the survey (e.g. personal interview, telephonic), auxiliary information for respondents, an effective respondent relations programme, a well designed questionnaire, and the use of active management to ensure regular follow-up on collection operations and adaptive data collection (ibid, 46).

For researchers, dealing with item nonresponse often involves reweighting or imputation methods. The latter ought to be based on careful analyses of the response mechanism in a manner analagous to how survey organisations investigate unit nonresponse (this is the focus of the next chapter in this book). This allows the item response process to be understood using the same general methods for understanding unit response.

2.3.7 Adjustment Error

Adjustment error is defined as the discrepancy between the sample of respondents and the post-survey adjustments necessary to ensure the sample represents the population of interest. These adjustments are efforts to improve the sample estimate in the face of coverage, sampling and nonresponse errors, and use some information about the target or frame population or response rate information on the sample to make adjustments (Groves et al., 2004, 59). Adjustments are usually made by creating appropriate weights, so the data quality concerns associated with adjustment error pertain to weighting and estimation. The key data quality element associated with adjustment error is accuracy (StatCan, 2009, 61).

The three reasonably standard weights associated with probability-based surveys are probability of selection weights, unit nonresponse and post stratification weights. The first weights observations in the survey by the inverse of their probability of selection. The second assigns a weight to missing units relative to observed units that match some known characteristics between the two (e.g. cluster, psu location). Post-stratification weights adjust demographic survey population totals in a given survey period to the most recent national demographic population totals on record. These weights can then be multiplied together to obtain a composite weight for each observation in the survey that will be included with the publicly released dataset.

The principles associated with creating weights and correct estimation procedures that affect adjustment error depend on the type of weight produced and the method by which the weights get accounted for in the estimation process. Accurate information at the sampling and response stages of the survey help with the creation of sampling and unit nonresponse weights. Sampling weights need to reflect the sample design, so if a multi-stage design has been used (including stratification and clustering for example), then the probability of selection weight needs to correctly reflect the prob-

abilities associated with each stage of selection. For the nonresponse weight, the observed sample is smaller in size than the original sample, so to compensate, re-weighting can be performed by adjusting the design weights by factors that account for each unit's probability of response (StatCan, 2009, 59). These factors are usually obtained using response models (ibid, 59).

If auxiliary data are available, an improvement to the precision of certain estimates can be achieved by a process known as calibration, which consists of adjusting the weights such that estimates of the auxiliary variables satisfy known totals (ibid, 59). The post-stratification weight is one such example, but more generally, desirable properties of calibration include (1) coherent estimates between different sources of data; (2) potential improvements to the precision of the estimates; and (3) potential reduction of unit nonresponse error and coverage error (ibid, 59). Final estimates of key statistical quantities of interest are then about correctly accounting for these weights in the estimation process.

2.4 Data Quality and Survey Errors in Statistics South Africa Household Surveys

Evident from the above discussion is that every component of survey error links through to data quality metrics. But it is also important to be aware of the broader efforts within the statistical organisation to produce the dataset from inception of the project to public-release. Therefore, in order to make an accurate assessment of microdata quality, the TSE framework is an important start.

We now investigate the quality of South African labour market household surveys from the mid 1990s to the mid 2000s. This was a unique period in the country's history during which many changes were taking place, including inside the national statistics office. The surveys considered are the October Household Surveys (OHS, 1995–1999) and the Labour Force Surveys (LFS 2000 September—2007 September). The variable of interest is employment income (a necessary choice when discussing the measurement side of TSE), and we will be tracking the evolution of the income question over time within the context of changing survey instruments and methodological innovations.

An analytically challenging part of this discussion is trying to understand the changing situational environment within Statistics South Africa (SSA) over the period of interest. In order to do this, the results of a personal interview with a retired sampling statistician—Professor David Stoker—will be utilised (see Daniels and Wittenberg, 2010). Prof Stoker worked with SSA in various capacities from 1985 onwards, and his institutional knowledge about what was happening at the time is thought to be unique.

As far as the surveys themselves are concerned, the OHS and LFS share the same mode, namely they are face-to-face household interview surveys where an interviewer asks a household member a set of questions from a questionnaire about

that member's activities and about other household members' activities. However, the OHS was always a single cross section, while the LFS was a biannual rotating panel commencing in February 2000 and extending until September 2007. In 2008, SSA changed the LFS to a quarterly panel, but stopped releasing questions about income to the public; hence, the QLFS will not be reviewed here.

2.4.1 Representation of the Population of Interest

In this section we evaluate the errors of nonobservation associated with the TSE framework, including coverage error, sampling error, nonresponse error, and adjustment error. As before, the time period of interest is 1995–2007. At the start of this period the newly formed geopolitical region of democratic South Africa had just been born out of an Apartheid state that excluded what were known as the Bantustans (Transkei, Bophuthatswana, Venda, Ciskei—the TBVC states). The challenge for the national statistics agency was therefore to help everyone understand this new country, and there was much urgency on the part of policymakers to know the socio-economic features of the new South Africa. While surveys like the OHS were conducted during this period to achieve these ends, survey documentation was often very poor, complicating attempts to understand everything that was going on at the time.

Coverage Error

The new geopolitical entity of South Africa required a new sampling frame, which took time to create. In fact, the 1996 Census was the first time that Statistics South Africa (SSA) had the opportunity to send fieldworkers to every part of the country. As such, it served as an opportunity to validate the existence of dwelling units in remote areas that had escaped previous enumeration attempts and only been observed by satellite imagery.

The next major effort to understand the limitations with the sampling frame was the 1996 Post Enumeration Survey (PES). A PES is an independent survey that allows comparisons to be made with Census results, permitting estimates to be made of coverage and content errors (Whitford and Banda, 2001). One of the major objectives of a PES is to develop a methodology for the calculation of the undercount or overcount of the Census, which can be differentiated by geographical area or demographic characteristics (e.g. age, race, sex).

Since the OHS 1995 was conducted before the 1996 Census, it is likely to suffer from the greatest degree of coverage error compared to all other surveys investigated in this document (OHS 1995—LFS 2007 September). However, SSA did release updated OHS 1995 weights based on the population totals in the 1996 Census (a few years after it was completed) in order to reduce this source of error.

The next major effort to update the sampling frame was the 2001 Census and the subsequent 2001 PES. The 2001 Census also experienced problems in the field with interviewers, such as interviewers stopping work because they had not been remunerated (this was reported in the local press at the time). However, between the Census and the PES, the national sampling frame would have been appropriately updated. The final concerted effort to update the sampling frame was the 2007 Community Survey, but that falls outside the scope of this document.

It is important to note that despite the discussion above, sampling frames are not just updated at discrete points in time. Because SSA undertake surveys every year, and employ fieldworkers to administer questionnaires, feedback from interviewers concerning the absence of existing dwelling units or the presence of new units takes place on a continuous basis. This information impacts the measure of size of each cluster the fieldworkers visit, and therefore has an important implication for the calculation of the correct selection probability of each dwelling unit or household within the cluster.

In summary then, the fact that a new geopolitical unit was created with the democratic South Africa in 1994 meant that the Statistics Agency had their work cut out for them. Coverage error was therefore likely to be largest in the mid to late 1990s, diminishing steadily as the frame became fully enumerated. Since SA is a developing country, we also expect migration patterns and new housing developments to have a significant effect on coverage error over time. This means that the sampling frame is likely to continue to change on an annual basis. The importance of using a combination of technology (e.g. GIS) and skilled interviewers with a virtuous feedback loop to the sampling statisticians then becomes the key to reducing coverage error.

Sampling Error

It is important to understand key developments in the sample design of the various surveys over time. The type of surveys evaluated (the OHS and LFS) also raise different questions with respect to sampling error: the OHSs were all single period cross-sectional surveys with complex probability-based designs, while the LFS was a rotating panel survey. Sampling error for a rotating panel is expected to be slightly different compared to a cross-section (see StatCan, 2009, 23–26).

There were important changes made to the sampling design of the OHS 1995 compared to all previous surveys conducted by SSA before that, namely that (1) the focus switched to households rather than dwelling units, (2) the number of households drawn within each EA was reduced while the number of EAs was increased, and (3) race stopped being used as an explicit variable upon which to stratify the sample (Daniels and Wittenberg, 2010). These were changes in the sample design that improved the representivity of the sample relative to the population, and increased the cost of the surveys (specifically in the case of increasing the number of EAs).

The OHS 1996 sample was produced in conjunction with the sample for the 1996 Post Enumeration Survey (SSA, 1996, Metadata), while the OHS 1997 was based on the administrative records of the 1996 Census, which are records kept by interviewers

for each EA they visit (Daniels and Wittenberg, 2010). The 1998 OHS was based directly on the Census 1996 (SSA, 1998, Metadata), while the OHS 1999 was based on the 1998 Master Sample. However, due to the concurrent implementation of the Census in 1996 and Post Enumeration Survey in 1996, the budget for the 1996 OHS was reduced and the sample size reduced substantially, thereby increasing sampling error.

The 1998 Master Sample then came to play a major role for many SSA surveys including the LFS Rotating Panel. SSA developed the first master sample in 1998, and then updated it in 2003 and 2008 (Daniels and Wittenberg, 2010). The master sample reserves certain clusters of households for certain planned surveys in the future as well as ad hoc surveys that may arise. The SSA 1998 master sample was reserved for the last of the OHSs, the LFS, the General Household Survey and the 2000 Income and Expenditure Survey (ibid, 2010). Anecdotally, the budget for the OHS in 1998 was also lower, possibly due to resources diverted to the development of the master sample, and this reduced the sample size of the OHS in 1998 accordingly, increasing sampling error in this year too.

The advantage of a master sample is that even though it is expensive to develop initially, it becomes more cost effective in the long-run because more than one survey can be based on it (Pettersson, 2005, 72). However, the disadvantage of a master sample is that because it fixes the households that will be selected in each EA for each survey at the time of development, it can become outdated the longer it is used.

The LFS experienced many problems initially with successfully implementing a rotating panel survey design. The first wave of the panel was in February 2000, but subsequent to that two problems arose: (1) the rotating part of the sample was improperly implemented, and (2) fieldworkers were not properly trained to do what they were supposed to in terms of interviewing the same household (Daniels and Wittenberg, 2010). The correct implementation of the rotating panel design only commenced in LFS 2002 February (ibid, 2010).

From a sampling point of view, a panel differs from a single cross-section in that while the sample for a rotating panel is nationally representative in the first wave, it can loose that representivity over time. The rotation of the sample is designed to reduce this loss of representation. Attrition can cause bias in panel surveys, but this was never rigorously explored by SSA over the life of the LFS, most likely due to the role that the rotating component of the panel played in frequently refreshing the sample.

Nonresponse Error

There are two components of nonresponse, namely unit and item nonresponse. Our focus here is on unit nonresponse only (Chapter Four of this book will focus on item nonresponse for employee income data).

Unit nonresponse occurred in every survey under review. However, SSA's description concerning how they dealt with unit nonresponse is completely absent for every OHS. The LFS is also silent on unit nonresponse until the LFS 2000 September, when

Table 2.1 Intended and realised sample sizes

Year	Intended sample size	Actual sample size	Percent
1995	30,000	29,700	99.0
1996	16,000	15,920	99.5
1997	30,000	29,811	99.4
1998	20,000	18,981	94.9
1999	30,000	26,134	87.1
2000	30,000	26,648	88.8
2001	30,000	27,372	91.2
2002	30,000	26,529	88.4
2003	30,000	26,835	89.5
2004	30,000	28,594	95.3
2005	30,000	28,418	94.7
2006	30,000	28,363	94.5
2007	30,000	27,981	93.3

it is only mentioned with respect to the weights (SSA, 2000, Metadata). Despite this, it is possible to track the extent of unit nonresponse. We do this in Table 2.1 below by showing the difference between the intended sample size for each survey from OHS 1995—LFS 2007, compared to the realised sample size (computed by evaluating the number of households in the datasets) released for each survey.

Table 2.1 shows that there are very high response rates in SSA's household surveys, particularly in the 1990s. Kerr and Wittenberg (2012) provide evidence that this was because SSA substituted for unit nonresponse in the early OHSs, yet there is no indication of this in the *Metadata* survey documentation that accompanies the surveys (see OHS and LFS Metadata, 1995–2007).

Adjustment Error

There are three principal weights used for adjustment purposes: (1) probability of selection, (2) unit nonresponse, and (3) post-stratification. The survey documentation for the OHS is only ever useful when it comes to understanding the first of these for households and individuals. From a reading of the Metadata files for each OHS, it seems that SSA never corrected for unit nonresponse using weights (see SSA, Metadata: OHS95-99). Unit nonresponse weights are only officially mentioned in the LFS 2000 September survey documentation (see SSA, 2000, Metadata).

The post-stratification weight is also never discussed or even hinted at in any OHS survey documentation (see SSA, Metadata: OHS95-99). The LFS 2000 February is the first survey in the series evaluated here to include a discussion of post-stratification and how it was conducted.

Adjustment error therefore seems to be possibly one of the largest sources of TSE in the OHSs. For the LFS, the weights seem to be fine. However, neither unit nonresponse weights nor post-stratification weights featured in the official documentation of the OHSs. Researchers have for some time been struggling to understand the apparent jumps in key weighted variable estimates over time using SSA's household surveys (see Branson and Wittenberg, 2007 and Branson, 2009). This goes at least part of the way to explaining why these apparent trend-breaking patters are found over time.

2.4.2 Measurement of the Construct of Interest

We now turn to the measurement side of the Total Survey Error framework and use the employment income variable to anchor the discussion. The income question is directed to employees only in the OHSs, but to both employees and self-employed in the LFSs. In the discussion below, we evaluate the employee income question only, thereby tracing the evolution of the question over time. The surveys instruments evaluated include the OHS 1995—OHS 1999, and the LFS 2000 February—LFS 2007 September.

Validity

The construct of interest for all surveys reviewed in this section is income earned in the main job for all individuals that were employed in the last seven days, except in the OHS 1995 where the "seven days" is not made explicit in the wording of the question. Throughout the OHSs and LFSs, income is always distinguished into various components in the instrument, including (a) salaries and wages, (b) bonuses and (c) income from overtime. The question thus requires the respondent to provide the sum of the three components of income in a single estimate. This amount is before tax.

Key features of the income question in the OHS and LFS are summarised below.

Table 2.2 Features of the income instrument

	OHS & LFS income question
Survey mode	Personal interview
Recall period	Weekly, monthly or annually
Anchoring cues	Main activities in last 7 days
Tax status	Before tax
Components	Salary, overtime, allowances, bonuses
Seasonal adjustment	No, unless annual (in which case it is implicit)

The extent to which this income question loses validity is negligible. The focus is on income in the main job, and consequently remuneration in that job would yield the correct distribution of salaries earned by the employed. If individuals have more than one job, then total income earned by the individual would be higher, but total income is a different construct to income earned in the main job. Consequently, results should be interpreted as such.

There is no mention in the survey documentation of SSA whether the questionnaire was ever pre-tested or how it fared when translated. This shows the paucity of information relating to data quality for many of these surveys. However, we can observe from the income questions themselves important changes to the wording over time. In 1995, the time period options for reporting income included daily, weekly and monthly, but that changed after 1998 to weekly, monthly and annually. This had a deleterious effect on aggregation and standardisation of income values for the sample. It also renders comparisons over time problematic because researchers have to make very arbitrary decisions about how to treat daily income.

Measurement Error

As noted above, Groves (1991, vi) differentiates measurement error into four components including the interviewers, the respondents, the questionnaire and the mode of data collection. The two components that are most important for the income question are interviewer effects and errors due to the psychological issues impacting respondents (viz. social sensitivity of the income question). The wording and the mode also play a role, though are likely less significant. The wording of the income question is identical in every SSA survey investigated except for the OHS95. Whatever weaknesses are associated with this wording are held constant across the surveys. Similarly so for the mode of data collection, since the OHSs and LFSs are both face-to-face surveys.

The impact of interviewers on respondents is multi-dimensional. Because income is such a socially sensitive question, respondents may be influenced by any number of psycho-social and socio-demographic factors, such as the race and gender of the interviewer and even the tone of voice used. As a consequence, interviewer training is very important when trying to solicit income information in face-to-face household interview surveys (Groves and Couper, 1998). Survey organisations consequently often try and match the race of the interviewer with the expected racial majority of the geographical areas of responsibility of the interviewer. Further training of interviewer conduct and behaviour within households is also frequently undertaken.

As far as the wording and sequencing of the income question is concerned, there are two parts to the question in all the OHSs and LFSs except 1996. The first is when the interviewer asks the respondent for the actual value of their income. A respondent is then faced with three options: (a) to provide the actual value, (b) to refuse to provide the value, or (c) to state that they don't know the value. Only if the respondent does not provide an actual value, is s/he presented with a list of income brackets. For a respondent to then decide to provide an answer after having failed to

do so at the first prompt suggests either that they did not want to reveal the precise value of their income and now have been persuaded to do so by the showcard with income brackets, or that they are unsure of the exact value of their income (or other people in the household's income that they are asked to provide a value for).

This latter feature of the question, where the respondent is asked to provide the income of other members who live in the household, potentially induces a considerable source of measurement error. One would expect that cohabiting or married partners would have better information about each others' income, but multiple unrelated employed people in one household may know very little about the income of other household members. The ratio of self-reporters to proxy reporters in the surveys are presented in Table 2.3.

An identifier for self-reporting was only included in the questionnaire from 1999 onwards. We can see from the table self-reporters generally constitute no more than sixty percent of the sample in any given year. This implies that the scope for measurement error due to proxy reporting is rather substantial. There is very little that can be done about this, save to be aware of it and control for it where possible.

The existence of a bracket reporting option in the income question is designed to reduce item non-response, but in so doing, an additional component of measurement error is introduced. This is the case simply because we now no longer know the exact wage of the respondent, but rather the range into which it falls. However, non-response is more expensive to deal with for survey organisations and statistically poses tougher challenges, so this trade-off between components of total survey error is important for the income question.

In surveys where point and interval options are presented to the respondent, the sequencing of the prompts and nature of the alternatives are important because they can aid recall and provide information about the response process. Often, the practises of survey organisations differ in important respects on this matter. SSA sequence the income question in the OHSs and LFSs to firstly ask the respondent for an exact value of their income before the interval prompt takes place. In the Health and Retirement Study (HRS) in the USA, however, the sequencing is the same as the Labour Force Survey (proceeding from an exact value to an interval estimate), but the nature of the prompt for the intervals is very different. Instead, the HRS has an unfolding bracket design where the respondent is first asked if they earn greater than $25,000. If they respond in the affirmative, the interviewer then proceeds to ask whether they earn a higher amount (>$50,000); if they respond in the negative, a lower value is prompted (>$5,000). This proceeds logically until a narrower interval is obtained (see Heeringa, 1995 for a discussion of the income variable in the the HRS instrument). The National Income Dynamics Study (2010–2011) in South Africa employs a similar unfolding bracket design to the HRS for all income questions.

The analytical implications of the different designs are non-trivial. As Vazquez-Alvarez (2006) and Melenberg et al. (2006) have demonstrated, the unfolding bracket design introduces anchoring bias. Anchor strategies are purposefully introduced into surveys to aid respondent recall (see Blair et al., 1991). However, they also introduce potential biases into the results. While the sequencing and format of the brackets in SSA's design is likely to be free from anchoring bias, it remains an open question

Table 2.3 Self and proxy reporting per survey year

Survey year	Proxy	Self reporter	Total
1999	11,647	13,619	25,266
%	46.1	53.9	100
2000	10,216	14,876	25,092
%	40.71	59.29	100
2001	11,299	13,733	25,032
%	45.14	54.86	100
2002	11,182	12,880	24,062
%	46.47	53.53	100
2003	9,873	13,791	23,664
%	41.72	58.28	100
2004	10,425	13,542	23,967
%	43.5	56.5	100
2005	10,011	14,946	24,957
%	40.11	59.89	100
2006	9,898	14,985	24,883
%	39.78	60.22	100
2007	10,668	13,971	24,639
%	43.3	56.7	100

whether it is an improved method. Casale and Posel (2005) note the non-randomness of the bracket subset of respondents, identifying differences between self- and proxy-reporting to be significant.

Table 2.4 shows the evolution of the distribution of response types in the Labour Force Survey for the employed, economically active population only. We restrict the analysis to this survey only and this particular subsample in order to demonstrate how the empirical magnitudes change when we hold the instrument constant.

From the table we can see that over time, the continuous subset of observations has reduced, but not monotonically. The percentage of bracketed response categories fluctuated around 20 percent in every year except 2000, when a disproportionate number of respondents provided a continuous response. This may have been due to greater training of interviewers by SSA to assure respondents of the confidentiality of the information. "Don't Know" and "Refuse" response options increased to about their steady state after the year 2000, when they were at their lowest. This again suggests that unusual effort was expended by the survey organisation in 2000 to obtain good quality income responses, and better interviewer training may have been the key here.

Table 2.4 Distribution of Response Types Per Survey Year

Response Type	2000	2001	2002	2003	2004	2005	2006	2007
Zero-Bracket	0.32	0.16	0.25	0.25	0.25	0.22	0.29	0.32
Zero-Cont.	0.02	0.01	0	0	0	0	0	0
Continuous	86.13	73.71	68.58	66.03	70.93	72.19	74.4	74.83
Bracket	9.93	20.13	23.94	25.87	21.8	21.84	20.55	20.01
Don't Know	0.39	2.54	3.24	2.6	2.74	2	1.34	1.48
Refuse	0.86	3.05	3.77	5.11	4.08	3.5	3.12	2.85
Unspecified	2.35	0.4	0.21	0.14	0.2	0.24	0.31	0.51
N	25,414	25,118	24,086	23,691	23,993	24,958	24,899	24,653

Processing Error

The impact of processing error on the survey is often difficult to detect for the income question specifically, but it has potentially significant implications. Because of the release of three variables into the public-use data for employee income (i.e. continuous income, categorical income and the time unit of reporting), processing error has the potential to exist when more than one response type exists for the same individual (we explore this further in the next chapter in this book). Other examples of processing error in the income question include:

- Incorrectly coding an income value, for example by inputting the data incorrectly or failing to input the data for the income question.
- Recording the actual income incorrectly.
- Recording the actual income value's time-frame incorrectly.

It is not always possible to identify all of these forms of processing error in the surveys, but some forms of error are easily identifiable from the variables released in the data. Furthermore, because processing error can impact all variables unevenly in a public-use dataset, it is important to check all variables of interest for processing error before analysis.

Sometimes processing error may be suspected when there are other ambiguities in the data. For example, one of the far-reaching implications of the wording of the income question in 1995, where the question prompts the interviewer to clarify from the respondent whether the amount of income reported is daily, weekly or monthly, is that when one multiplies the number of respondents who reported a daily value for their income by their income, the resulting values are extremely high. On the one hand, this is an artifact of poor question wording; on the other hand, it could be interviewer error. Thankfully the income question changed permanently and for the better subsequent to 1995, but it does render comparisons with that year problematic.

2.5 Discussion

For South Africa during the mid to late 1990s, there were extraordinary demands on SSA. On the one hand it had to define and enumerate a new sampling frame for a revised geopolitical entity. On the other, there were pressing demands by policymakers for information about the new SA, and this pressure likely reduced the time available for thorough documentation and quality control. The mid 1990s was marked by poor operational standards, suggesting that SSA was still very much finding its feet as an institution, itself undergoing internal restructuring as an organisation.

For the representation side of the TSE framework then, we saw that researchers could do very little about coverage error, even though it is likely an important source of error in the OHSs. The 1996 Census and 1996 Post Enumeration Survey played a very important role in defining the new sampling frame. However, it reduced the budget available for the OHS in 1996, which resulted in a reduced sample sizes in that year.

The 1996 Census and 1996 PES helped statisticians develop the first Master Sample in 1998, which was then used to define the Labour Force Survey sample and many other household survey samples in SA. The switch from the OHS to the rotating panel of the LFS introduced new sampling errors, for rotation was improperly implemented, suggesting once again that SSA was undergoing a process of learning about this new survey instrument.

Fieldworkers play a very important role in updating the measure of size of Enumerated Areas (EA) drawn in the master sample as new dwelling units are added or destroyed. As the master sample gradually becomes outdated, improper enumeration or failure to re-enumerate can introduce a form of coverage error. Inbetween updating the master sample, then, fieldworkers also have an impact on this source of error.

For the probability of selection, (unit) nonresponse and post-stratification adjustments, survey organisations usually provide weights that must be taken into account when analysing the data. However, the weights in SSA datasets seemed to be problematic and certainly not subject to sufficient methodological documentation until later waves of the LFS. The weights always combined at least the probability of selection weight with a post-stratification weight (in the OHSs), and also with the unit nonresponse weight (in the LFSs), to form one composite weight differentiated by individual and household. Because the process was never described in relevant documentation, researchers were never aware of exactly what SSA did in this regard. The weights that were released to the public generated population totals on key variables of interest that were often unstable and highly variable when the datasets were stacked over time.

For item nonresponse on individual variables like income, Stats SA have never provided single or multiple imputations of missing data. It therefore falls to researchers to evaluate the patterns of missing data on variables of interest, and then to develop solutions like single or multiple imputation strategies to deal with this form of potential bias in public-use datasets. This issue is explored later in this book.

For the measurement side of the TSE framework, validity of the constructs in the questionnaires are usually established by pre-testing exercises. But there is no

record of this in the documentation throughout the period of 1995–2007. For specific variables like income, the design of the question is usually targeted at reducing item non-response (by including the income brackets as a follow-up prompt), but it does so at the cost of introducing measurement error on the value of income reported. From a survey design point of view, this can be interpreted as a trade-off between non-response bias and measurement error attributable to the instrument. In other words, it is preferable to have some measurement error on the income variable than to have non-response on it, which is much more difficult to understand or treat appropriately if it is non-ignorable non-response. Non-ignorable non-response cannot be understood effectively without incorporating and budgeting for a specific study of non-respondents to be undertaken by the survey organisation. However, this was never done with SSA's OHSs and LFSs.

The actual wording of the income question did change over time, however, despite no clear documentation of pre-testing questions. In fact, the income question changed with almost every OHS until it stabilised in the LFS. The time units for income reporting eventually moved away from daily, weekly and monthly (up until 1998, though in 1995 an annual option was also available) to weekly, monthly and annually (from 1999 onwards). "Don't know" as a response option was added to the question in 1999, and "Refuse" was added as a further response option from the commencement of the LFS. The ranges of the income brackets changed between 1995 and 1996 and 1997, after which those ranges remained constant all the way through to the 2007 LFS. Finally, the self employed were asked a different income question in the OHSs, while they were asked the same income question in all of the LFSs.

Measurement error attributable to the interviewer was anecdotally rife throughout these surveys due to poor fieldworker practises (e.g. recruitment and training). One can only speculate about whether and how interviewers influenced respondents, thereby introducing another form of measurement error, but this is impossible to quantify. Finally, because of the release of three variables in the public-use data for income (i.e. continuous income, categorical income and the time unit of reporting), processing error was introduced into the data when more than one response type existed for the same individual. This gradually reduced over time though, suggesting more careful data cleaning or interviewer training on this question. We discuss the scope processing error further in the next chapter in this book.

2.6 Conclusion

At the heart of any discourse on scientific method is debate about data quality. For producers of data, modern expectations are that greater disclosure of the limitations of data is required. For consumers of data, judicious analyses of that data mandates a thorough understanding of what the data is intended to measure, versus what it can be stretched to accommodate. Scientific research often shapes policy dialog, and so another interest group begins to weigh in on data quality debates. Unfortunately, debates that are ostensibly about data quality can often hide disingenuous attempts

to thwart results based on sound data, particularly in the policy domain. The need for a clear framework for investigating data quality is therefore a cogent one.

The main contribution of this chapter has been to (1) adapt the TSE framework into one that recognised the limited agency of researchers to assess data quality; and (2) integrate the TSE paradigm with the data quality paradigm. This helped create a framework for investigating microdata quality that was sensitive to the capacity of agents to diagnose data quality in the first place, while at the same time recognising the pressures that shape data quality within data production organisations.

It is important to recognise that improvements to data quality did happen over time with SSA labour market surveys, partly as a natural consequence of the learning process from previous mistakes and partly because of the involvement of researchers and policymakers who communicated their data quality concerns to Stats SA. As researchers focussed specific effort on only a few variables in the surveys, they often uncovered deficiencies in the data that were much harder for the survey organisation to detect. Consequently, improving data quality is an iterative process that should ideally promote a virtuous cycle of interaction between producers and consumers of data. For producers of data, the preparation and publication of detailed data quality frameworks is recommended in much the same way as Statistics Canada and SSA have gone about developing them. These frameworks are also excellent documents to inform users about issues of relevance to survey organisations, such as confidentiality issues.

The advantage of using a coherent framework to discuss data quality is that it directs attention to components of the data production process and the likely data quality elements that led to that error. However, for researchers as consumers of data, the TSE framework is insufficient in itself to inform efforts to rigorously interrogate data quality, for it is rarely possible to identify those errors or quantify their magnitude in public-use datasets. In the absence of clear data quality documentation for each survey instrument, considerable thought therefore needs to be given to the likely errors that exist and their impact on analyses. For example, comparing poverty estimates between the mid 2000s and the mid 1990s using the LFS and OHS is likely an exercise riddled by coverage errors that researchers can do very little about. Yet these numbers often dominate the policy discourse. Under such circumstances, it is far better to acknowledge uncertainty more explicitly and to consider the bounds of sensitivity of key estimates to alternative assumptions about the data generating process.

References

Bhorat, H. (1999). The October household survey, unemployment and the informal sector: A note. *South African Journal of Economics, 67*(2), 320–326.

Blair, J., Menon, G. & Bickart, B. (1991). Measurement effects in self vs. proxy responses to survey questions: An information-processing perspective. In Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.) *Measurement error in surveys*. New Jersey: Wiley.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology, 25*(2), 139–149.

Branson, N. (2009). Re-weighting the OHS and LFS national household survey data to create a consistent series over time: A cross entropy estimation approach. SALDRU Working Paper Number 38. Cape Town: Southern Africa Labour and Development Policy Research Unit (SALDRU)

Branson, N., & Wittenberg, M. (2011). *Re-weighting South African national household survey data to create a consistent series over time: A cross entropy estimation approach* (SALDRU Working Paper 54). Cape Town: Southern Africa Labour and Development Research Unit (SALDRU).

Branson, N., & Wittenberg, M. (2007). The measurement of employment status in South Africa using cohort analysis, 1994–2004. *South African Journal of Economics, 75*(2), 313–326.

Canberra Group. (2001). *Expert group on household income statistics: Final report and recommendations*. Ottawa: The Canberra Group

Canberra Group. (2011). *Canberra Group handbook on household income statistics* (2nd ed.). Geneva: United Nations Economic Commission for Europe.

Casale, D., & Posel, D. (2005). *Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa*. Mimeo, Durban: University of Kwazulu-Natal.

Daniels, R. C., & Wittenberg, M. (2010). *Sampling methodologies in Statistics South Africa household surveys: A conversation with David Stoker*. Mimeo, Cape Town: Data First, University of Cape Town.

Daniels, R. C., & Rospabe, S. (2005). Estimating an earnings function from coarsened data using an interval censored regression procedure. *Studies in Economics and Econometrics, 29,* 1.

Flinn, C. J., Kulka, R., Moffitt, R., & Wolpin, K. I. (2001). Introduction to the journal of human resources special issue on data quality. *Journal of Human Resources, 36*(3), 413–415.

Glewwe, P. (2005). Overview of the implementation of household surveys in developing countries. In *Household sample surveys in developing and transition countries*. Mimeo: Department of Economic and Social Affairs. New York: United Nations Statistics Division.

Groves, R. M. (1991). Measurement error across the disciplines. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys*. New Work: Wiley.

Groves, R. M. (2004). *Survey errors and survey costs*. New York: Wiley Press.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New Jersey: Wiley.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley Press.

Heeringa, S. G. (1995). Application of generalized iterative Bayesian simulation methods to estimation and inference for coarsened household income and asset data. In *The Proceedings of the Section on Survey Methods* (pp. 42–51). American Statistical Association.

Heeringa, S. G., & Groves, R. M. (2006). Responsive design for household surveys: Tools for actively controlling survey nonresponse and costs. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 169,* 439–457.

Kerr, A., & Wittenberg, M. (2012). *The impact of changes in Statistics South Africa's enumeration practise on average household size*. Oxford: Centre for the Study of African Economies Conference Paper.

Kish, L. (1965). *Survey sampling*. New York: Wiley Press.

Krotki, K. P. (2008). Sampling error. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (Vol. 2). Thousand Oaks: Sage Publications.

Melenberg, B., van Soest, A., & Vazquez-Alvarez, R. (2006). *Identification and estimation with partial respondents and anchoring effects*. Tilburg: CentER Discussion paper series: 01–57.

Pettersson, H. (2005). Design of master sampling frames and master samples for household surveys in developing countries. In *Household sample surveys in developing and transition countries*. Mimeo: Department of Economic and Social Affairs. New York: United Nations Statistics Division.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Rubin, D. B. (1976). Inference and missing data. *Biometrica, 63,* 581–592.
Statistics South Africa (SSA). (1995). *October household survey metadata*. Pretoria: SSA.
SSA. (1996). *October household survey metadata*. Pretoria: SSA.
SSA. (1998). *October household survey metadata*. Pretoria: SSA.
SSA. (2000). *Labour force survey metadata*. Pretoria: SSA.
SSA. (2005). *Labour force survey metadata*. Pretoria: SSA.
SSA. (2009). *South African statistical quality assessment framework*. Pretoria: SSA.
SSA. (2010). *South African statistical quality assessment framework* (2nd ed.). Pretoria: SSA.
Statistics Canada. (2003). *Statistics Canada quality guidelines* (4th ed.). Ottawa: Statistics Canada.
Statistics Canada. (2009). *Statistics Canada quality guidelines* (5th ed.). Ottawa: Statistics Canada.
Vazquez-Alvarez, R. (2006). *Anchoring bias and covariate nonresponse*. Mimeo, Version: July, 2006. St Gallen: St Gallen University.
Whitford, D. C., & Banda, J. P. (2001). Post enumeration surveys (PES's): Are they worth it? In *Symposium on global reveiw of 2000 round of population and housing censuses: Mid-decade assessment and future prospects*. Department of Economics and Social Affairs, United Nations Secretariat. New York: Statistics Division.
Wittenberg, M. (2004). The mystery of South Africa's ghost workers in 1996: Measurement and mismeasurement in the manufacturing census, population census and october household surveys. *South African Journal of Economics,72*(5), 1003–1022.
Wittenberg, M. (2006). Research note: Errors in the October Household Survey 1994 Available from the South African Data Archive. *South African Journal of Economics,74*(4), 766–768.
Yansaneh, I. S. (2005). Introduction. In *Household sample surveys in developing and transition countries*. Mimeo: Department of Economic and Social Affairs. New York: United Nations Statistics Division.
Yu, D. (2007). The comparability of the Statistics South Africa october household surveys and labour force surveys (Stellenbosch Economic Working Papers: 17/07). Stellenbosch: University of Stellenbsoch.
Yu, D. (2009). *The comparability of labour force survey and quarterly labour force survey* (Stellenbosch Economic Working Papers: 08/09). Stellenbosch: University of Stellenbsoch.

Chapter 3
Questionnaire Design and Response Propensities for Labour Income Microdata

3.1 Introduction

The income question in household surveys is one of the most socially sensitive constructs. Two problems that arise with social sensitivity concern the probability of obtaining a response and the type of response provided. In survey error terms, this translates into an important relationship between questionnaire design (construct validity) and item non-response. In turn, these affect the statistical distribution of income that has both univariate and multivariate implications. Consequently, the interrelationship between questionnaire design and response type is crucial to understand when conducting analyses of the income variable.

This chapter discusses the design of the employee income question and evaluates the characteristics of respondents who report their incomes as exact values, bounded values, and three additional response types that we will initially group into item nonresponse: (a) those who state they don't know their income or that of the proxy individual on whose behalf they are reporting, (2) those who refuse to answer the question, and (3) responses that are coded unspecified responses in the public-use dataset. The focus is therefore on the response process for a particular variable, which is conditional on the respondent having already agreed to participate in the survey.

In all of Statistics South Africa's (SSA) Labour Force Surveys (LFS), which began in 2000, the employee income question commences by asking individuals what the exact value of their income is. If they refuse to answer or state that they don't know, respondents are then presented with a showcard that displays ascending bounds of income categories. Here they are required to pick an income category that most likely captures the correct income value. If they refuse a second time or repeat that they don't know the value, the final response is recorded as such. The treatment of nonresponse groups in the income question differed across the October Household Surveys (here we focus on the OHS 1997–1999). In 1997 and 1998, there were no options for don't know and refuse, whereas in 1999 only an option for don't know was included in the questionnaire. This resulted in a large number of unspecified

R. C. Daniels, *How Data Quality Affects our Understanding of the Earnings Distribution*,
https://doi.org/10.1007/978-981-19-3639-5_3

income responses in the publicly released OHSs, which confound the understanding of the nonresponse mechanism.

Only in the LFS were options introduced into the employee income question to differentiate nonresponse into both don't know and refuse response types, yet there were also always a positive number of unspecified responses in the LFS 2000–2003. The introduction of new response groups to the income question allows us to examine the impact of these questionnaire design changes on the response propensities of participants in the survey. From this, we can understand the item nonresponse mechanism far more precisely, and this has profound implications for imputation strategies that become the subject of the next chapter in this book.

The factors that influence respondents to provide a particular kind of response become important for two main reasons: firstly, it helps shed light on the possible socio-cultural factors that influence social sensitivity *or* social desirability, and secondly it provides insight into the correlates of bounded responses and nonresponse. An important part of the analytical process required for understanding nonresponse is to attempt to diagnose whether that data is ignorable for the type of analysis envisaged. For applied purposes, ignorability determination amounts to establishing whether the data are missing at random or not. Analysing response propensities therefore also helps to characterise the missingness mechanism. Response propensity models are traditionally employed by survey organisations when investigating the determinants of survey participation and unit nonresponse (see Groves and Couper, 1998). The innovation in this chapter is to investigate item nonresponse process analgously.

The chapter proceeds as follows: firstly, different designs of the employee income question in household surveys is discussed. This provides insight into the trade-offs of varying approaches to asking respondents about their incomes, a traditionally very sensitive question and one where evasive behaviour by the respondent is common. Secondly, we discuss the methodology for analysing item response propensities. We draw from the survey participation literature for this purpose, and discuss suitable models to tailor the approach to item nonresponse. Finally, the results are presented and discussed, before the conclusion summarises.

3.2 Questionnaire Design and the Income Question

3.2.1 The Response Process and the Cognitive Burden of Answering Income Questions

Like any survey question, the decision by the respondent to provide an answer to the income question is broadly influenced by (1) whether they can answer, and (2) whether they will answer. Psychological research has demonstrated that respondent knowledge is a matter of degree rather than a dichotomy of knowing and not knowing, where respondent knowledge can be classified in terms of four cognitive states: whether that knowledge is available, accessible, generatable (i.e. able to be cued), or

inestimable (Beatty and Herrmann, 2002, 73). Given this, it would be reasonable to assume that an important objective of questionnaire design should be to structure the sections and questions in such a way as to improve respondent recall, which means framing the instrument and using anchoring strategies to be as supportive as possible in assisting recall.

The design of the questionnaire, including section and question presentation order, is therefore a non-trivial issue when it comes to the quality of responses to questions (Schwarz and Hippler, 1991). Response propensity is not only affected by respondent attributes such as age, race and gender, but also by factors such as the survey mode, interviewer training, question topics and structure, and institutional dimensions (e.g. public or private statistical agency or marketing company) of the survey (Dillman et al., 2002).

For the income question, key goals for the design of the question are not only to reduce item nonresponse, but also to minimise misreporting, under-reporting and measurement error. Hurd et al. (2003) note that questions about incomes are among the most difficult to answer in household surveys for several reasons, including that (1) respondents may be reluctant to reveal information they consider private and sensitive; (2) cognitive issues make it difficult for respondents to accurately report their income, especially when that reporting is done for other household members; (3) the time period for which a source of income is asked in the questionnaire may be quite different to the time period the respondent usually receives that income; and (4) taxes may or may not be included in different sources of income. Hurd et al. (2003) conclude that all of these issues can lead to significant bias (particularly in the case of under-reporting) and measurement error.

In the case of the employee income question, many of these negative potential outcomes are mitigated by the introduction of a follow-up prompt that applies if a respondent initially states that they don't know or refuse to provide a value. The follow-up then asks the respondent to identify some range of values into which their (or the other household member on whose behalf they are reporting) income falls. The objective of this follow-up prompt is to provide an anchoring strategy for the respondent in the form of a lower and upper bound to income, but it also reduces the social sensitivity of the question because it reduces the level of information disclosure. The precise type of follow-up prompt differs between surveys, and there is some discussion in the literature about the relative merits of alternative questionnaire designs.

Anchoring is an important principle that facilitates respondent recall by triggering indirect cues in the cognitive response process that bear on the target judgement (Frederick et al., 2010). However, Jacowitz and Kahneman (1995) note that the disadvantage of using an anchor to prompt the respondent into some form of indirect answering of quantitative estimation questions (such as income), is that it introduces the possibility of anchoring bias. Anchoring bias is when respondents provide a value for their income that is closer to the value of the anchor itself, which introduces uncertainty surrounding the reliability of the answer. Jacowitz and Kahneman (1995) develop a simple quantitative methodology to measure anchoring bias. They find that anchoring effects are "surprisingly large", sometimes evident in the origi-

nal evaluation of the anchor as high or low (in the questionnaire design phase), and inversely related to respondents' confidence in their judgements but substantial even in judgements made with high confidence. For the income from employment question, the extent of anchoring bias is partly related to the exact form of the income follow-up prompt, to which we now turn.

3.2.2 Different Types of Income Questions

In household face-to-face interview surveys the employee income question differs mainly with respect to the nature of the follow-up prompt that follows an initial request for an exact amount (of either gross income or net income). This follow-up prompt can differ in three primary ways:

1. Using a show card presented by the interviewer with bracketed responses. This is where the respondent points to an amount on the show card that lies within a predetermined range, say between R1000 and R2000). The highest range of the bracketed response options is usually an open-ended interval with no defined upper bound (Juster and Smith, 1997).
2. Using an unfolding bracket. This is where the respondent is first asked if their income is above a given amount per month, say R1000. If it is, then the interview probes further to ask if it is less than a higher amount, say R2000. The unfolding bracket proceeds logically until an appropriate lower and upper bound is established. This type of follow-up prompt was first introduced in the PSID Wealth Modules of 1984 and 1989 (Juster and Smith, 1997).
3. Using respondent-generated intervals. This is where the respondent is asked to self-identify the lower and upper bounds of their income for a given time period. This is a newer type of follow-up prompt that has not yet entered into widespread survey use, though experimental evidence has showed promising results (Press, 2004; Press and Marquis, 2001).

There are several different dimensions to take into account when discussing the merits of alternative designs. However, all three question types share the commonality that they reduce item nonresponse on the question by providing an alternative response option to an exact response. In order to distinguish the relative merits between the question types, we focus on (1) how they affect the response process, and (2) their analytical implications.

Schwartz and Paulin (2000) conducted an experiment to assess the merits of these three questions types to respondents. Eligibility to participate in the experiment was based on whether a respondent received any money in wages or salary in the past twelve months. An instrument similar to the Consumer Expenditure Quarterly Interview Survey in the USA was developed, with different types of bracketing techniques used including show cards, unfolding brackets and respondent generated intervals (RGIs). Upon completion of the mock interview, a cognitive interview was conducted to evaluate respondents' subjective experience of the process. It was found

that across experimental groups, the show-card conventional bracketing technique received the highest overall preference rating and it was rated the easiest with which to reach an answer, possibly due to the fact that it is the only question with a combination of a visual aid (ibid, 967). This was followed by the RGI technique, with unfolding brackets selected as the least popular technique.

Schwartz and Pualin (2000, 969) suggest that while respondent preference may not be an issue for surveys that rely on only one interview, for longitudinal surveys this factor may become more important. Here, conventional brackets and RGIs are considered to be preferable by the authors. An important finding was also that conventional brackets were likely to have been considered preferable by high-income respondents because there was limited disclosure if their income was in the highest, open-ended bracket. With RGIs, however, high income respondents had to disclose a lower and upper bound that lead to the (self-selected) bounds becoming wider as income increased.

In the final analysis, Schwartz and Paulin (2000) suggest that RGIs are likely to lead to higher data quality on income questions because, unlike the conventional bracket which is essentially a recognition memory task, the RGI technique is a two-step memory task. Here, the respondent must firstly estimate the actual amount and then decide how to bound that amount. Their experiment suggested that one way respondents chose to limit the complexity of the RGI task was to skip it and instead provide an exact value. It was noted (ibid, 969) that exact values are statistically preferred to range responses for income questions because they are more precise, and consequently RGIs would improve data quality.

Analytically, the existence of the bracketed subset raises the issue of anchoring bias. For RGIs and the conventional show-card bracket question, anchoring bias (or entry-point bias) is non-existent, but for the unfolding bracket design it is potentially substantial. For salary income though, Hurd et al. (2003) find that there is little evidence of anchoring bias in the Health and Retirement Study (HRS) in the USA, but Juster et al. (1999) find that there is evidence of anchor bias in measures of saving and income from components of wealth. However, Vasquez-Alvarez (2003) postulates different types of anchoring effects for the HRS's (1996) salary income variable when it is treated as a covariate in a model of differences in smoking prevalence between the sexes, and finds evidence that anchoring biases play a significant role in model inferences. The detection of anchoring bias is a non-trivial issue and much work remains to be done on this topic (see especially Juster et al., 2007).

While conventional show-card brackets and RGIs are not subject to anchor biases, they are not without their problems. Show-cards can only be administered in face-to-face interview surveys, whereas unfolding brackets and RGIs can be presented telephonically too. RGIs are the most recent innovation to questionnaire design for financial data. Press and Tanur (2004) find that the interval length between the lower and upper bounds of RGI questions is directly related to the respondent's confidence in their answer, and that sometimes question wording has a direct relationship to the response rate, and to accuracy of the population parameter. Press and Tanur (2005) suggest that to improve the accuracy of RGIs it is helpful to have respondents provide confidence scores about how sure they are of their answers. RGIs also impose specific

estimation tasks concerning interval estimation at the individual level, as opposed to show-cards and unfolding brackets where the length of the interval is standardised in questionnaire design.

The relevance of this discussion for our purposes is that the choice made by respondents about how to answer the income question matters. The precise nature of the follow-up prompt for income helps overturn initial refusals to the income question and therefore conveys information about the response process. Questions then arise about whether groups of respondents with particular characteristics behave in similar ways and are more likely to disclose their incomes with the follow-up question. This can help shed light on the socio-cultural and ethno-linguistic determinants of social sensitivity or social desirability. Social desirability is when respondents want other people to know what incomes they earn, as a type of demonstration effect.

3.2.3 Analysing Response Groups in the Income Question

Common to all employee income question types is a three-fold differentiation of response groups into exact responses, bounded (bracketed) responses and nonresponse (don't know and refusals).[1] In this section we discuss how models of survey participation can be used to develop response propensity models for individual questions like employee income.

Traditionally, survey methodologists develop response propensity models to understand survey participation (or unit nonresponse), often decomposing nonparticipation into noncontacts and refusals (see De Leeuw and de Heer, 2002). This literature provides an important basis for adapting the models to item nonresponse. Groves and Couper (1998) note that there are four hypotheses about survey participation: (1) the opportunity cost hypothesis; (2) the exchange hypothesis; (3) the social isolation hypothesis; and (4) the concept of authority and survey cooperation.

The opportunity cost hypothesis states that people will participate in surveys if they don't have anything better to do. For example, employed people may have less discretionary time than unemployed people. The exchange hypothesis relates to the fact that people generally feel more obligated to participate if they are given an unconditional gift. The social isolation hypothesis suggests that more isolated individuals have a lower probability of survey participation. An example of this is when an individual is a victim of crime and chooses to close their home off to outsiders. Finally, a survey organisation can use its authority to encourage participation. This is possible for a national statistics agency in particular, but may be less so for a marketing company.

[1] Note that our treatment of "Don't Know" responses as a form of nonresponse takes its precedence from Rubin et al. (1995). However, this definition imposes no constraints on the analysis, and later in this chapter we consider "Don't Know" as a partial form of response because it reveals at least some information about income, as opposed to refusals.

The dependent variable in survey participation models is usually binary, coded zero for conducting the interview and one for not participating (either refusals or non-contacts, but not both). The explanatory variables include variables for environment (e.g. central city urban or suburbia, population density, crime rate, percent under twenty years old); social isolation (including race, mixed ages (e.g. greater than 69 years old), single person household, children less than 5 in the household; residential exchange in last five years); and social exchange (owner occupied house, monthly rental, house value).

Models of response behaviour also incorporate more elaborate individual factors. For example, Johnson et al. (2002) describe the impact of culture on nonresponse. They suggest that cultural variability matters for nonresponse for everything from survey question comprehension, to memory retrieval, judgement formation and response editing processes. As a consequence, it is also important to factor these variables into response propensity models, though it is unlikely that every relevant variable in this respect will be available in public-use datasets.

3.2.4 Questionnaire Design Changes in SA Labour Market Household Surveys

We evaluate employee income in South Africa's two major household interview labour market surveys: the October Household Surveys (OHS; 1997–1999), and the Labour Force Surveys (LFS; 2000–2003 September waves only). The OHS was a repeated cross-sectional survey, while the LFS was a biannual rotating panel survey. Only the September Waves of the LFS are chosen in order to allow the series to be more comparable with the OHS. Since the LFS is a rotating panel survey, it poses no methodological problem to take only one wave in a given year because each wave of a rotating panel is designed to estimate the population of South Africa at the time of going to field. The rotation part of the panel ensures that a portion of the sample changes in every Wave of the survey (Cantwell, 2008).

In both of these surveys, the employee income question developed by Statistics South Africa (SSA) had a show-card follow-up for bracketed responses, but evolved over time with respect to its treatment of nonresponse. In the OHS 1997 and 1998, there were no options for don't know and refuse; in the OHS 1999 don't know was added as an option for the first time; only with the commencement of the LFS in 2000 were both don't know and refuse added to the question.

We want to exploit these changes in questionnaire design to evaluate how they affected the capacity to understand the response process for employee income. Figure 3.1 displays the employee income question in the LFS 2000 that became the standard after much trial and error in the 1990s.

For both the OHS and LFS, the surveys required a single adult respondent to answer the income question for every member in the household. When responses are provided for household members other than the respondent, this is called proxy

4.15.a What is ……'s total salary/pay at his/her main job?

Including overtime, allowances and bonus, before any tax or deductions.

Give amount in whole figures, without any text or decimals

If refusal or don't know ? Go to Q 4.15.c

Only if amount given in 4.15.a

4.15.b Is this

1 = Per week
2 = Per month
3 = Annually

Only if refusal or don't know in 4.15.a

***4.15.c** Show the categories. Make sure the respondent points at the **correct income column** (weekly, monthly, annually) on **Show card 3** and mark the applicable code.*

Weekly	Monthly	Annually
01 = None	01 = None	01 = None
02 = R1 - R46	02 = R1 - R200	02 = R1 - R2 400
03 = R47 - R115	03 = R201 – R500	03 = R2 401 - R6 000
04 = R116 - R231	04 = R501 – R1 000	04 = R6 001 - R12 000
05 = R232 - R346	05 = R1 001 - R1 500	05 = R12 001 - R18 000
06 = R347 = R577	06 = R1 501 = R2 500	06 = R18 001 - R30 000
07 = R578 - R808	07 = R2 501 - R3 500	07 = R30 001 - R42 000
08 = R809 - R1 039	08 = R3 501 - R4 500	08 = R42 001 - R54 000
09 = R1 040 - R1 386	09 = R4 501 - R6 000	09 = R54 001 - R72 000
10 = R1 387 - R1 848	10 = R6 001 - R8 000	10 = R72 001 - R96 000
11 = R1 849 - R2 540	11 = R8 001 - R11 000	11 = R96 001 - R132 000
12 = R2 541 - R3 695	12 = R11 001 - R16 000	12 = R132 001 - R192 000
13 = R3 696 - R6 928	13 = R16 001 - R30 000	13 = R192 001 - R360 000
14 = R6 929 or more	14 = R30 001 or more	14 = R360 001 or more
15 = Don't know	15 = Don't know	15 = Don't know
16 = Refuse	16 = Refuse	16 = Refuse

Fig. 3.1 The income question: labour force survey 2000 September

reporting, which has been subject to some attention in the literature due to the anticipated increase in measurement error associated with a proxy reporter (see Blair et al., 1991). The intuition behind this is simple: a proxy reporter is less likely to know the exact value of the income of other members of the household. While this may be less likely in the case of cohabiting partners in an intimate relationship where the intra-household allocation of resources is shared, it is increasingly likely in multiple adult households either in the same extended familial group or unrelated individuals living in the same household.

One way to account for this is to include a variable for self or proxy reporting directly into the analysis (see for example, Casale and Posel, 2005). However, the ability to do so was not present in the majority of October Household Surveys and only became part of the questionnaire in 1999. The differences between the questionnaires over time therefore has an important bearing on the degree to which we can understand the response process.

The final major difference in the questionnaires between the OHS and the LFS is that in the OHS more general information is provided about the household including their household conditions and exposure to crime for example. In fact, when the OHS ended in 1999, two surveys were designed to replace it: the Labour Force Survey (LFS) and the General Household Survey (GHS, although the GHS was only implemented some years later). The LFS contained all the labour market information from the previous OHS questionnaire with improvements to sections like the income question, while the remainder of the OHS questionnaire was directed to the GHS. Note that despite the differences in the length of the overall questionnaires between the OHS and LFS, the income question appears at roughly the same point in each questionnaire, implying that respondent fatigue by the time they reached the employee income question during the interview was not altered too drastically between the two survey instruments.

The evolution of the survey instrument and the income question in these surveys provides us with a valuable opportunity to evaluate how changes to questionnaire design impacted the response process.

3.3 Methodology

The principle of developing response propensity models for an individual question like income shares its motivation from the analagous requirement to understand the response process for the survey more generally. We begin by describing the evolution of the employee income question and the resulting structure of the data released to the public. Thereafter, the response propensity models are developed before estimation, specification and testing are discussed.

3.3.1 *Response Propensity Models for the Employee Income Question*

Models of survey participation propensity, such as those in Groves and Couper (1998), De Leeuw and de Heer (2002) and Johnson et al. (2002), model the process as a function of (1) variables that reflect the possible perceptions of the respondent to the relative burden of participating in the survey, in combination with (2) variables that reflect the capacity of the survey organisation to shift the perception of the respondent about that burden.

Unlike survey participation propensities, however, response propensities to particular questions in a survey already have buy-in from the respondent about survey participation. Consequently, modelling the process is dependent on the features of the variable(s) of interest. Another way of saying this is that survey participation and response propensities on individual questions are always related in that item nonresponse is conditional upon unit response.

For the income from employment question, we saw from the literature that there are two primary concerns: the cognitive burden of answering the income question, which is partly related to recall and social sensitivity issues; and the expected correlates of income itself, since both bounded response and nonresponse is thought to be related to higher income levels. We therefore also need to incorporate variables that best predict this effect. Here we are limited by the questionnaires themselves.

In the OHS and LFS questionnaires, the following variable groups of interest can be identified in some or all of the instruments:

- Variables reflecting the personal characteristics of the respondent, including sex, race and education. These characteristics are also correlated with income in South Africa (particularly race and education).
- Variables reflecting the cognitive burden of retrieving information about income, including self-reporter, the head of the household, whether the respondent is cohabiting with a romantic partner, household composition variables (number of children, adults and retirees), and household size.[2]
- Variables reflecting the willingness to disclose income (possibly shaped by the social environment of the respondent), including the first language of respondent, whether the respondent felt unsafe in their neighbourhood, and an indicator for urban households.
- Variables that are thought to be highly correlated with income, including total household expenditure, vehicle ownership, home ownership and dwelling type.

Important variables that would help shed light on the response process are interviewer codes and any diagnostic information about the interview itself (often called paradata). However, none of this information is available in any of the public-use versions of the OHSs or LFSs.

[2]The number of retirees will be omitted in order to prevent a perfectly collinear relationship between the household composition variables and household size.

The above variables are included in all of the response propensity models when they become available in the survey questionnaires. Because the same variables are utilised in every survey year, it is important to note that we invoke the assumption that the response process is stationary over time. This implies that, a-priori, we do not expect changes to the direction of influence of the covariates over time. However, their direction of influence can change depending on the response type under investigation. We discuss each variable's rationale for inclusion in the section on model specification and testing below.

3.3.2 Questionnaire Design Changes and the Resulting Structure of Income Data in Publicly Released Datasets

An important difference between the OHSs and LFS was that in the OHS, self-employed individuals answered a different income question to employees, whereas in the LFS both employees and the self-employed were asked the same question. In order to standardise the sample to employees only, we drop all self-employed from all surveys and further restrict the sample to the economically active population (16–64 years old).

In the OHS97 and OHS98, the time period for reporting income was daily, weekly and monthly, whereas in 1999 (and, thankfully, every year since then), the periods changed to weekly, monthly and annually. In all of SSA's public datasets, employee income is differentiated into three variables: (1) a continuous variable that reflects the range of exact income responses; (2) a categorical variable that reflects the ascending bounded income ranges of the bracketed subset; and (3) a variable for the time unit of income recorded. These three variables need to be used to derive a single income variable for analysis.

The two surveys of interest are the OHS (1997–1999) and LFS (2000 September–2003 September). During the OHS, the income question changed (the don't know option was added in 1999 and the time period of reporting changed from daily, weekly and monthly in 1997 and 1998 to weekly, monthly, annually in 1999), and new questions were added to the questionnaire that can help explain the response process (e.g. the introduction of self versus proxy reporting in 1999). The OHS also asked more general questions about the neighbourhood the respondent was living in and their experience of crime, whereas the LFS omitted these questions from the questionnaires. While in the OHS, both the employee income question and the questionnaire changed, in the LFS, neither the employee income question nor the questionnaire changed on key variables of interest.

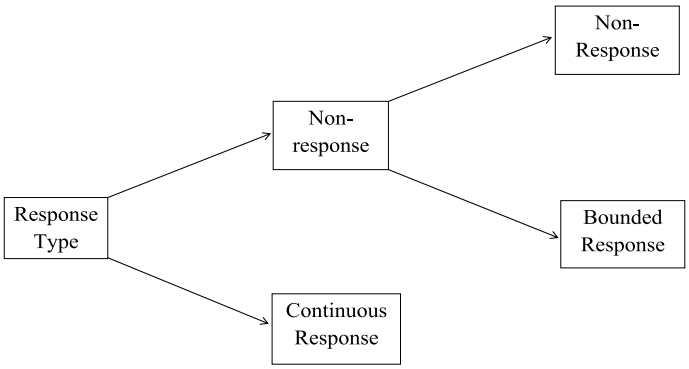

Fig. 3.2 The employee income response process in OHS 1997 and 1998

3.3.3 Estimation, Specification and Testing

Estimation

We can think of response propensity models for employee income as modelling a latent variable for the *unwillingness* to disclose income. This variable is not directly observed, but we do observe the response type for the income question, which gives us information about the level of information disclosure the respondent is willing to provide. An important estimation task is then to adequately account for the sequential nature of the response process that reveals the level of information disclosure.

In the income question, the interviewer first asks the respondent for an exact income value; if they refuse or state that the don't know, the interviewer asks a follow-up question where a showcard is presented to the respondent with bounded income ranges. The respondent can then choose a bracket into which their income falls. Only if the respondent states that they don't know or refuses again, is the final response coded as don't know or refuse.[3]

Because the income question itself evolved over the survey years under investigation (particularly between 1997–2000), the sequential nature of the response process differs over time. Figures 3.2 and 3.3 depict this.

From Fig. 3.2, we see that the respondent can first provide an exact income value or state that they don't know or refuse (collectively grouped as "nonresponse" in the figure). The interviewer then prompts the respondent to answer again, this time with a bounded response follow-up question presented with a showcard. If the respondent refuses again or states that they don't know, the OHS 1997 and 1998 data record an unspecified response for that individual, which we know can be either don't know

[3]Note that we assume the showcard that the interviewer presents to the respondent only has the bounded income ranges printed, rather than the additional options to state that they "Don't Know" or "Refuse", which is present in the questionnaire as per Fig. 3.1. This would ensure that the interviewer does not inadvertently prompt the respondent for a "Don't Know" or "Refuse" response by presenting it on the showcard.

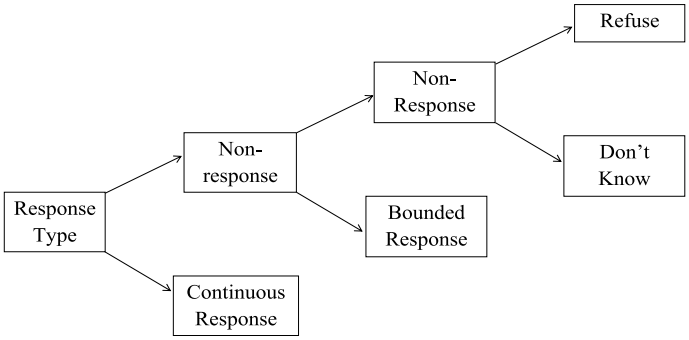

Fig. 3.3 The employee income response process in LFS 2000–2003

or refuse, but which cannot be identified as such from the questionnaire and so is conflated into a grouped "nonresponse" option that concludes the response process for these survey years.

In the OHS 1999, don't know was provided in the income question for the first time, and hence the sequential structure of the response process has an additional branch that decomposes the final "nonresponse" option into don't know and unspecified. Here, unspecified responses are confounded with refusals because no option for refuse is present in the OHS99 questionnaire. In the LFS 2000–2003, we have the same sequential structure as the OHS 1999, but this time the final "nonresponse" option is decomposed into its exhaustive subsets of refusals and don't know responses. Figure 3.3 presents this sequential structure.

A suitable characterisation of this kind of problem is the sequential response model of Maddala (1983). Adapting this model to the problem of the employee income question as depicted in Fig. 3.3, define the outcome variable Y to have four possible alternatives:

- $Y = 1$ if the individual provides an exact response, which equates to full information disclosure;
- $Y = 2$ if the individual provides a bounded response, which equates to partial information disclosure;
- $Y = 3$ if the individual provides a "Don't Know" response, which equates to even less information disclosure; and
- $Y = 4$ if the individual provides a "Refuse" response, which equates to full non-disclosure.

The probabilities of each outcome in the sequential response model can be written as:

$$
\begin{aligned}
P_1 &= F(\beta_1' x) \\
P_2 &= [1 - F(\beta_1' x)]F(\beta_2' x) \\
P_3 &= [1 - F(\beta_1' x)][1 - F(\beta_2' x)]F(\beta_3' x) \\
P_4 &= [1 - F(\beta_1' x)][1 - F(\beta_2' x)][1 - F(\beta_3' x)]
\end{aligned} \tag{3.1}
$$

where F is the cumulative distribution function and the betas are parameters to be estimated.

As Maddala (1983, 49) notes, this kind of model is easy to analyse because the likelihood functions can be maximised by maximising the likelihood functions of dichotomous models repeatedly. By doing this, note that we therefore make the assumption that the probability of choice at each stage of the response model is independent of the choice at the previous stage. In other words, the independence of irrelevant alternatives (IIA) assumption of more general polytomous discrete choice models is applicable here too.

Despite the invocation of the IIA assumption, however, note that unlike the multinomial response model, the sequential response model estimates dichotomous models that *combine* multiple outcomes against a *changing* base outcome sequentially until the stages of the sequence are exhausted. Therefore, as implied by Fig. 3.3 and Eq. 3.1, the first stage of the sequence is estimated combining bounded responses, don't know responses and refusals, $\{Y = 2 + Y = 3 + Y = 4\}$, against the base outcome of a continuous response, $\{Y = 1\}$. The second stage of the sequence is estimated combining don't know and refusals, $\{Y = 3 + Y = 4\}$, against the base outcome of a bounded response $\{Y = 2\}$; and the third stage of the sequence is estimated as $\{Y = 4\}$ against the base outcome of a don't know response, $\{Y = 3\}$.

In other words, the parameter β_1 in Eq. 3.1 is estimated from the entire sample by dividing it into two groups, continuous responses and initial nonresponse (to the first exact income question); β_2 is estimated from the subsample of remaining response types divided into bounded responses and final nonresponse (to the follow-up income question); and β_3 is estimated by dividing the subsample of final nonresponse into refusals and don't know responses.

In this context, the IIA assumption is entirely reasonable because the respondent has to refuse or state that they don't know twice: once to the initial income question for an exact response, and a second time to the follow-up question that presents a showcard. The third stage simply decomposes nonresponse into refusals and don't know, exhausting the possible response alternatives. Hence the IIA assumption is reasonable to defend.

Buis (2011) discusses a modern application (and some limitations) of the sequential response model, and we use the estimator he developed called the sequential logistic model, implemented in *Stata* version 12 using the package written by Buis (2012, Version 1.1.15).

Specification

In this section we discuss variable selection over the different survey years, possible omitted variables and the possibility of measurement error in the explanatory variables. Recall from Sect. 3.3.1 that we have four broad variable groups: (1) cognitive burden of answering income variables; (2) willingness to disclose variables; (3) personal characteristics of respondent; and (4) correlates of income variables. The rationale for including each variable under these themes is presented in Table 3.1.

Across the survey years from 1997–2003, we observe almost all of these variables, but in some years certain variables are not available or they change from categorical to continuous. For example, an identifier for self reporter (versus proxy reporting) only becomes available from 1999 onwards, while the variable for feeling unsafe in the neighbourhood you live is only available in 1997 and 1998.

The variable for total household expenditure changes from continuous in 1997 and 1998 to categorical in 1999. It then changes again in 2000, when it was not asked at all in the LFS 2000 (September) because of the concurrent 2000 Income and Expenditure Survey that was administered to the same households. For this survey year, we merge in the continuous variable from the IES 2000. For all LFS after that, expenditure was asked in the same way as the OHS 1999, when a bounded expenditure range was presented to respondents. Note that only in the years when there is a categorical expenditure variable are there options for don't know and refuse to the question.

It is important to note that for the variable 'first language of respondent', the rationale for including it in the models is to capture socio-cultural influences of social sensitivity to reporting income. In other words, we are interested in whether it affects the willingness to disclose income. However, it is very difficult to predict a-priori what the direction of the coefficients will be, for very little research has been done into this topic in South Africa. In order to ensure that we do not get spurious results in this respect, we are insulated by the fact that the response propensity models will be run over multiple, independent samples of individuals in the South African population over multiple time periods from 1997–2003. Consequently, we get a chance to observe the stability of the findings for language over time.

Note that two different language variables are constructed for the analysis: one that introduces dummies for all eleven official SA languages, and one that keeps Zulu, Xhosa, English and Afrikaans, but aggregates the more regional languages together (including Ndebele, Northern Sotho, Southern Sotho, Tswana, Swazi, Venda, Tsonga and Other language). The rationale for the latter is that the cell sizes for some of these regional languages get very small when included with all of the other covariates. Zulu is SA's most spoken first language, and we consequently use it as the reference category in all regression models.

A similar problem exists with the race variable. In contemporary discourse in SA, race is still disaggregated into the main classifications of the Apartheid era, namely African/Black (hereafter referred to only as African), Coloured, Indian/Asian (hereafter referred to only as Indian), and White. An option for the respondent to report their race as "Other" was present in all survey years from 1997–2003. However,

Table 3.1 Explaining response type: covariate selection

Variable	Rationale for inclusion	Testing
Household head	If respondent is HHH, more likely to know about incomes in the hh	Cognitive Burden (CB)
Self reporter	If a respondent is SR, more likely to know exact income	CB
Cohabiting status	If respondent in a cohabiting relationship, more likely to know spouse or partner's income	CB
HH composition	Tests effects of number of kids (<=15) & adults (16–64) relative to the # of seniors (65+) in hh (reference group). The expected sign here is that an additional adult should increase CB of reporting	CB
Household size	The larger the size of hh, the less likely respondent knows all incomes	CB
Male	Personal characteristics of respondent or proxy	Personal Characteristics (PC)
Age + age squared	Personal characteristics of respondent or proxy	PC
Race	Personal characteristics of respondent or proxy	PC/CI/WD
Education	Education category of respondent or proxy	PC/CI
First language (1)	Dummies for 11 official languages in SA. Captures possible socio-	Willingness to disclose (WD)
	cultural influence to disclose income, though effects ambiguous	
First language (2)	Simplified from above to four main SA first languages: Zulu, Xhosa, Afrikaans & English. All others combined into "Other"	WD
Wealth approximation	Derived from interaction of home ownership dummy with dwelling type:	Correlate of Income (CI)
	(1) Owned formal dwelling, including brick house, semi-detached house, flat or retirement unit (2) Unowned formal dwelling, same dwelling types as above (3) Sub-let room or dwelling, including room in main dwelling or structure in backyard (shack or room), not interacted with ownership (4) Mud hut or shack in squatter settlement, not interacted with ownership	
Expenditure	Total household expenditure: continuous in 97,98 & 00; categorical in 99, 2001–2003	CI
Owns vehicle	Dummy for whether respondent owns vehicle or not. Reflects stock of wealth	CI
Felt unsafe in neighbourhood	If respondent feels unsafe, less likely to disclose income (only available in 97 & 98)	WD
Urban	Testing the effect of location. Has possible effect on willingness to disclose income	WD

the number of individuals in the employed economically active subpopulation who report their race as "other" is very low, ranging from a minimum of zero in 1997 to a maximum of 49 in 2001. We therefore set "other race" to missing in the regression models due to the small cell sizes associated with it, and rather estimate race as a dummy variable for the four main racial groups only, with African as the reference group.

On the question of the construct of race, it should be noted that there is very likely to be some measurement error on this variable. This is because the race question in all survey years (1997–2003) has a reporting option called "African/Black". During and even after Apartheid, the convention among supporters of certain political parties including the African National Congress was to follow the Black Consciousness movement's recommendation to label all historically disadvantaged groups "Black". So, for example, Indian/Asian and Coloured people who were historical supporters of the liberation struggle during Apartheid were (and still are) far more likely to report their race as "Black" compared to the Apartheid classifications given to them (especially among older generations). There is very little we can do about this form of measurement error in the data, other than note it for reference.

It should also be noted that important omitted variables in this analysis include information about the interviewer that administered the questionnaire to the respondent, such as their race, age and gender, and information about the behaviour of the respondent in the interview, such as whether they were hostile or not. However, it is rare that this information is released by the survey organisation to the public, so very little can be done to compensate for these omitted variables other than to acknowledge their importance.

The response propensity models developed in this chapter are not models that allow for causal inference. However, the stability of the signs and effect sizes of coefficients, over independent samples of the employed economically active population of South Africa from 1997–2003, does provide very useful insight into the stability of the correlates of the response process.

3.4 Results

In this section we report the main findings. We commence by conducting a descriptive analysis of the distribution of different response types to the income question, before evaluating the probability of a bounded income bracket response as income increases. We then present the response propensity models. All results are not weighted because we are interested in the characteristics of the sample itself, rather than the population.

3.4.1 A Descriptive Analysis of Employee Income Response Type

Table 3.2 shows the distribution of income subsets when the exact income variable is combined with the bounded income variable to form one derived monthly employee income variable that will henceforth be used for analysis.

The percentage of exact responses in each survey year ranges from 87 percent in 2000 to 54% in 1999. This suggests that interviewer effort and training on socially sensitive questions may yield high dividends. Anecdotal evidence of greater effort by Statistics SA to train interviewers in 2000 is given in Daniels and Wittenberg (2010).

Bounded responses vary from 9% of the sample in 2000 to 37% of the sample in 1998. However, there is no clear trend in the response propensity of this subset over time, though it does rise consistently after 2000.

If we sum the responses for Don't Know, Refuse and Unspecified, we can evaluate the percentage of the sample for each year that represent the group of item nonrespondents for the income question. This number ranges from approximately 3% in 2000 to about 7% in 2003. This suggests that the bracket follow-up prompt is very successful at reducing nonresponse for employee income. The percentage of Don't Know responses doesn't seem to have a discernible trend, but the percentage of Refusals increases steadily from the LFS 2000–2003.

For the bounded subset of observations, preliminary insight into the response mechanism can be obtained by evaluating the probability of a bounded response within each income category. Here, all observed income responses (including the

Table 3.2 Distribution of response types: OHS97–LFS03

Year		Exact	Bounded	Don't know	Refuse	Unspecified	Total
1997	Obs	16 186	6 758	.	.	942	23 886
	Percent	68	28	.	.	4	100
1998	Obs	7 637	4 720	.	.	628	12 985
	Percent	59	36	.	.	5	100
1999	Obs	11 735	8 055	1 588	.	548	21 926
	Percent	54	37	7	.	3	100
2000	Obs	18 745	2 033	72	144	461	21 455
	Percent	87	9	0	1	2	100
2001	Obs	15 948	4 065	521	578	77	21 189
	Percent	75	19	2.5	2.7	0.4	100
2002	Obs	14 469	4 684	651	664	40	20 508
	Percent	71	23	3.2	3.2	0.2	100
2003	Obs	13 759	4 998	485	891	23	20 156
	Percent	68	25	2.4	4.4	0.1	100

exact subset) are converted into bounded ranges before the probability is calculated. Table 3.3 presents the results.

The table shows the percentage of respondents who provide a bounded response when all income observations are grouped into income categories. Don't know, refuse and unspecified responses are omitted from the calculations. A value of 0.98 as the first number for the zero income category in 1997 therefore implies that 98% of respondents who replied that their income was zero did so only when prompted by the interviewer for a bracketed response. There were 46 observations in total for this reporting option in 1997, 98% of which answered inside the bracket bound. The zero income category is somewhat peculiar to the SSA income question and generally has a low number of observations, ranging from 2 in 1998 to 46 in 1997.

For income categories above zero, there is a near monotonic increase in the probability of reporting a bounded response as income itself increases, and this finding holds for almost every survey year. In other words, social sensitivity increases as income increases. Two notable exceptions to the monotonicity finding are in 1998 and 1999, where the highest probability of a bracket response is in the R11,001-R16,000 range in both years. Finally, the total probability of a bounded response in each survey year is presented at the bottom of Table 3.3, where we see it is lowest in 2000 at 10% and highest in 1998 at 38%. This considerable fluctuation may be due to interviewer training on the approach to the income question, as 2000 is considered to be the year that a substantial investment in interviewer training by Statistics SA was made (Daniels and Wittenberg, 2010).

The overall conclusion from this section is that, in general, the probability of a bounded response increases as income increases. This is most likely due to the social sensitivity of income and the higher cognitive burden of answering the income question as an individual's remuneration increases and possibly becomes more complex (e.g. has benefits added or deductions subtracted). We now turn to multivariate analysis to evaluate the predictors of the various response types.

3.4.2 Sequential Response Propensity Models

In this section we report results for the sequential response propensity models over two time periods: (1) 1997–1999, and (2) 1999–2003. In the first period, a two-stage sequential logistic response model is estimated for response type as per Fig. 3.2. The inclusion of OHS99 here means we do not decompose nonresponse into don't know and unspecifieds initially. Instead, we do this in the second time period, when we also analyse the LFS. Here, a three-stage sequential logistic response model is estimated as per Fig. 3.3 and Eq. 3.1. For all models, odds ratios are reported for the coefficients. The results are unweighted because we are interested in the sample itself. Standard errors are robust and clustered at the level of the primary sampling unit.

Table 3.3 Probability of a bounded response within each monthly income category: OHS97–LFS03

Income category	Probability	1997	1998	1999	2000	2001	2002	2003
R0	Prob.	0.98	1.00	0.96	0.86	0.88	1.00	1.00
	Obs	46	2	28	42	24	34	34
R1-200	Prob.	0.16	0.21	0.27	0.07	0.12	0.16	0.17
	Obs	1 497	861	1 404	1 165	1 057	933	551
R201-500	Prob.	0.17	0.20	0.25	0.04	0.08	0.07	0.09
	Obs	3 487	2 160	3 689	3 794	3 346	3 165	2 176
R501-1000	Prob.	0.21	0.29	0.30	0.04	0.12	0.10	0.10
	Obs	4 200	2 057	3 625	4 122	3 844	3 592	4 187
R1001-1500	Prob.	0.28	0.38	0.39	0.08	0.19	0.19	0.17
	Obs	3 848	1 946	2 927	2 776	2 629	2 293	2 176
R1501-2500	Prob.	0.33	0.43	0.45	0.09	0.19	0.22	0.21
	Obs	4 290	2 226	3 235	3 610	3 458	3 143	3 092
R2501-3500	Prob.	0.40	0.58	0.58	0.12	0.30	0.35	0.36
	Obs	2 198	1 132	1 666	1 639	1 792	1 664	1 745
R3501-4500	Prob.	0.45	0.54	0.65	0.18	0.35	0.49	0.48
	Obs	1 286	828	1 041	1 057	1 192	1 175	1 211
R4501-6000	Prob.	0.45	0.58	0.65	0.19	0.36	0.46	0.52
	Obs	1 011	533	922	1 102	1 234	1 304	1 378
R6001-8000	Prob.	0.46	0.61	0.68	0.20	0.37	0.49	0.53
	Obs	542	249	540	624	662	836	975
R8001-110000	Prob.	0.58	0.67	0.68	0.27	0.50	0.58	0.61
	Obs	272	156	282	365	405	518	642
R11001-16000	Prob.	0.68	0.79	0.70	0.29	0.52	0.65	0.68
	Obs	155	85	215	204	203	273	335
R16001-30000	Prob.	0.66	0.53	0.57	0.35	0.59	0.69	0.69
	Obs	82	58	129	133	120	172	201
> R30000	Prob.	0.73	0.16	0.25	0.75	0.66	0.82	0.78
	Obs	30	64	87	145	47	51	54
Total	Prob.	0.29	0.38	0.41	0.10	0.20	0.24	0.27
	Obs	22 944	12 357	19 790	20 778	20 013	19 153	18 757

Two-Stage Sequential Logistic Response Model

We now present the findings for the two-stage sequential response models used for the OHS 1997, 1998 and 1999. For 1999, don't know responses are combined with unspecifieds. The first-stage results are reported in Table 3.4 and the second-stage results are reported in Table 3.5. Recall that the first stage of the sequential logistic model evaluates initial nonresponse to the exact income question, whereas the second stage evaluates final nonresponse compared to bounded responses (see Fig. 3.2). Odds

ratios are reported for all model coefficients, and the effects are discussed for each group of explanatory variables (see the "Testing" column in Table 3.1 for a recap of the variable groups).

Table 3.4 shows the odds ratios for the first stage of the system of equations that represent the sequential response model of equation refeq:rp1, for survey years 1997–1999. Subsequent stages of the model are presented in the tables below. Regardless of the stages of the model, however, it is important to note that the specifications differ slightly between 1997 and 1999 due to changes in questionnaire design. Specifically, the variable "felt unsafe in neighbourhood" appears in 1997 and 1998, but is absent from 1999 onwards. Similarly, the variable for self reporter only appears in 1999. While this renders strict comparison of the stability of predictors over time impossible, it does give us insight into how questionnaire design changes impacted the capacity to diagnose the response process.

Evident from Table 3.4 is that for the cognitive burden variables, none are repeatedly significant across the survey years except household head, and the direction of influence also changes for the number of kids and the number of economically active individuals (aged 16–64 years old) within the household between the survey years. Individuals in cohabiting relationships have lower odds of reporting initial nonresponse, but this effect in only significant in 1998. A self-reporter is significant, but only appears in 1999 and so its repeated effect cannot be assessed yet. In 1999, a self reporter to the income question reduces the odds of initial nonresponse by approximately 29%.

Variables reflecting the personal characteristics of the respondent show a little more stability. Men have higher odds of *not* reporting an exact response, and this effect is significant in every year. The turning point of age is calculated as the coefficient on age divided by two times the coefficient of age squared, and is presented at the bottom of the table. Note that while the turning point is calculated using the log of the odds, the coefficients in the table itself are odds ratios (this convention will be maintained for the rest of this chapter). Note that while the odds ratios in the table are rounded to the third decimal place, the signs for the log of the odds of the coefficients on age squared are all negative. This implies that the shape of the relationship between age and the probability of initially refusing to answer the income question in all three years increases up to the turning point, after which it decreases.

Important to note is that in 1998, the turning point lies outside the upper bound of the sample of economically active individuals (64 years old), suggesting a monotonic relationship between age and response type for this survey year. In 1997 and 1999, however, that relationship is quadratic with a turning point reached at about 52 years of age. Therefore, in 1997 and 1999 individuals are increasingly likely to refuse the initial income question up until 52, whereafter they become more likely to provide an exact income response.

The race dummies show changes in direction of influence across the years for Indian and Coloured people, where the odds ratio suggests a negative relationship for these two groups relative to Africans in 1997, but this changes to a positive relationship in 1998, then changes again to negative in 1999 for Indian people. A

Table 3.4 First-stage response propensity: initial nonresponse compared to exact responses: OHS 1997-OHS 1999

Covariate	OHS97	OHS98	OHS99
Household head	0.842***	0.877***	0.933*
Self reporter			0.708***
Number kids	0.984	1.044	0.957
Number 16–64yrs	1.085	1.095	0.971
Household size	0.963	0.927	1.029
Cohabiting	0.946	0.858**	0.952
Male	1.185***	1.083*	1.101***
Age	1.032***	1.027***	1.032***
Age squared	1.000**	1.000*	1.000**
Coloured	0.871	2.090***	1.261
Indian	0.898	1.913**	0.729
White	1.715***	1.940***	1.839***
Primary education	1.261***	1.423***	0.994
Secondary education	1.762***	1.734***	1.420***
Further education	1.734***	1.828***	2.031***
Tertiary education	2.121***	2.196***	1.934***
Afrikaans	0.650***	0.595**	0.981
English	0.985	0.872	1.345*
Ndebele	0.434***	0.849	1.083
Xhosa	0.665***	0.548***	1.466***
N.Sotho	0.639***	0.768	1.013
S.Sotho	0.544***	0.756**	0.987
Tswana	0.616***	0.845	1.078
Swazi	0.708**	0.708*	1.217
Venda	0.470***	0.362***	1.815***
Tsonga	0.515***	0.913	1.138
Other	0.927	0.607	1.192
Unowned formal dwelling	0.856**	0.924	0.771***
Sub-let	1.054	0.943	0.771***
Informal dwelling	0.913	0.87	0.776***
Owns Vehicle	1.204***	1.356***	1.412***
Log hh expenditure	1.234***	1.328***	
Expen: R400-R799			0.983
R800-R1199			1.072
R1200-R1799			1.263***
R1800-R2499			1.324***
R2500-R4999			1.369***
R5000-R9999			1.438***
> R10000			1.266
Felt unsafe in neighbourhood	1.101	1.111	
Urban	1.557***	1.438***	1.760***
Constant	0.032***	0.029***	0.183***
Age turning point	52	67	53
Estimation sample	22 624	12 076	19 522

Reference: Number >65yr; African; no education; Zulu; expen R0-R399;

Dwelling = owned formal dwelling. Significance: * = 10%, ** = 5%, *** = 1%

Table 3.5 Second-stage response propensity: final nonresponse compared to bounded response: OHS 1997-OHS 1999

Covariate	OHS97	OHS98	OHS99
Household head	0.601***	0.921	0.552***
Self reporter			0.106***
Number kids	0.866	0.808	0.584***
Number 16–64yrs	0.915	0.896	0.683***
Household size	1.162	1.237	1.711***
Cohabiting	0.941	1.300*	0.739***
Male	1.159*	1.11	1.625***
Age	0.963	1.008	1.018
Age squared	1.001	1.0	1.0
Coloured	0.792	0.962	0.565
Indian	0.932	0.704	1.027
White	0.978	1.455	0.722
Primary education	0.988	1.053	0.589***
Secondary education	1.015	0.969	0.9
Further education	1.08	1.096	0.888
Tertiary education	1.352	0.923	0.896
Afrikaans	1.032	1.153	1.261
English	1.205	1.346	1.321
Ndebele	1.209	0.789	0.326*
Xhosa	0.609*	1.458	0.627***
N.Sotho	0.792	1.813	0.531***
S.Sotho	0.843	0.784	0.413***
Tswana	0.888	1.336	0.736
Swazi	0.447**	0.413	0.350***
Venda	1.221	1.885	0.219***
Tsonga	0.724	0.882	0.586*
Other	1.719	3.326*	2.691
Unowned formal dwelling	0.757	0.814	0.878
Sub-let	0.534*	0.609	1.046
Informal dwelling	1.196	1.223	0.651**
Owns vehicle	1.117	1.206	1.022
Log hh expenditure	0.845**	1.129	
Expen: R400-R799			0.624***
R800-R1199			0.526***
R1200-R1799			0.456***
R1800-R2499			0.282***
R2500-R4999			0.303***
R5000-R9999			0.344***

(continued)

Table 3.5 (continued)

Covariate	OHS97	OHS98	OHS99
> R10000			0.180***
Felt unsafe in neighbourhood	1.027	0.974	
Urban	0.478***	1.091	1.23
Constant	1.181	0.019***	0.248**
Age turning point	38	41	45
chi2	692	678	806
Effective subsample size	7 110	4 937	8 348
Estimation sample	22 624	12 076	19 522

Reference: Number >65yr; African; no education; Zulu; expen R0-R399;
Dwelling = owned formal dwelling. Significance: * = 10%, ** = 5%, *** = 1%

stable effect is observed for White people, where the odds of nonresponse is always greater than Africans. Education shows predictable effects given its correlation with income, with the odds of nonresponse increasing as education increases (relative to those with no education).

For the willingness to disclose variables, we see that rarely does any language have the same direction of influence across survey years, and sometimes the same language has statistically significantly negative odds in one year (relative to Zulu speakers), and statistically significantly positive odds in another year (e.g. Xhosa and Venda). This suggests that linguistic differences are ambiguous predictors of the first stage sequential response process.

For the neighbourhood safety variable, which is only available in the OHS97 and OHS98, we see that it is associated with about ten percent higher odds for nonresponse reporting, but the coefficient is not statistically significant in either year. On the other hand, an urban location is always statistically significant and always has greater odds for nonresponse reporting compared to exact response reporting.

For 1997 and 1998, variables that are thought to be correlated with income show the expected signs and significance, except the dwelling ownership and type variables. For 1999 the dwelling type variables show predicted effects and are significant. The reference category is an owned formal dwelling, a strong signal of wealth, so we would expect respondents who live in unowned formal dwellings, sub-let arrangements or informal areas to have lower odds of initial nonresponse, which is indeed the case. For those who own a vehicle, another stock of wealth variable, the odds of not providing an exact response are always higher than those who do not own a vehicle, and this result is statistically significant across the three years. Living in an urban area is a positive and significant predictor of nonresponse reporting in each year.

For household expenditure, when it is measured as a continuous variable, the results suggest that a one percentage point increase in expenditure increases the odds of nonresponse by 0.23% in 1997 and 0.33% in 1998. However, there seems to be

a nonlinear effect of expenditure on income reporting type, which is discernible only when expenditure is reported in brackets. Here, we see that while almost every expenditure category has higher odds for nonresponse and bounded response reporting relative to the R0-R399 expenditure category, the highest, open-ended expenditure category (>R10,000) has a lower effect size than the second highest category (R5,000-R9,999), and is not statistically significant (we return to this in the three-stage sequential response model below).

We now turn to the second stage of the sequential logistic response model. Here we are comparing nonresponse to bounded response, with the same set of explanatory variables as the first stage model. Nonresponse in 1999 conflates don't know responses with unspecified, whereas in 1997 and 1998 there are only unspecified responses for this subset.

What we're looking for in this second stage response model is any stable change in direction of the effects previously observed, which will tell us that the response process has changed as the response options evolve into the second income question. Important to note is that because we now exclude the exact subset of responses, the effective subsample size differs from the estimation subsample. The effective subsample includes only the bounded responses and nonresponse subsets of respondents in the second stage of the sequential model.[4]

Evident from Table 3.5 is that there are far fewer statistically significant coefficients across the entire range of predictors compared to the first stage model, except in 1999. In 1998 only two coefficients are significant, namely cohabiting and other language. At first consideration, the lack of significance doesn't seem to tell us much about this stage of the response process. But it is important to note that a lack of significance for so many covariates in the second stage suggests a very different response process to the follow-up employee income question. This would be equivalent to stating that the observed wealth effect in the first stage has been removed in the second income question, and that now both nonresponse and bounded response groups are indistinguishable on this set of predictors.

However, some caution is perhaps prudent here, for the findings in 1999 in particular are quite different to 1998 and 1997. The predictors themselves are also different, for in 1997 and 1998, self-reporter is not available while feeling unsafe in neighbourhood is available. The latter is insignificant in both years, as it was in the first stage response model (see Table 3.4), suggesting perhaps that it is an irrelevant variable in both stages of the employee income response process. On the other hand, self-reporter is highly significant in 1999, and is clearly a more relevant variable in these models. We shall examine this in more detail for the LFS surveys below.

In 1999, Table 3.5 shows that the cognitive burden variables are very important predictors of final nonresponse. A household head reduces the odds of nonresponse by about 45%, while a self-reporter reduces the odds of nonresponse ten-fold. Since

[4]Note that the effective subsample size is not available using Buis's (2012) algorithm for the sequential logistic response model. Here, and in every other table presented in this chapter, the effective subsample size is estimated by fitting separate logistic regression models to each stage of the sequential response process. The validity of doing so is given by Maddala (1983), and discussed in Sect. 3.3.3 above.

household size is held constant, the interpretation of the coefficients on the number of children and adults in the household is relative to them replacing a senior citizen (65 years or older). Thus, if a child was to replace a senior, it would reduce the odds of nonresponse by 42%, while an adult (aged 16–64) would reduce the odds of nonresponse by 32%.

The coefficient on household size reflects the addition of one more senior citizen because the number of children and adults are being held constant. Therefore, the addition of one senior citizen increases the odds of final nonresponse by 71%. The presence of senior household members is clearly correlated with greater reluctance to provide an income response, or greater confusion about that income (leading to a higher incidence of don't know responses).

Also in 1999, for the personal characteristics variables, cohabiting with a romantic partner reduces the odds of nonresponse by 26%. Men have odds that are 63% higher than women for final nonresponse, but the age, race and education variables are generally insignificant.

This is the first indication that the correlates of income variables may no longer be playing the powerful role in explaining the response process that they did in the first-stage model. If we consider the coefficients and significance of housing, vehicle ownership and expenditure variables, this effect would seem to be reinforced. Consequently, it suggests that variables that are correlated with income do not explain final nonresponse (alternatively we may simply not be able to measure this effect accurately). This is a very important finding, but preliminary at this point. We explore this further in the three-stage models below.

For the willingness to disclose variables, the effects for language is once again ambiguous, even though many of the coefficients are significant in 1999. Living in an urban area is significant in 1997, but the direction of influence changes across the survey years.

In summary, we can see that there are very different factors explaining the first stage of the sequential response model compared to the second stage. The qualifier on these findings, is that nonresponse in the final stage confounds don't know and refuse, providing limited insight into the construct of nonresponse itself. Below we are unconstrained by this, and explore the three-stage models for 1999–2003.

Three-Stage Sequential Logistic Response Model

In this section we present results for the three-stage models for the survey years 1999–2003. The first stage evaluates the determinants of initial nonresponse compared to exact responses; the second stage evaluates the determinants of final nonresponse against bounded responses, and the third stage decomposes nonresponse into refusals compared to don't know responses.

For the OHS 1999, which doesn't have an option for refusals in the questionnaire, we use the response group coded "unspecified" in the public-use dataset as the indicator of interest. This group of unspecified responses presumably conflates refusals with processing error. By analysing the predictors of this response type along with

the LFS, we have an opportunity to see if the same relationships hold over time. Note, however, that because of the lack of the refuse option in the OHS 1999, it is not strictly comparable to the LFS in the third-stage of the sequential response model, and we will interpret the results accordingly. For the first two stages of the model, the lack of a refuse option doesn't prejudice the comparability of the output.

Table 3.6 shows that for the cognitive burden variables, there are many significant effects, particularly during 2000–2002, but less so in 1999 and 2003. The household head variable is significant in every year until 2003, when its direction of influence changes. A self reporter is always significant and always reduces the odds of nonresponse. The household composition variables are not repeatedly significant across all survey years, but the direction of influence of additional kids or economically active people (16–64 years old) is almost always lower than the reference category of seniors. The household size variable is also not significant in 1999, 2002 and 2003. Cohabiting individuals reduce the probability of nonresponse, but the variable is only significant in 2000 and 2002. The importance of self-reporters in this section is noteworthy relative to the findings in 1997–1999.

For personal characteristics, men always have slightly higher odds of nonresponse, but this is not significant in every year. The coefficients on age are significant in every survey year except 2000, and for those years when it is significant, the turning point is approximately 47 years of age. The sign of the coefficients once again suggest an inverted-u shape to the relationship between age and response propensity, with the probability of refusing to answer the first income question increasing until 47, after which it decreases.

The race variables are fascinating. Coloured and White people have higher odds of nonresponse compared to Africans (though only the coefficients for Whites are significant in every year), but Indian people have significantly lower odds of nonresponse compared to Africans. This suggests that, all else equal, people of Indian or Asian descent in SA actually have a preference for reporting an exact response. Thus, rather than there being a socially sensitive dimension to the exact income question, for Indian people there seems instead to be a socially desirable dimension to it—a possible demonstration effect.

The education category dummies show the expected directional influence given their correlation to income, with effect sizes generally increasing over time. Thus, tertiary education respondents have much higher odds of initial nonresponse compared to those with no education. After primary school, all of the education categories have coefficients that are statistically significant in every year, suggesting stable direction of the effects relative to the base of no education (except in 1999), even though the coefficients are quite different in magnitude.

For other variables that are correlated with income—including housing type and ownership, vehicle ownership and total household expenditure—the coefficients are also always in the expected direction and always significant (with one or two exceptions) in every survey year. This is perhaps the most important affirmation that, for initial nonresponse at least, it is strongly related to higher income levels. The exception to this is the finding for Indian people, who are on average the second wealthiest population group in South Africa after Whites, but here demonstrate behaviour that

Table 3.6 First-stage response propensity: initial nonresponse compared to exact responses: 1999–2003

Covariate	OHS99	LFS00	LFS01	LFS02	LFS03
Household head	0.931*	0.883*	0.901**	0.910**	1.059
Self reporter	0.706***	0.863**	0.653***	0.662***	0.702***
Number kids	0.957	0.868*	0.847**	0.904	0.922
Number 16–64yrs	0.966	0.856**	0.921	0.938	1.03
Household size	1.033	1.178**	1.133**	1.09	1.048
Cohabiting	0.944	0.876**	0.924	0.871***	0.933
Male	1.100***	1.185**	1.109**	1.186***	1.063
Age	1.029**	1.011	1.047***	1.068***	1.037***
Age squared	0.9997**	0.9999	0.9995***	0.9993***	0.9996**
Coloured	1.275	1.394	1.742***	1.396*	1.680***
Indian	0.771	0.382***	0.480***	0.498***	0.613**
White	1.862***	1.954***	1.699***	2.203***	2.433***
Primary	0.988	1.207	1.161	1.553***	1.206
Secondary	1.426***	1.522***	2.228***	3.024***	2.393***
Further	2.025***	1.929***	3.594***	4.911***	4.209***
Tertiary	1.990***	2.335***	3.794***	5.492***	4.559***
Afrikaans	0.979	1.168	1.13	0.798	0.577***
English	1.370*	1.548	1.962***	1.461**	1.288
Xhosa	1.482***	1.115	1.473***	1.145	0.844*
Other	1.089	0.996	1.1	1.187**	0.796***
Unowned formal dwelling	0.767***	0.616***	0.969	0.853**	0.767***
Sub-let room or dwelling	0.767***	0.605***	0.655***	0.781**	0.666***
Informal area dwelling	0.764***	0.583***	0.657***	0.733***	0.684***
Expen: R400-R799	0.973		0.977	1.140*	1.345***
R800-R1199	1.056		1.251**	1.413***	1.906***
R1200-R1799	1.242***		1.357***	1.722***	2.077***
R1800-R2499	1.276***		1.372***	2.196***	2.198***
R2500-R4999	1.320***		1.260**	2.225***	2.739***
R5000-R9999	1.410***		1.313**	2.593***	3.144***

(continued)

Table 3.6 (continued)

Covariate	OHS99	LFS00	LFS01	LFS02	LFS03
> R10000	1.215		1.540**	2.777***	2.754***
Log hh expenditure		1.187***			
Owns Vehicle	1.438***	1.041	1.238***	1.494***	1.454***
Urban	1.709***	1.569***	1.203**	1.185**	1.337***
Constant	0.206***	0.007***	0.033***	0.018***	0.036***
Age turning point	48	57	46	47	46
Estimation sample	19 802	20 083	20 030	19 550	19 417

Reference: Number >65yr; African; no education; Zulu; expen R0-R399; dwelling = Owned formal dwelling. Significance: * = 10%, ** = 5%, *** = 1%

suggests a cultural difference in their attitude to social sensitivity. Because we are controlling for the partial effect of language and race in these models (note that in these three-stage sequential logistic models, a more aggregated language variable (see Table 3.1) is used to ensure large enough cell counts for the models to run), the finding for Indian people can be interpreted as a socio-cultural effect, and is highly noteworthy.

We now turn to the second stage of the sequential response model, which evaluates final nonresponse (including refusals combined with don't know responses) compared to bounded response. Table 3.7 presents the results.

Evident from the table is that the cognitive burden variables are important predictors of final nonresponse compared to bounded response. The household head and self reporters always have lower odds of nonresponse, and these coefficients are statistically significant in every year except in 2003 for the household head. However, for the household composition variables, the effects are not significant in 2000 and 2001, though the coefficients go in the same direction as every other year. Similarly, for household size, in 2000 and 2001 the effects are in different directions and not significant, whereas they are both positive and significant in other years. For cohabiting status, 2000 and 2003 have insignificant results and the effect is in different direction in 2000, while for the remaining years they reduce the odds of nonresponse and are significant.

The results for personal characteristics variables, including gender, age, race and education are rarely consistently statistically significant over all years, and the coefficients for language show no consistent direction of influence over time. The failure of age to play a significant role in the second stage of the response process (except in 2001) is identical to the second stage of the response models for OHS97-99 presented in Table 3.5 above, suggesting that it plays a diminished or non-existent role in explaining further nonresponse beyond the first stage of income reporting.

Table 3.7 Second-stage response propensity: final nonresponse compared to bounded responses: 1999–2003

Covariate	OHS99	LFS00	LFS01	LFS02	LFS03
Household head	0.576***	0.505***	0.711***	0.677***	0.925
Self reporter	0.252***	0.687*	0.508***	0.434***	0.536***
Number kids	0.658***	0.938	0.852	0.781*	0.652***
Number 16–64yrs	0.719***	0.958	0.898	0.876	0.766**
Household size	1.556***	1.002	1.176	1.264*	1.438***
Cohabiting	0.726***	1.122	0.741***	0.677***	0.957
Male	1.424***	1.188	1.216**	1.546***	1.067
Age	1.006	0.935	0.967	1.059**	0.987
Age squared	1.0000	1.0008	1.0005	0.9994*	1.0001
Coloured	0.871	1.761	1.375	1.613	0.877
Indian	1.575	3.485	0.54	0.736	1.272
White	1.037	1.969	1.35	2.180**	1.479
Primary	0.640***	1.314	0.596*	1.212	1.433
Secondary	0.946	1.41	0.985	1.188	1.869*
Further	0.831	1.595	0.79	1.179	1.910*
Tertiary	1.125	1.604	1.072	0.867	1.909*
Afrikaans	0.963	4.625*	1.075	1.848	1.646
English	1.185	6.339**	2.054*	1.795	1.779
Xhosa	0.759*	3.236*	1.421	1.882**	1.206
Other	0.612***	2.644*	1.603**	2.123***	1.116
Unowned formal dwelling	0.897	0.639	0.889	0.912	0.793*
Sub-let room or dwelling	1.018	0.684	1.024	1.633**	1.031
Informal area dwelling	0.624***	0.627	1.039	0.756	0.788
Expen: R400-R799	0.683**		0.693*	0.945	0.791
R800-R1199	0.568***		0.660**	0.678*	0.531***
R1200-R1799	0.502***		0.916	0.841	0.348***
R1800-R2499	0.306***		0.794	0.648*	0.420***
R2500-R4999	0.312***		0.669*	0.733	0.362***
R5000-R9999	0.388***		0.466***	0.715	0.321***

(continued)

Table 3.7 (continued)

Covariate	OHS99	LFS00	LFS01	LFS02	LFS03
> R10000	0.212***		0.395**	0.461**	0.424***
Log hh expenditure		0.664***			
Owns Vehicle	1.137	0.989	1.340*	1.183	1.054
Urban	1.084	0.544	0.995	1.673***	1.645***
Constant	0.374*	7.741	0.330*	0.018***	0.150***
Age turning point	697	42	34	48	67
Effective subsample	8 628	1 986	4 538	5 361	5 839

Reference: Number >65yr; African; no education; Zulu; expen R0-R399; dwelling = Owned formal dwelling. Significance: * = 10%, ** = 5%, *** = 1%

The housing wealth dummies are also almost never significant, nor the vehicle ownership variable (except in 2001). However, the expenditure variables are frequently significant, especially in the highest income category which is significant in every year. The direction of the effect is surprising though, for it seems that as total household expenditure goes up, the odds of nonresponse go down. The coefficient on the log of expenditure also suggests lower odds for nonresponse reporting as expenditure increases.

The take-home message from the second stage of the response model is that the odds of final nonresponse do not seem to increase with income. The most consistent effects over time are for the cognitive burden variables, notably self reporter followed by household head. The lack of explanatory power in the wealth variables suggests that the follow-up employee income question that presents the showcard to the respondent is very successful in persuading higher income individuals to disclose their earnings, albeit as a bounded response. This would suggest that any remaining nonresponse should no longer be unambiguously positively correlated with income. We now turn to exploring this in the third stage of the sequential response model.

Table 3.8 shows the results of the third stage response model, where the dependent variable decomposes final nonresponse into refusals compared to don't know responses, except in 1999 where unspecified responses confound refusals with other possible sources of missing data, such as processing error or measurement error. However, there are generally no stable predictors over time in this stage of the response process despite a standardised instrument between 2000–2003. Small sample sizes also suggest weaker power in these models.

In this table we also start seeing very large effect sizes for certain variables. The large coefficient sizes are potentially indicative of small cell sizes in this stage of the response model, leading to near perfect prediction of the outcome. To get some idea about whether it is a small sample size that is driving this, the effective sample size at the bottom of the table is useful to consult, as is Table 3.2 above, which provides the

counts of each response type that constitute the dependent variables in these models. As far as the effective subsample size is concerned, the results for 2000 demonstrate that it has the smallest sample of nonresponse groups, and is very different to every other survey year. We evaluate further diagnostics of these models in the next section of this chapter below.

Among the cognitive burden questions, only self-reporter is repeatedly significant (except in 2000), and it increases the odds of refusing by the largest order of magnitude. The strength of the self-reporter variable is unsurprising though because those respondents who are proxy reporters are much less likely to know the incomes of other household members, whereas self-reporters are much more likely to refuse on social sensitivity grounds. Hence the large coefficients are to be expected here, though a magnitude of 33 times the odds (in 2001) is surprising in light of the relatively large effective sample size (of 864 observations, roughly equally distributed between don't knows and refusals—see Table 3.2).

For personal characteristics variables, there is no stable effect for age, sex or race, with odds ratios often below one for a given year and then above one for the next year. For age and age squared, it is not meaningful to discuss the turning points as the results are insignificant for all survey years. Education categories have odds ratios generally greater than one, and in 2002 the results are large and significant. The very large coefficients for education in 2000 suggest small cell sizes in this year in particular.

For the willingness to disclose variables, language is again inconsistent over time, while living in an urban location is almost always significant, but the direction of influence on the odds change from negative to positive and back again over time.

For the correlates of income, the results for expenditure in 2002 and 2003 suggest an increasing chance of refusing as expenditure increases, but the results are not always significant at the lower expenditure categories. However, owning a vehicle and housing wealth is almost never significant, suggesting an absence of a wealth effect on the odds of refusing.

The overall conclusion to this stage of the response model is that self-reporting is the major explanatory factor impacting upon the probability to refuse to answer the income question. The wealth effect seems to be absent, while a positive but non-monotonic relationship with household expenditure seems to be present, a slightly contradictory set of results.

Finally, an important concern that arises in each of the sequential response models, but particularly in the case of the third stage models where the effective sample size is smallest, is the interrelationship between covariate nonresponse on expenditure and nonresponse on income. If these two forms of missingness are correlated, then it is possible for a simultaneity problem to exist that could lead to biased results. We now turn to evaluating this question along with other diagnostic tests of the response models.

Table 3.8 Third-stage response propensity: refuse compared to don't know responses: 1999–2003

Covariate	OHS99	LFS00	LFS01	LFS02	LFS03
Household head	1.058	2.948	1.028	1.638*	1.075
Self reporter	8.207***	1.634	33.729***	17.120***	27.691***
Number kids	1.342	0.954	0.776	1.064	1.38
Number 16–64yrs	1.08	0.845	0.837	0.908	1.009
Household size	0.787	1.007	1.324	0.879	0.713
Cohabiting	1.187	0.479	1.465	2.520***	2.530***
Male	0.662**	0.564	0.732	0.767	1.1
Age	0.923	1.108	0.981	0.923	1.043
Age squared	1.0010	1.0001	1.0003	1.0012	0.9992
Coloured	1.077	14.883**	3.634*	0.615	0.354
Indian	1.176	27.157	1.57	0.674	1.872
White	0.82	17.466**	3.505*	0.993	0.533
Primary	1.278	8.865	0.756	5.184**	6.878
Secondary	0.976	59.648*	1.299	6.145**	10.712
Further	1.048	78.110*	2.075	5.309**	9.881
Tertiary	1.952	12.933	2.167	6.618**	9.612
Afrikaans	1.862	0.78	3.166	1.583	3.04
English	3.883*	0.756	5.945**	1.201	4.959**
Xhosa	2.449***	0.504	3.178**	0.839	1.08
Other	2.136**	0.494	2.683*	0.673	0.503
Unowned formal dwelling	1.139	1.611	1.379	0.839	1.058
Sub-let room or dwelling	1.052	0.179	3.321**	1.191	1.52
Informal area dwelling	1.114	4.408	1.049	0.613	1.538
Expen: R400-R799	1.433		1.318	3.575*	3.501*
R800-R1199	1.568		2.005	4.803**	7.495***
R1200-R1799	1.45		3.003**	7.160***	5.024**
R1800-R2499	1.45		2.314*	6.314***	4.282*
R2500-R4999	1.215		2.201	7.512***	8.196***
R5000-R9999	1.64		1.546	8.164***	6.600**

(continued)

Table 3.8 (continued)

Covariate	OHS99	LFS00	LFS01	LFS02	LFS03
> R10000	1.226		8.531**	8.307***	8.318**
Log hh expenditure		1.738			
Owns Vehicle	1.054	2.274	1.781*	1.536	1.426
Urban	0.561**	0.130*	1.048	3.083***	2.274**
Constant	0.833	0.000**	0.011**	0.016***	0.004**
Age turning point	40	511	32	33	26
chi2	817.1	556.0	1195.3	1749.3	1710.4
Effective subsample	1 088	123	704	864	950
Estimation sample	19 802	20 083	20 030	19 550	19 417

Reference: Number >65yr; African; no education; Zulu; expen R0-R399; dwelling = Owned formal dwelling. Significance: * = 10%, ** = 5%, *** = 1%

3.4.3 Diagnostics of the Sequential Response Models

In this section we evaluate model fit and the sensitivity of the results above to simultaneous income and expenditure missing data. This helps shed light on the limitations of the analysis, and provides some useful insights for further research.

Model Fit

In this section we discuss model fit for the sequential logistic response models estimated in the main text of this chapter by presenting Hosmer-Lemeshow (H-L) statistics. The sequential logistic model fitted to the data is estimated as a system of equations in the algorithm by Buis (2012). Theoretically, however, it is also possible to derive the same results by fitting binary logistic models to each stage of the sequential response process. This is immediately evident from Eq. 3.1 above.

The H-L test results in Table 3.9 are calculated as post-estimation statistics after binary logistic models for each stage of the sequential response models are fitted to the data. The pseudo R^2 values from those models are also presented as a further model diagnostic.

The table shows the response stage for each year investigated, the number of observations involved in the post-estimation procedure after each binary logistic model is fitted in order to calculate the H-L statistic, the number of groups used, the H-L statistic itself with p-value, and the pseudo R^2. Large H-L statistics and small p-values indicate a lack of fit of the model.

The results from Table 3.9 suggest that the models *do not* fit the data well in the first stage of the sequential response process in every survey year except 2002. This is unsurprising because multiple response groups are collapsed into the dependent variables of the first stage models, namely bracketed responses, don't know, refuse and/or unspecifieds, which are all compared against exact responses (the base outcome in the first stage). It is only from the second stage of the response process that the models begin to fit well.

For the second and third stages the H-L tests suggest that we fail to reject the null of good model fit in all survey years except in the third stage of 2001 (at the 5% significance level). It should be noted that the small sample size in 2000 indicates weak statistical power of the H-L test in this year, but for every other year this is unlikely to be the case.

However, the pseudo R^2 values suggest that the specification of the models best explain the variance of only the third stage of the response process: that is, predictors

Table 3.9 Hosmer-Lemeshow (H-L) test for model fit and pseudo r squared in logistic regression of each sequential response stage

Year-response stage	No. Obs	No. Groups	H-L chi^2	Pr. > chi^2	Pseudo R^2
1997–1	22 624	10	14.07	0.080	0.085
1997–2	7 110	10	12.31	0.138	0.044
1998–1	12 076	10	19.71	0.012	0.109
1998–2	4 937	10	6.99	0.538	0.028
1999–1	19 802	10	14.05	0.080	0.098
1999–2	8 348	10	10.67	0.221	0.132
1999–3	1 088	10	5.18	0.738	0.201
2000–1	20 083	10	13.39	0.099	0.095
2000–2	1 986	10	11.07	0.198	0.078
2000–3	123	10	7.95	0.438	0.399
2001–1	20 030	10	39.36	0.000	0.119
2001–2	4 538	10	11.58	0.171	0.060
2001–3	704	10	16.52	0.036	0.411
2002–1	19 550	10	11.2	0.191	0.170
2002–2	5 361	10	11.98	0.152	0.086
2002–3	864	10	13.66	0.091	0.376
2003–1	19 417	10	26.6	0.001	0.188
2003–2	5 839	10	5.14	0.743	0.055
2003–3	950	10	9.82	0.278	0.440

Response Stage 1: missing + bracket compared to continuous

Response Stage 2: missing compared to bracket

Response Stage 3: refuse compared to don't know

of refusals compared to don't knows. For the first and second stages, the pseudo R^2 is typically very weak. Important to note here is that on statistical grounds, the pseudo R^2 is not a particularly informative statistic for discrete (and particularly binary) dependent variable regression models due to the limited variation in the dependent variable itself. Nevertheless, its magnitude does impart some information on how the response models perform.

The Sensitivity of Model Estimates and Inferences to Omitted Expenditure

It is important to conduct an analysis of simultaneous nonresponse on employee income and expenditure because these two variables are correlated and expenditure is an explanatory variable in every response propensity model. The role of the total household expenditure variable in these models is to provide us with a correlate to individual employee income, but the capacity of this variable to do its job effectively is reduced if nonresponse on it occurs jointly with nonresponse on income.

It should be noted that while employee income is measured at the individual level for the employed economically active population, expenditure is measured at the household level. Therefore, the extent to which these two variables are correlated will be higher in smaller households.

Table 3.10 presents the percentages of joint nonresponse for each survey year and the denominator subsample size in the percentage calculations.

The changing form of the expenditure variable over time provides for different levels of detail in this analysis. Firstly, when total household expenditure is a continuous variable, then the only form of nonresponse that we observe on it is an unspecified response. This is compared against the number of don't know, refuse and unspecifieds on income. The number jointly observed as nonresponse on expenditure and income then enters into the numerator of the percentage calculation, while the total number of don't know, refuse and unspecified responses for employee income enters the denominator. From this we see that for the OHS97, OHS98 and LFS00, simultaneous nonresponse on income and expenditure accounts for between 17 and 26% of all nonresponse.

These numbers can be further decomposed when a bounded expenditure bracket is asked for rather than an exact response, because additional response options exist in the expenditure question for don't know and refuse. As with income in the OHS99, the expenditure question also does not have an option for "refuse", which was only introduced in the LFS questionnaires. The most important row of Table 3.10 for the OHS99 and LFS00-03 is the last one, in which all forms of nonresponse on expenditure is compared to all forms of nonresponse on income. Here we see that simultaneous nonresponse is in fact much larger than for the continuous expenditure variable in every year investigated, averaging about 30% of all nonresponse on income in the LFS, but rising to a very high 47% in the OHS99.

The first-order impact of nonresponse on expenditure in the regression models is to reduce the estimation sample size by the number of nonrespondents on expenditure.

Table 3.10 Jointly observed nonresponse subsets for expenditure and income

Survey year	OHS 97	OHS98	LFS00	
Percent missing on ln expen & NR on income	25.5	17.7	19.1	
Subsample size of NR on income	942	628	677	
Survey Year	OHS99	LFS01	LFS02	LFS03
Percent DK on expen category & DK on income	42.6	22.1	20.7	15.7
Subsample size of DK on income	1588	521	651	485
Percent R on expen category & R on income	n/a	28.5	28.6	31.8
Subsample size of R on income	n/a	578	664	891
Percent DK+R expen category & DK+R+				
Unspecified on income	46.5	31.1	31.0	28.6
Subsample size of DK+R+Unspec on income	2136	1176	1355	1399

In the limiting case, if all nonrespondents on household expenditure were the highest income earners, then the loss of covariate information for these cases could introduce biases into the sequential response models. But since the numbers here are quite low, this concern is mitigated to some extent, particularly in the first and second stages of the sequential logistic response models where the subsample sizes are always in the several thousands for each survey year.

However, expenditure nonrespone becomes non-trivial in the third stage of the sequential response models when the outcome variable is refusals (for the LFS, unspecifieds in 1999) compared to don't know responses. From Table 3.10, we can see the potential estimation sample sizes for the outcome variable sometimes involves observations counts in the hundreds. Here, nonresponse on household expenditure will play an important role because it reduces the estimation sample size for all other covariates too, and to the extent that these covariates also help predict refusals and don't know responses in the income question, the explanatory power of the models—and for refusals compared to don't know responses in particular—is compromised.

We therefore re-estimate the three-stage sequential response model of Sect. 3.4.2, omitting the expenditure variables from each year. Table 3.11 presents the results for

Table 3.11 Third-stage response propensity: refuse compared to don't know responses omitting expenditure

Covariate	OHS99	LFS00	LFS01	LFS02	LFS03
Household head	0.854	1.451	1.007	1.279	1.135
Self reporter	7.747***	2.264	31.363***	19.114***	29.059***
Number kids	1.396*	0.805	1.11	1.251	1.231
Number 16–64yrs	1.02	1.116	1.041	1.071	1.007
Household size	0.772	0.954	0.943	0.731	0.767
Cohabiting	1.255	1.745	1.475*	2.555***	2.623***
Male	0.903	0.743	0.82	0.774	1.043
Age	0.926*	0.971	0.969	0.953	0.969
Age squared	1.001*	1.001	1	1.001	1
Coloured	1.111	3.538	2.19	1.813	0.387
Indian	1.512	11.982	1.401	1.516	1.515
White	1.446	7.106**	2.434	2.184	0.793
Primary	1.343	25.812*	0.954	0.897	33.520***
Secondary	1.123	78.826**	1.38	1.472	39.249***
Further	1.118	108.974**	1.987	1.113	37.352***
Tertiary	1.756	58.753	1.559	1.315	33.630***
Afrikaans	1.581	4.634	3.571*	0.732	2.763
English	2.299	4.106	5.325**	0.538	4.834**
Xhosa	1.706	2.993	1.926	0.675	0.676
Other	1.756*	1.068	1.905	0.506	0.534
Unowned formal dwelling	1.03	0.61	1.11	0.670*	1.13
Sub-let	0.944	0.141**	1.878	0.972	1.096
Informal dwelling	0.878	2.031	0.553	0.446	0.803
Owns Vehicle	1.198	1.771	2.167***	1.911***	1.985**
Urban	0.645**	0.174**	1.285	2.607***	1.973**
Constant	1.032	0.002*	0.071*	0.433	0.019***
chi2	935.286	685.396	1275.421	1797.115	1788.431
N	21433	20419	20754	20198	19959
Gain in Obs cf Table 3.8	1631	336	724	648	542

Reference: Number >65yr; African; no education; Zulu; expen R0-R399; dwelling = Owned formal dwelling. Significance: * = 10%, ** = 5%, *** = 1%

the third stage of the response model only.[5] By way of summary, in the first and second stages of the model, almost all coefficients were in a similar direction. More common was that the significance levels changed, and this occurred for about 10% of the coefficients, though never consistently over time. However, for the third stage of the model, there are important changes in the direction of influence of coefficients and in statistical significance.

Table 3.11 shows the results of the third stage of the sequential response model when expenditure is omitted from the specification. At the bottom of the table, we introduce a row that shows the gain in estimation sample size attributable to omitting expenditure from the model. This number ranges from 336 in 2000 to 1631 in 1999, the latter clearly more likely to influence results than the former.

Comparing the results of this stage of the model with its counterpart in Table 3.8 shows somewhat similar findings, but given that the main finding in Table 3.8 was that there were no stable findings across the years, this is not particularly informative. One identical effect in Table 3.11 is for the self reporter variable, where the coefficient sizes are again very large and significant in the same four years as in Table 3.8 (i.e. 1999, 2001–2003).

In the two years when the expenditure category is always significant in Table 3.8, namely 2002 and 2003, the effect of omitting expenditure is to deflect its influence into other variables in the model. In 2002, vehicle ownership and unowned formal dwellings becomes significant when they were not before. On the other hand, the education variables reduce in magnitude and become insignificant when expenditure is omitted.

One interesting effect in Table 3.11 is for education in 2003, where the coefficients have now nearly doubled in magnitude and become significant (compared to Table 3.8). To the extent that education is picking up a correlate of income effect, the omitted expenditure variable may be influencing the results for education. However, because this only happens in 2003, it is not possible to generalise the result. Nevertheless, it does suggest that the effect of omitting expenditure in the sequential response models is not trivial, and may cause more problems than it solves in certain survey years.

3.5 Conclusion

The main objective of this chapter was to carefully establish the interrelationship between questionnaire design and response propensities in order to identify the characteristics of respondents that have the highest probability of not responding to the employee income question. Analytically, an important part of the analysis was to assess the stability of the effects over multiple time points. Two periods were distinguished: (a) 1997–1999, which allowed us to evaluate how improvements to the income question affected our understanding of the response process, and how the

[5] For the first and second stages of the sequential response model excluding expenditure, results will not be presented (but are available from the author.).

addition of the self-reporter option and omission of unsafe neighbourhood influenced our understanding of income response type; and (b) 2000–2003, which allowed us to evaluate the stability of groups of predictors over time given a fixed instrument. The latter ensured that the findings were not exclusively due to transient empirical fluctuation in any given year.

Improvements to the design of the income question unambiguously positively impacted the ability to understand nonresponse on it. This was particularly so for decomposing final nonresponse into both refusals and don't knows. In 1999, when only the don't know option was provided, unspecified responses seemed to mimic the patterns associated with those who refuse to answer the question for the first two stages of the sequential response models, but by the third stage began to differ in the signs and significance of important covariates. The addition of a self-reporter indicator in the questionnaire was equally important for explaining final income nonresponse in all survey years, except 2000 which was clearly an anomaly in the history of Statistics South Africa's surveys.

The sequential logistic response model proved to be a suitable estimator for response propensities to employee income when it was measured by an initial exact prompt followed by a showcard bracketed follow-up prompt. The overall results from the first stage of the sequential response models was that initial nonresponse was strongly associated with variables correlated with income. This result was stable over almost every survey year from 1997–2003. There was also an interesting social desirability or demonstration effect discernible for people of Indian/Asian descent in this first stage response process, though this was most apparent in the LFS.

However, in the second stage, there seemed to be a reversal of the finding that response propensities were correlated with income. Instead, a rise in the importance of household characteristics and self-reporting was apparent. What this implied was that the follow-up income question actually overturned initial refusals from higher earning respondents, and therefore neutralised the correlate of income effect in the (non)response process.

The third-stage response propensities showed that, with or without expenditure included in the specification, the results were unstable across the years except for self-reporting, which was large and significant in every survey year except 2000. A small sample size is the most likely explanation for the anomalous results in 2000. Notable for this stage of the response models was the strength of the Hosmer-Lemeshow tests and pseudo r-squared statistics. But the fact that no subset of predictors remained consistently statistically significant across the years suggests some variation in this part of the missingness mechanism over time.

Finally, it should be remembered that a limitation with this analysis is the inability to observe variables related to (1) the characteristics of the interviewer conducting the survey, and (2) the respondent's behaviour during the survey. These (omitted) variables could have helped better explain the final refusal response in particular.

References

Beatty, P., & Herrmann, D. (2002). To answer or not to answer: Decision processes related to survey item nonresponse. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse.* New Jersey: Wiley.

Blair, J., Menon, G. & Bickart, B. (1991). Measurement effects in self versus proxy responses to survey questions: An information-processing perspective. In P. P., Biemer, R. M., Groves, L. E., Lyberg, N. A., Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys*, New Jersey: Wiley.

Buis, M. (2012). seqlogit: Stata module to fit a sequential logit model, Version number 1.1.15. http://maartenbuis.nl/software/seqlogit.html

Buis, M. (2011). The consequences of unobserved heterogeneity in a sequential logit model. *Research in Social Stratification and Mobility, 29*(3), 247–262.

Cantwell, P. J. (2008). Rotating Panel Design. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods.* Thousand Oaks: Sage Publications.

Casale, D., & Posel, D. (2005). *Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa*. Mimeo, Durban: University of Kwazulu-Natal.

Daniels, R. C., & Wittenberg, M. (2010). *Sampling methodologies in Statistics South Africa household surveys: A conversation with David Stoker*. Mimeo, Cape Town: Data First, University of Cape Town.

De Leeuw, E., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse.* New Jersey: Wiley.

Dillman, D. A., Eltinge, J. L., Groves, R. M., & Little, R. J. A. (2002). Survey nonresponse in design, data collection, and analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse.* New Jersey: Wiley.

Frederick, S., Kahneman, D., & Mochona, D. (2010). Elaborating a simple theory of anchoring. *Journal of Consumer Psychology, 20*(1), 17–19.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New Jersey: Wiley.

Hurd, M., Juster, T. F., & Smith, J. P. (2003). Enhancing the quality of data on income: Recent innovations from the HRS. *The Journal of Human Resources, 38,* 758–772.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21*(11), 1161–1166.

Johnson, T. P., O'Rourke, D., Burris, J., & Owens, L. (2002). Culture and survey nonresponse. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse.* New Jersey: Wiley.

Juster, F. T., Cao, H., Couper, M., Hill, D., Hurd, M., Lutpon, J., Perry, M., & Smith, J. (2007). *Enhancing the quality of data on the measurement of income and wealth*. Mimeo, Michigan Retirement Research Center, Ann Arbor: University of Michigan.

Juster, T. F., & Smith, J. P. (1997). Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association, 92*(440), 1268–1278.

Juster, F. T., Smith, J. P., & Stafford, F. (1999). The measurement and structure of household wealth. *Labour Economics, 6,* 253–275.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

Press, S. J., & Tanur, J. M. (2005). An overview of the respondent-generated intervals (RGI) approach to sample surveys. In *American Statistical Association (ASA) Section on Survey Research Methods, Proceedings* (pp. 3487–3493).

Press, S. J. (2004). Respondent-Generated Intervals (RGI) for Recall in Sample Surveys. *Journal of Modern Applied Statistical Methods, 3*(1), 104–116.

Press, S. J., & Marquis, K. H. (2001). Bayesian estimation in a US Census Bureau survey of income recall using respondent-generated intervals. *Research in Official Statistics, 1,* 151–168.
Press, S. J., & Tanur, J. M. (2004). Relating respondent-generated interval questionnaire design to survey accuracy and response rate. *Journal of Official Statistics, 20*(2), 265–287.
Rubin, D. B., Stern, H. S., & Vehovar, V. (1995). Handling "Don't Know" survey responses: The case of the Slovenian plebiscite. *Journal of the American Statistical Association, 90*(431), 822–828.
Schwartz, L., & Paulin, G. (2000) Improving response rates to income questions. In: *American Statistical Association (ASA) Section on Survey Research Methods, Proceedings* (pp. 965–970).
Schwarz, N., & Hippler, H.-J. (1991). Response alternatives: The impact of their choice and presentation order. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys.* New Jersey: Wiley.

Chapter 4
Univariate Multiple Imputation for Coarse Employee Income Data

4.1 Introduction

Employment income data are coarsened as a result of questionnaire design. In the previous chapter we saw that Statistics South Africa (SSA) ask two employment income questions: an exact income question with a showcard follow-up. In public-use datasets, this results in two income variables: a continuously distributed variable for exact income responses and a categorical variable for bounded income responses with separate categories for nonresponse. It is the task of the researcher to then generate a single income variable that effectively deals with this mixture of data types. Following Heitjan and Rubin (1991), we call a variable with this mixture of data types "coarse data".

Coarse income data pose non-trivial implications for researchers concerned with analysing that data. The primary problem that arises from an inconsistent treatment of this variable is that parameter estimates may be biased and dependent on the particular researcher's choice of method to overcome the problems posed by the instrument's design and resulting data structure. This leads to potentially erroneous inferences on important univariate parameters of the income distribution, including quantiles and moments.

Multiple imputation is potentially an effective solution for coarse data problems (Heitjan and Rubin, 1990; Heitjan, 1994). It involves substituting coarse data values with plausible draws of those values multiple times. Multiple imputation has been applied to coarse wealth data by Heeringa (1995, 2002), and it has been applied to coarse earnings data by Daniels (2008) and Vermaak (2010). Ardington et al. (2006) conducted multiple imputation for total income. However, because multiple imputation is effectively a simulation-based technique (Schafer, 1999), it is very dependent on the setup of the imputation process and can frequently perform sub-optimally for reasons that may not be easy to identify. Van Buuren et al. (1999), Royston (2004), White et al. (2007) and Graham et al. (2007) discuss various aspects of the multiple imputation process that can affect the reliability of imputed draws

R. C. Daniels, *How Data Quality Affects our Understanding of the Earnings Distribution*,
https://doi.org/10.1007/978-981-19-3639-5_4

and statistical inference, such as covariate selection, the imputation algorithm itself and the numbers of imputations needed for reliable inference.

In this chapter the imputation algorithm is simplified by imputing univariately for coarse income data only, rather than also imputing covariate missing data. This has both advantages and disadvantages. The main disadvantage is that it removes all units with covariate nonresponse from the estimation sample, which is equivalent to treating covariate nonresponse as missing completely at random (MCAR). The cost of doing this is dependent on the application, with Allison (2000) noting that more sophisticated treatments of covariate nonresponse can impose equally stringent (but often more opaque) assumptions on the data. However, a distinct advantage of multiple imputation is that imputed draws can be made for many variables with missing data simultaneously, making it computationally efficient. There is, therefore, a definite trade-off in ignoring covariate nonresponse.

The main advantage of imputing multiple times for a single variable is that it allows us to be far more precise about exactly which aspects of the multiple imputation algorithm lead to implausible results. The two primary dimensions of the imputation algorithm that will be explored are specification of the prediction equations and sensitivity of the results to the number of imputations. The reason we need this precision is because, as shown in the previous chapter on questionnaire design and response propensities, we saw that respondents who chose to answer the bounded income question generally were higher income individuals. However, when we accounted for predictors of higher incomes in the sequential response propensity models, it was revealed that the final nonresponse subset had refusals that were largely indistinguishable from don't know responses on observable covariates. It was this finding that led to the suggestion that final nonresponse was likely an ignorable form of nonresponse.

In this chapter a key objective is to assess where in the income distribution the bounded, refuse, don't know and unspecified subsets of the employment income question lie when we generate plausible values of their incomes using multiple imputation. The coarse data framework allows us to characterise the nature of the problem in a theoretically sound manner. The simplified univariate multiple imputation algorithm then allows us to test the sensitivity of inferences to covariate selection and the number of imputations. The usefulness of doing this is that we learn how robust imputations are to mis-specification. Lessons learnt from this process can then feed into more complex multivariate missing multiple imputation exercises.

In order to examine the performance of the imputation algorithm, we test four different specifications of the prediction equations: one that is completely mis-specified to establish a baseline of how wrong the imputed draws can be; one with covariates selected identically to the response propensity models of the previous chapter; one with Mincerian earnings function based covariates; and one with a combination of response propensity and Mincerian earnings function covariates, which we treat as the first-best specification method for reasons discussed below.

Data for this exercise is identical to the previous chapter: the October Household Surveys (OHS, 1997–1999) and Labour Force Surveys (LFS, 2000–2003 September

Waves only). As with the previous chapter, the sample is restricted to economically active (16–64 year old) employees only. We can therefore also observe how improvements to the income question over time affect the imputation process.

4.2 Preliminaries

4.2.1 Coarse Income Data

A variable with continuous, bounded and missing observations is not simply an example of nonresponse, but in fact a more complicated problem known in the literature as "coarse data". The theory of coarse data stems in part from the theory of missing data, which was principally developed by Rubin (1976, 1987). However, "coarse data" is in fact a generalisation of the various ways that data may not reflect their true values, and includes as special cases rounded, heaped, censored, partially categorised and missing (i.e. completely coarse) data (Heitjan and Rubin, 1991).

Two principal papers established the theory of coarse data: Heitjan and Rubin (1991) and Heitjan (1994). To show the direct precedents to missing data theory, it is useful to note that the theory of coarse data generalised Rubin's (1976, 1987) theoretical phraseology–an association partially mandated by the result that missing data was simply one form of coarsening. As a consequence, the concepts of missing completely at random" (MCAR), "missing at random" (MAR), and "not missing at random" (NMAR) were distinguished from "coarsened completely at random" (CCAR) and "coarsened at random" (CAR). Heitjan and Basu (1996) explicitly differentiate between these five concepts, but the epistemological extensions provided by coarse data theory are particularly useful to income in public-use micro datasets.

For the purposes of this discussion, coarse data is defined to consist of a combination of continuous data (assumed not to be coarsened at all), bounded data (bracket responses), and item missing data. We formally define what this means for the univariate statistical distribution of income, commencing with the missing data framework and then incorporating the more general coarse data framework.

Following Little and Rubin (2002, 12), we define the complete data matrix as $Y = (y_{ij})$ and the missing data indicator matrix $M = (M_{ij})$. Y is differentiated into an observed and unobserved component, Y_{obs} and Y_{mis}. The distribution $f(\cdot)$ of missingness is conditional upon Y and unknown parameters ϕ, denoted $f(M|Y, \phi)$. If $f(M|Y, \phi) = f(M|\phi)\ \forall\ Y, \phi$, the unobserved data are said to be Missing Completely at Random (MCAR). Here, missing data do not depend on the observed or unobserved components of the complete data matrix. If $f(M|Y, \phi) = f(M|Y_{obs}, \phi)\ \forall\ Y_{mis}, \phi$, the unobserved data are said to be Missing at Random (MAR), a more restrictive condition than MCAR because now the missing data depend on the observed data. If the missing data M depend on the missing values in the data matrix, the mechanism is called not missing at random (NMAR). The missing data mechanism is said to be "ignorable" if the unobserved data are thought

to be MCAR or MAR; in this case, a separate model for the mechanism that causes non-response is not needed (i.e. can be ignored). The missing mechanism is said to be "non-ignorable" if the unobserved data are NMAR.

The coarse data framework incorporates missing data as a type of coarsening, but is also generalisable to bounded data such as income reported in brackets. To see the extensions, we again rely on Little and Rubin's (2002, 127–129) formulation of the problem. Let Y be the complete data matrix in the absence of coarsening with sample space Ψ, and let $f(Y|\phi)$ denote the density of Y for the complete data with unknown parameters ϕ. The observed data are now thought to consist of a subset of the sample space Ψ in which Y is known to fall. This subset is a function of Y and a coarsening variable G that determines the bounds of Y_{obs}, so that $Y_{obs} = Y_{obs}(Y, G)$.

To see the extension to bracketed responses such as those present in income microdata, note that the characterisation of $Y_{obs} = Y_{obs}(Y, G)$ assumes that the observed data fall within *known* upper and lower bounds and not outside these bounds. Since the bounds are assumed known, the coarse data framework is flexible enough to be applied not only to bracketed response types, but also to data that is thought to be imprecisely coarsened, such as rounded data, heaped data, or otherwise partially categorised data (see Heitjan and Rubin, 1991). In each case the coarsening mechanism needs to be precisely modelled.

To incorporate missing data into this framework, call the unobserved data completely coarsened, and allow plausible values of that data to lie within the sample space Ψ of Y. In this case, G is simply the missing data indicator matrix. Thus:

$$y_{obs,ij} = \begin{cases} \{y_{ij}\}, & \text{the set consisting of the single true value, if } G_{ij} = 0 \\ \Psi, & \text{the sample space of Y, if } G_{ij} = 1 \end{cases} \tag{4.1}$$

From this, the data Y_{obs} are called coarsened at random (CAR) if $f(g|y_{obs}, y_{mis}, \phi) = f(g|y_{obs}, \phi)$ for all y_{mis}.

To apply the framework to a mixture of continuous responses, bounded responses and missing data, we follow Heeringa's 1995 example and simply allow G to precisely define whether the data are observed as continuous, bracketed or missing. To make the framework specific to the income question in the OHS and LFS, we will characterise the coarsening process to match what is found in the public-use datasets.

$$y_{obs,ij} = \begin{cases} \{y_{ij}\}, & \text{if } G_{ij} = \{0\} \\ [y_L \leq y_{ij} < y_U), & \text{if } G_{ij} = \{1, 2, \ldots, 14\} \\ \Psi, & \text{if } G_{ij} = \{15, 16, 17\} \end{cases} \tag{4.2}$$

Here, $G_{ij} = \{0\}$ indicates that y_{ij} is observed as a set consisting of the single true (exact) income value; $G_{ij} = \{1, 2, \ldots, 14\}$ indicates that y_{ij} falls within the lower bound y_L and upper bound y_U of one of the fourteen possible brackets in the OHS and LFS income questions; and $G_{ij} = \{15, 16, 17\}$ indicates that y_{ij} is observed as "Don't Know", "Refuse" or "Unspecified", and would then fall within the sample space of Y.

A key implication of the coarse data framework is that the variable G itself is measurement error free (Heitjan and Rubin, 1991; Wittenberg, 2008). This effectively implies that if a respondent reports their income to be within a given bracket, it cannot lie outside of those bounds. It also implies that if a respondent provides an exact income response, that response is assumed to be precisely reported. One of the implications of this relates to the imputation process for it implies that plausible draws of income for the bracketed subset of observations have to lie within the lower and upper bounds of those brackets, while draws for the missing data can be made over the sample space of income.

The Special Case of Unspecified Responses in the Coarse Data Framework

In Statistics SA's household surveys between 1997 and 2003, nonresponse to the employee income question was often recorded in the public-use data as an unspecified response. This response type exists even when there are options for don't know and refuse in the questionnaires. In 1999, the don't know option was introduced to the question for the first time, before both don't know and refuse options were added in 2000. Despite this, in each of the LFS, unspecified responses still exist for the subsample of employed economically active individuals. This represents a form of either processing or measurement error because don't know and refuse exhaust the possible nonresponse types in the income instrument.

Because of this, the nature of the coarsening mechanism for unspecified responses is opaque. Unspecified responses in the OHS 1997 and 1998 are the only identifiable form of nonresponse because the income question does not present any options to the interviewer for recording a don't know or refuse response. Therefore, we are forced to treat those as nonresponse. In 1999, the unspecified responses are confounded with refuse responses. But in the LFS, unspecified responses are identifiable as a form of processing error.

Observations that are deemed to be a result of processing error cannot simply be included in the coarse data framework as applied here, for it represents a mutually exclusive error mechanism in the data. We deal with this below by firstly exploring the extent of processing error in the data and then conducting independent multiple imputations for these observations.

The Special Case of Zero Income Brackets

An idiosyncratic feature of the bounded income question in all of the surveys analysed in this chapter (OHS97-LFS03) is that it has a zero income option in the showcard. The existence of zero income brackets is thought to be related to false income reporting by Vermaak (2010), who imputes a proportion of these responses based on an assessment of the share that seem plausibly zero. The coarse data framework does not allow measurement error in the coarsening process to exist. Therefore, simply

imputing the zero responses without a theoretical basis for doing so is arbitrary. Vermaak (2010) seems to include the self-employed in her subsamples of economically active individuals, which increases the number of zero responses substantially. This is easy to do in the LFS because the same question is asked to both the employed and the self-employed, whereas in the OHS the income question was different for self employed individuals. We restrict the sample here to employees only in all survey years.

Zero income values can exist as a valid response type for the subsample of economically active employees because respondents can be off work on unpaid leave. We evaluate the prevalence of zero income responses below, but keep all such observations in the data without imputing them.

4.2.2 Multiple Imputation

Multiple imputation has gained recognition as one of the most effective methods for handling multivariate item nonresponse in public-use datasets. However, its use requires a clear understanding of its limitations. The coarse data framework is very useful for characterising the possible ways in which observed data may differ from their true values, and while it incorporates missing data as a type of coarsening, its extension to other data problems such as measurement error is limited on theoretical grounds. Recent advances in multiple imputation theory do indeed pose solutions to data measured with error (see, particularly, Ghosh-Dastidar and Schafer, 2003), but associated with this is (1) a necessary change in the operation of imputation algorithms and, (2) a modification of the combination rules required for valid statistical inference from multiply imputed datasets (Reiter and Raghunathan, 2007).

Multiple imputation has to address the pattern of coarsening present in a dataset. It was traditionally envisaged as a tool for data base constructors whose use of the methods was assumed to be independent from the data analyst's (Rubin, 1996). However, as the algorithms became more widely available and as more researchers became familiar with the methods, its use has burgeoned across the social and life sciences to a vast array of different applications. Indiscriminate use of multiple imputation is clearly discouraged by the major proponents of the method. As Schafer (1999) points out, multiple imputation is neither the only principled method for handling missing values, nor is it necessarily the best. Indeed, "(f)rom a statistical standpoint, …a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests…" (Schafer, 1999, 3). This view echoes Rubin's (1996: 475), who reminds all that the "actual objective (of multiple imputation) is valid statistical inference not optimal point prediction under some loss function, and replacing the former with the latter can lead one badly astray".

One of the important implications of the coarse data framework discussed in Sect. 4.2.1, and directly implied by Eq. (4.2), is that the type of coarsening is defined to be precise; in other words, there can be no measurement error in the coarsening

variable (G). The use of the coarse data framework thus places particular restrictions on the manner in which multiple imputation can be conducted. Its utility lies in the fact that it provides clear rules for multiple imputation for the data structure resulting from the income question in the surveys considered.

There are examples in the literature of multiple imputation being used to deal with other forms of survey error. In particular, Ghosh-Dastidar and Schafer (2003) demonstrate how multiple imputation theory can be extended to the case of nonresponse and measurement error (without a validation study). They call their process multiple edit multiple imputation (MEMI), and note that producing MEMI requires assumptions about the distribution of the ideal data, the nature of nonresponse, and a model for the measurement error mechanism. This approach can also be adapted to suit other uses of multiple imputation, such as anonymising confidential survey information (ibid, 2003). However, in each case both the imputation algorithms and the rules for estimation and inference from the multiply imputed datasets differ, and have to be derived for the intended application.

4.3 Setup of the Problem

In this section we firstly discuss the data preparation tasks needed before working with the employee income variables. Here, the existence of bounded zero responses and processing error will be evaluated. We then develop an appropriate multiple imputation algorithm for coarse income data and identify the rules for estimation and inference given the nature of the coarse data problem and the imputation process.

4.3.1 Data Preparation

Zero Income Responses

Since the subsample of interest is economically active employees, zero income responses ought not to exist in general, unless the person is off work temporarily and on unpaid leave. However, in each survey year, there are a positive number of zero responses in the OHS and LFS. Moreover, the majority of zero responses are reported in the bounded income question in the OHS and LFS questionnaires, rather than the exact income question.

Table 4.1 presents the number of observations reported in each response type. Evident from the table is that the number of zero responses is usually very small, ranging from two in 1998 to forty-five in 1997. Most of these are reported in the bounded income question.

Of those employees who reported a zero income response (either in the bounded question or the exact question), the percentage that also reported that they have been absent from work in the past week due to illness ranges from zero in 1997–1999 to

Table 4.1 Distribution of response types: OHS97—LFS03

Response Type		1997	1998	1999	2000	2001	2002	2003
Exact	Obs	16 185	7 637	11 735	18 739	15 945	14 469	13 759
	Percent	67.76	58.81	53.52	87.34	75.25	70.55	68.26
Exact-Zero	Obs	1	.	.	6	3	.	.
	Percent	0.00	.	.	0.03	0.01	.	.
Bounded	Obs	6 713	4 718	8 028	1 997	4 044	4 650	4 964
	Percent	28.10	36.33	36.61	9.31	19.09	22.67	24.63
Bounded-Zero	Obs	45	2	27	36	21	34	34
	Percent	0.19	0.02	0.12	0.17	0.10	0.17	0.17
Don't Know	Obs	.	.	1 588	72	521	651	485
	Percent	.	.	7.24	0.34	2.46	3.17	2.41
Refuse	Obs	.	.	.	144	578	664	891
	Percent	.	.	.	0.67	2.73	3.24	4.42
Unspecified	Obs	942	628	548	461	77	40	23
	Percent	3.94	4.84	2.50	2.15	0.36	0.20	0.11

29% in 2000, 42% in 2001, 53% in 2002 and 24% in 2003. There is no question for whether individuals are on unpaid leave for other reasons, however, so we cannot investigate this phenomenon. Because there are legitimate reasons for zero income reporting, we keep all zero responses in the subsamples of employees for each survey year and do not impute any of them.

Processing Error and/or Measurement Error in the Data

Two anomalies exist in Statistics SA's OHS and LFS: (1) instances where both an actual and a bracketed value are observed for the same individual; and (2) observations that are coded as "Unspecified" (i.e. missing), when in fact response options already exist in the questionnaire for the respondent to reply that they "Don't Know" or "Refuse" to answer the question. It is impossible to tell from the data or the survey documentation whether these anomalies are by design or whether they constitute a form of processing or measurement error, but they need to be addressed before imputation can take place.

To formalise the problem, consider that the universe of potential outcomes for income responses consists of a continuous (exact) income subset, a bounded subset, and a missing (don't know, refuse or unspecified) subset. These three subsets are mutually exclusive because a bracketed outcome is only observed if the respondent chose *not* to answer the actual income prompt from the interviewer. A missing outcome is only observed if the respondent chose not to answer *both* the actual and the bracketed response prompt.

Let the event that an exact income response is reported by the respondent be denoted $P(A)$, the event that a bounded response is reported be denoted $P(B)$, and the event that a missing response be reported be denoted $P(M)$. For these three events to be mutually exclusive, $P(A \cup B \cup M) = P(A) + P(B) + P(M) = 1$, and $P(A \cap B \cap M) = 0$; $P(A \cap B) = 0$; $P(A \cap M) = 0$; $P(B \cap M) = 0$. A first form of (either processing or measurement) error can then be defined to exist if any of these outcomes are violated.

Because the design of the income question evolved between the OHS 1997-LFS 2000, $P(M)$ is not defined by don't know and refuse for every survey year. We therefore need to decompose $P(M)$ into its observable parts: don't know responses (denoted $P(D)$), refusals (denoted $P(R)$), and unspecified responses (denoted $P(U)$). Across the survey years we will then observe missing responses as:

- $P(M) = P(U)$ for OHS 1997 and 1998;
- $P(M) = P(U) + P(D)$ for OHS 1999;
- $P(M) = P(D) + P(R)$ for LFS 2000–2003.

A second form of error can be defined to exist only for the LFS if $P(M) = P(D) + P(R) + P(U)$, where $P(U) \neq 0$. This is because don't know and refuse responses in the LFS complete the possible forms of nonresponse for the employed, economically active population. In the OHS 1999, unspecified responses cannot be identified as a form of error because those responses confound refusals in the same way that unspecified responses confounded both don't know and refusals in the OHS 1997 and 1998.

Table 4.2 presents the extent of these errors in the OHS97-LFS 2003. In order to estimate the subsets correctly, we use the raw data from the surveys of interest before any transformations of the variables are made.
In the table, the column for 2000 is repeated for presentation purposes only, simply to show (1) how the transition from the OHS to the LFS proceeded, and (2) how all of the LFSs compare.

We can see from the table that the sum of the probabilities do not always add up to one; this is the first clue that something is amiss. The first form of error exists for the OHS97-LFS00, but only for the subset $P(A \cap B)$. That is, we sometimes jointly observe values for exact and bounded income for the same respondents in these public-use datasets, which should not be happening.

The findings for 1997 and 1999 are noteworthy because of the magnitude of the error in the data, at 68 and 53%, respectively (obtained from the "Sum" row in the table). For both years, these numbers match the percentage of actual income observations in the survey. This suggests that for each exact income observation, there is also a bounded observation. It is unclear why this is the case, or what motivation Statistics SA could possibly have had in doing this. One potential reason is that it is not a form of error at all, but rather that the survey organisation intentionally did this for some reason (it was not apparent from a reading of the survey organisation's accompanying literature and metadata whether or why this was done).

In order to investigate this further, we checked the consistency between the exact values that were also observed as brackets by transforming actual income into a new

Table 4.2 Subsets of interest in the observed income data

Income response subsets	1997	1998	1999	2000
N (employed EAP)	23 886	12 985	21 926	21 455
(1) Exact Responses: P(A)	0.6779	0.5881	0.5352	0.8737
(2) Bounded Responses: P(B)	1.0000	0.4888	0.8981	0.0951
(3) Nonresponse: P(M)	0.0000	0.0000	0.0724	0.0101
(4) Complement: $(A \cup B \cup M)^{\complement}$	0.0000	0.0484	0.0250	0.0215
Sum: (1) + (2) + (3) + (4)	1.6779	1.1253	1.5307	1.0003
$P(A \cap B)$	0.6779	0.1253	0.5307	0.0003
$P(A \cap M)$	0.0000	0.0000	0.0000	0.0000
$P(B \cap M)$	0.0000	0.0000	0.0000	0.0000
Income Response Subsets	2000	2001	2002	2003
N (employed EAP)	21 455	21 189	20 508	20 156
(5) Exact Responses: P(A)	0.8737	0.7527	0.7055	0.6826
(6) Bounded Responses: P(B)	0.0951	0.1918	0.2284	0.2480
(7) Nonresponse: P(M)	0.0101	0.0519	0.0641	0.0683
(8) Complement: $(A \cup B \cup M)^{\complement}$	0.0215	0.0036	0.0020	0.0011
Sum: (5) + (6) + (7) + (8)	1.0003	1.0000	1.0000	1.0000
$P(A \cap B)$	0.0003	0.0000	0.0000	0.0000
$P(A \cap M)$	0.0000	0.0000	0.0000	0.0000
$P(B \cap M)$	0.0000	0.0000	0.0000	0.0000

monthly income variable, and then converting that variable into a bracketed variable with the same bounds as the SSA's bounded variable. The result was that about 85% in 1997 and 99% of actual income observations in 1999 were in the correct monthly income bracket. For 1998, only 16% of actual income observations were in the correct bracket. While it is true that the extent of this error is mitigated to some extent when there is a match between the variables, the existence of two data points on income for the same person should never, as a rule, exist.

We do not observe this form of error for the other possible subsets, namely $P(A \cap M)$ or $P(B \cap M)$, in any of the datasets. This is unsurprising, for the actual placement of the "Don't Know" and "Refuse" options in the public-use dataset is as an option in the bounded income variable, making it impossible to confuse these subsets (when they enter the data electronically).

It is clear from the table, though, that SSA really improved their performance on this dimension of the problem over time, with this form of error dropping to zero by the LFSs. That said, the LFS2000–2003 all have non-zero complements to $P(A \cup B \cup M)$, which ought to no longer exist given that the income question had specific response options for don't know and refuse. Consequently, a second form of error exists, and is non-zero in each LFS dataset. It is substantial in the OHS 1999 and LFS 2000, at approximately 2.5 and 2%, respectively, of the sample of employed economically active individuals.

The first type of error discussed for these datasets can easily be dealt with by generating a new derived income variable from the combined actual and interval variables in the raw data, and overwriting the bracketed responses with the exact responses. The rationale for doing this is that exact responses are preferred to bounded responses from an information content point of view (see Schwartz and Paulin, 2000). For the second type of error, we deal with it differently across the survey years: the observations are kept in the OHS 1999 because they are confounded with refusals; but they are omitted for imputation purposes from the LFS, where the nonrespondent subset is fully defined by don't know and refusals. However, we will evaluate and impute these response types separately in the analysis below to examine their distribution.

4.3.2 The Imputation Algorithm

There are several important steps required for the development of appropriate multiple imputation methods. These include:

- Correctly characterising the nature of the missing data, called the "missingness" mechanism. Little and Rubin (2002, 4–8) identify several such patterns, including univariate nonresponse, multivariate nonresponse (e.g. item nonresponse and unit nonresponse), monotone missing (e.g. attrition in longitudinal studies), general patterns of missing data (e.g. item nonresponse on many variables in a single dataset), file matching missing data problems, and latent-variable patterns with variables that are never observed. An important relationship exists between the pattern of missing data and the imputation procedure, with univariate and monotone missing data patterns allowing for the simplest imputation algorithms to be implemented (White et al., 2007).
- Based on the missing mechanism, choosing an appropriate multiple imputation algorithm. An important requirement of this choice is ensuring that the imputation method is "proper", which means that it must account for uncertainty in the parameters of the imputation model (White et al., 2011). This is necessary because Rubin's Rules for combining datasets only yield valid standard errors if the imputations adequately reflect the uncertainty in drawing values for the missing data.
- Specifying the imputation model: variable selection. As White et al. (2011) point out, covariates for each prediction equation in the imputation algorithm have to be carefully chosen to help increase the plausibility of the missing (coarsened) at random assumption. Van Buuren et al. (1999) suggest that variable selection ought to include:
  - Variables that are required in the complete data model of interest;
  - Variables that appear to determine missingness;
  - Variables that explain a considerable amount of the variance of the target variable, which helps to reduce the uncertainty of the imputations.

- Specifying the imputation model: model form. An important concept in the imputation literature is the idea of a "congenial" imputation model. White et al. (2011) state that instead of aiming to find the true imputation model, an alternative approach relies on finding an imputation model that is congenial to the analysis model but not necessarily correctly specified. In this way, inference on multiply imputed data can approximate maximum likelihood estimates (for large numbers of imputations) (ibid, 385).
- Choosing sufficiently large numbers of multiple imputations for the missing data in order to reflect the uncertainty present in the imputation process. Traditional multiple imputation theory used the oft-cited rule-of-thumb of five imputations, but more recent studies suggest that many more multiple imputations may be needed—in the order of one hundred for certain applications (Graham et al., 2007).
- Conducting complete-case analysis from multiply imputed data using the correct combination rules. Depending on the problem under investigation, these combination rules may differ to Rubin's Rules (Reiter and Raghunathan, 2007).
- Testing the sensitivity of the results. This can be done in different ways, since each step described above imposes a certain structure on the imputation process, the sensitivity of which can be investigated. Carpenter et al. (2007) use a weighting approach after imputation to test the validity of the MAR assumption for each imputed dataset. However, this requires a specific model for how imputations depart from MAR. Sensitivity analysis can also be conducted using an uncongenial imputation model, which Kenward and Carpenter (2007) suggest. This involves specifying an imputation model that differs from the analysis model. We incorporate this suggestion into the analysis below.

It is important to note that in this chapter we are concerned with multiply imputing for coarse income data only, which sets the pattern of coarseness as univariate. Consequently, we are not interested in multivariate coarsening or the effect of coarse data on the earnings covariate vector. An important consequence of this is that the multiple imputation algorithms simplify tremendously because the process of drawing plausible values from the conditional distribution of each variable with coarse data is restricted by design to one conditional distribution—income.

Practically, this means our task is to develop a univariate multiple imputation algorithm. This has two implications: (1) it is no longer necessary to characterise the coarse data mechanism in a multivariate sense (e.g. to establish whether it is monotonic or a general multivariate coarse data pattern); and (2) it is no longer necessary to use a sequential regression multiple imputation approach to the problem because there is only one variable with coarse data.[1] For this purpose we utilise

[1] The two most common sequential imputation algorithms are variants of Van Buuren et al. (1999) multiple imputation by chained equations (MICE) algorithm, and Raghunathan et al. (2001) sequential regression multiple imputation (SRMI) algorithm. Royston's (2004, 2005, 2007, 2009) imputation by chained equations (ICE) algorithm is similar in principle to Van Buuren et al. (1999) procedure, while StataCorp (2011) developed a flexible multiple imputation package that can perform monotonic multiple imputation, fully conditional specification procedures (such as MICE, ICE and SRMI), and explicit Bayesian algorithms that allow the user to specify prior and posterior

the interval regression-based multiple imputation procedure developed by Royston (2007) and modified by StataCorp (2011).

4.3.3 Estimation and Inference from Multiply Imputed Data

Multiple imputation was suggested as a potential solution to missing data problems by Rubin (1976), and the rules for inference from multiply imputed datasets came to be known as Rubin's Rules. These essentially state that analyses of multiply imputed datasets should be conducted based on standard complete-data techniques, but parameter estimates must be combined across datasets.

Formally, Rubin's Rules are presented as follows (we follow Royston's, 2004 exposition): Let $\hat{\theta}_m, W_m, m = 1, \ldots, M$ be M complete-data estimates and their associated variances for an estimated parameter θ. The mean of θ is then calculated as:

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m. \tag{4.3}$$

The variance of θ has both a within component and a between component. The within component of the variance is:

$$\bar{W}_M = \frac{1}{M} \sum_{m=1}^{M} W_m. \tag{4.4}$$

The between component of variance is:

$$B_M = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta}_M)^2. \tag{4.5}$$

Combining the within and between-components then leads to the formula for total variance:

$$T_M = \bar{W}_M + \frac{M+1}{M} B_M, \tag{4.6}$$

The reference distribution for confidence intervals and significance tests is a t distribution,

$$(\theta - \bar{\theta}_M) T_M^{-1/2} \sim t_v,$$

distributions, amongst others. The algorithm in StataCorp (2011) also has the functionality to be restricted to the type of univariate multiple imputation procedure utilised in this chapter.

with degrees of freedom,

$$\nu = (M-1)\left(1+\frac{1}{M+1}\frac{\bar{W}_M}{B_M}\right)^2.$$

In the analysis below, we obtain parameter estimates for the marginal distribution of post-multiply imputed income using these rules for a variety of different parameters.

4.4 Results: Univariate Multiple Imputations for Coarse Income

In this section we conduct univariate multiple imputation for coarse income data. Our objective is to draw plausible values for both the bracketed and missing subsets in each survey year. The multiple imputation algorithm employed for this purpose is based on an interval regression procedure developed by Statacorp in *Stata Release 12* (2011). The algorithm allows for imputed draws to be restricted to the income bracket lower and upper bounds, and it simultaneously allows for imputed draws for missing data to be unrestricted. The sensitivity of estimates and inferences to a range of different specifications of the prediction equations of the imputation algorithm is tested. Four models are developed for this purpose:

1. Model 1: multiply imputing five times with an intentionally mis-specified covariate vector that includes gender and language as the only predictors. The purpose of doing this is to create a baseline set of imputations that provide insight into how badly things can go wrong due to covariate mis-specification.
2. Model 2: multiply imputing five times with prediction equations using covariates that explain the response process only (these are the same as the response propensity models of Chapter Three). The purpose of doing this is to create an "uncongenial" set of imputations, in the sense that the imputation model differs from the intended analysis model (Kenward and Carpenter, 2007).
3. Model 3: Multiply imputing five times for univariate income with Mincerian earnings function covariates only. These include age and experience (including their squares), other personal characteristics variables (including race and gender, but not language), hours worked, occupation, trade union membership, industry, and province. The purpose of this model is to create a set of imputations that would be "congenial" to analysing earnings, even though variables that explain the response process are largely absent.
4. Model 4: multiply imputing five times using both Mincerian earnings equation covariates and response propensity covariates. On a-priori grounds, this algorithm is treated as first-best because it conforms to the recommendations of Van Buuren et al. (1999, see Sect. 4.3.2 for discussion).

4.4.1 *Quantiles and Moments Across Four Imputation Models*

The results for weighted univariate income parameter estimates for each imputation model are presented in Table 4.3. The table shows parameter estimates of the multiply imputed nominal employment income variables ("Yimp"), for each of the four imputation models discussed above and the estimation sample size ("Est.N") in each survey year. Quantile estimates are calculated post-imputation for each of m imputed income variables using Rubin's Rules (see Eq. (4.3) above). For this section, the variance of the estimates are omitted, but they will be evaluated in detail below in Sect. 4.4.6.[2]

Results from the table are discussed thematically. The following issues are of relevance:

- The difference in parameter estimates across imputation methods.
- The difference in the estimation sample size across imputation methods.
- The difference in the upper and lower tails of each distribution.

Evident from Table 4.3 is that up until the median, the differences between the imputations are relatively trivial. This is expected, for we know that the probability of a bounded response increases as income increases, so any difference in imputed draws for this subset will only make its presence felt higher up the income distribution. That said, an important feature of the imputation algorithm is that it limits the range of imputed draws to the bounds of each income category. For the highest income category, however, this is an open ended interval with no upper bound. Therefore, imputations for respondents in this group have no upper limit.

At the top of the income distribution, we see substantial differences between the distributions. At the 99th percentile, the OHS 1999 has the widest range between the four imputation models. The mis-specified method of model 1 leads to substantially higher estimates than any other model. The differences between distributions in model 2 (that has response propensity covariates) and model 3 (that has earnings function covariates) is also substantial, but the difference in estimates between model 3 and the first-best imputation model 4 (which combines response propensity and earnings function covariates) is much lower.

In fact, in every survey year and for every quantile other than the minimum, the first-best imputation model always generates distributions with the lowest estimates. The importance of this is particularly stark for the maximum values in each distribution. Important to note here is that in survey years where an exact income value is extreme, such as in 1999 and 2000, the imputed values rarely exceed this outlier, except for the mis-specified imputation model one in 1999, where an imputed draw

[2] Note that the variance of a quantile has to be computed manually after m multiple imputations using Rubin's Rules (see Eq. (4.4) to (4.6) above). The total variance of a quantile contains only a between-imputation component of variance (see Eq. (4.5) above), but Rubin's total variance formula in Eq. (4.6) still has to be used to calculate the variance of a quantile because of the $(m + 1)/m$ adjustment for finite m.

Table 4.3 Quantiles of four different models for imputed income

Year	Variable	min	p5	p10	p25	p50	mean	p75	p90	p95	p99	max	Est.N
1997	Yimp-model1	0	211	350	863	1 796	4 054	4 000	8 705	14 512	37 724	307 832	23 868
	Yimp-model2	0	204	350	804	1 700	3 688	3 665	7 871	12 918	33 526	202 582	23 303
	Yimp-model3	0	206	350	803	1 709	3 433	3 660	7 548	12 028	27 451	177 681	23 206
	Yimp-model4	0	201	348	800	1 656	3 287	3 516	7 278	11 457	26 572	127 069	22 805
1998	Yimp-model1	0	206	304	800	1 951	5 600	4 980	12 213	21 397	61 051	511 400	12 985
	Yimp-model2	0	201	300	772	1 809	5 210	4 673	11 532	19 980	54 850	598 968	12 574
	Yimp-model3	0	200	300	681	1 608	3 910	3 971	8 836	14 488	35 832	370 000	11 619
	Yimp-model4	0	200	300	652	1 586	3 756	3 803	8 270	13 741	33 601	370 000	11 356
1999	Yimp-model1	0	216	337	785	2 000	7 549	5 869	15 441	28 147	88 298	1 559 224	21 915
	Yimp-model2	0	213	311	700	1 757	6 376	4 970	13 008	23 760	72 413	1 522 138	20 365
	Yimp-model3	0	216	312	700	1 796	6 041	5 014	12 879	22 483	61 965	1 522 138	20 575
	Yimp-model4	0	200	300	678	1 702	5 697	4 738	12 137	21 297	56 636	1 522 138	19 562
2000	Yimp-model1	0	217	318	665	1 521	5 890	3 500	7 037	11 146	27 474	4 726 242	20 993
	Yimp-model2	0	217	304	652	1 500	5 824	3 486	6 965	10 934	25 572	4 726 242	20 734
	Yimp-model3	0	217	305	652	1 500	5 804	3 500	7 000	10 921	24 686	4 726 242	20 725
	Yimp-model4	0	216	300	652	1 500	5 678	3 358	6 611	10 157	23 446	4 726 242	20 538
2001	Yimp-model1	0	250	350	748	1 800	4 120	4 383	8 999	14 894	37 827	500 000	21 112
	Yimp-model2	0	248	350	700	1 738	3 751	4 000	8 161	13 277	32 644	500 000	20 486
	Yimp-model3	0	250	350	702	1 738	3 681	4 000	8 098	12 934	30 953	500 000	20 599
	Yimp-model4	0	242	350	700	1 700	3 471	4 000	7 855	11 972	28 095	500 000	20 156
2002	Yimp-model1	0	250	350	763	1 919	5 399	5 012	11 827	20 190	55 296	500 797	20 467
	Yimp-model2	0	250	350	737	1 800	4 896	4 957	11 010	19 021	45 871	396 532	19 834
	Yimp-model3	0	250	350	750	1 842	4 448	4 844	10 159	16 738	38 143	380 000	19 994
	Yimp-model4	0	250	350	701	1 800	4 122	4 580	9 618	15 558	34 388	380 000	19 549
2003	Yimp-model1	0	300	480	856	2 000	5 925	5 653	13 120	22 415	59 026	726 726	20 130
	Yimp-model2	0	300	477	846	2 000	5 300	5 145	12 161	20 446	51 422	321 882	19 599
	Yimp-model3	0	300	495	854	2 000	5 048	5 226	11 904	19 330	45 200	240 975	19 805
	Yimp-model4	0	300	472	818	2 000	4 697	5 000	11 027	17 980	40 299	212 935	19 359

is larger than the maximum in that year. But there is nothing generalisable from this observation, for in 2001 where an exact income value also represents the maximum, the imputation model one does not exceed it. The relationship between outliers in the observed distribution and multiple imputation is therefore important to be aware of.

The differences between the four imputation models at the maximum are substantial in 1997, 1998, 2002, and 2003. This suggests that specification of the imputation algorithm is most significant to the upper tail of the income distribution. The fact that the model 4 estimates are the lowest for each parameter across the entire distribution suggests that covariate selection based on explaining both the outcome variable of interest (income) and the response process leading to coarse data (response propensities), is crucial for plausible draws of income, but even more important for the highest income earners.

However, it is not clear that a congenial imputation model that only focuses on earnings covariates (model three) is substantially worse than model four. Model two is slightly more volatile across the survey years, suggesting that choosing covariates that explain the response process alone is not an optimal way of specifying multiple imputation algorithms. Finally, the reduction in the estimation sample size for model 4, although relatively modest, is nevertheless an important limitation associated with increasing the number of covariates in the prediction equations.

4.4.2 The Distribution of Multiply Imputed Bounded Income Values

In this section we compare the subsets of multiply imputed income. We restrict the analysis initially to the first-best imputation model only. The kernel densities of the five multiply imputed bounded income distributions are presented in Fig. 4.1. The density for exact income responses is on the same graph. The solid lines represent the bounded distributions and the dashed line the continuous distribution for exact responses.

We can see from Fig. 4.1 that the densities of imputed draws for the bracketed subset are always to the right of the actual income response distribution. This is entirely expected from the analysis in Chapter Three, where we saw that the probability of a bounded income response increases as income increases.

The densities for each of the five imputed draws are very similar, and generally have similar skewness and kurtosis. This is to be expected given the bounds of the brackets, which restrict where in the distribution the draws can be made. An important observation concerns the maxima of the imputed draws for the bracketed subset of income respondents. In 1997 and 2003 we see clearly that the maximum monthly income value in the data is generated by the imputed draws for bounded income.

It is also apparent that the minimum income values are determined by respondents who answer the bracketed section of each questionnaire. It should be remembered

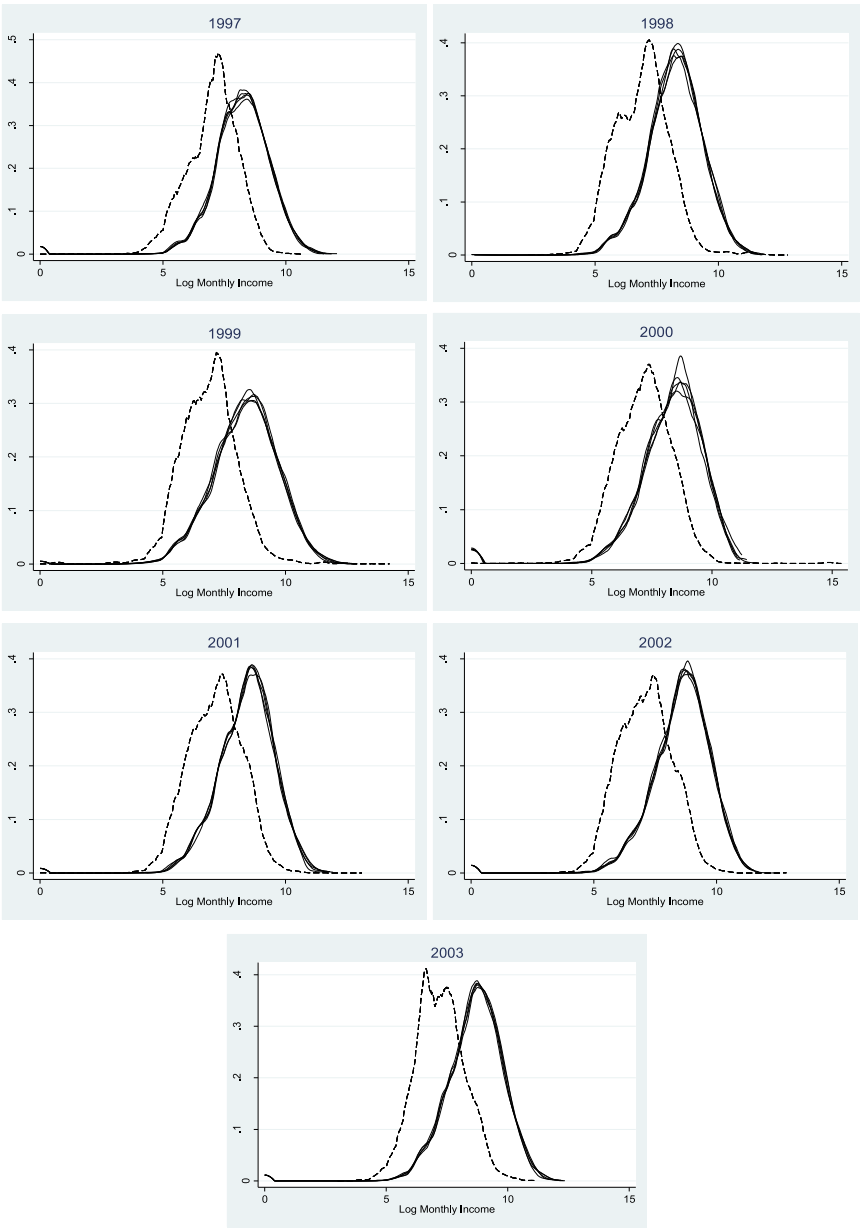

Fig. 4.1 Multiply imputed bracketed income (solid line) compared to observed continuous income (dashed line): 1997–2003

that the lowest bracket in each questionnaire is zero. And in each survey year we observe a non-zero count of such responses. This is highest in 2000, but is also noticeable in 1997, 2001–2003, where it clearly affects the kernel densities. The existence of zero values for employee income is not unreasonable given the fact that the income question asks respondents about their labour market activities in the week preceding the interview, during which respondents could be earning no income.

4.4.3 The Distribution of Multiply Imputed Missing Income Values

The kernel densities of multiply imputed draws for the nonresponse subset (combining unspecifieds, don't know and refusals as appropriate to the survey year) of observations are compared to the observed responses (bounded and continuous) in Fig. 4.2. As before, each of the five multiply imputed income distributions are plotted on the same graph for each year. The densities for imputed draws of missing income observations are the solid lines while observed income has dashed lines.

We can see from this figure that the distribution of imputed missing values changes over time, relative to the distribution of observed responses. In 1997 the densities for the missing income respondents generally overlaps that of the observed respondents. This suggests that respondents who didn't answer the income question had similar predicted values of income compared to respondents who did provide an answer to the question, based on observables in the public-use dataset. That begins to change immediately after 1997, however, where in 1998 it becomes clear that the missing subset of respondents had predicted income values discernibly more to the right than the observed subsets of income respondents.

The location of the densities for the missing subset of observations gradually moves further to the right over time. To explain this trend, it is noteworthy to remember that we are observing the *nominal* distribution of monthly income over time. It is therefore reasonable to expect that the distribution of income in the population itself would shift to the right over the time frame.

4.4.4 The Distribution of Multiply Imputed Refusals and Don't Know Income Values

In this section we evaluate the distributions of multiply imputed refusals and don't know income values. The time frame is restricted to 2000 and beyond, since these response options only appear in the questionnaires from 2000 onwards.

The kernel densities for the multiply imputed draws of refusals are plotted with a solid line while draws for don't know responses are plotted with dashed lines. Because imputed draws for refusals and don't know responses are of particular interest, we

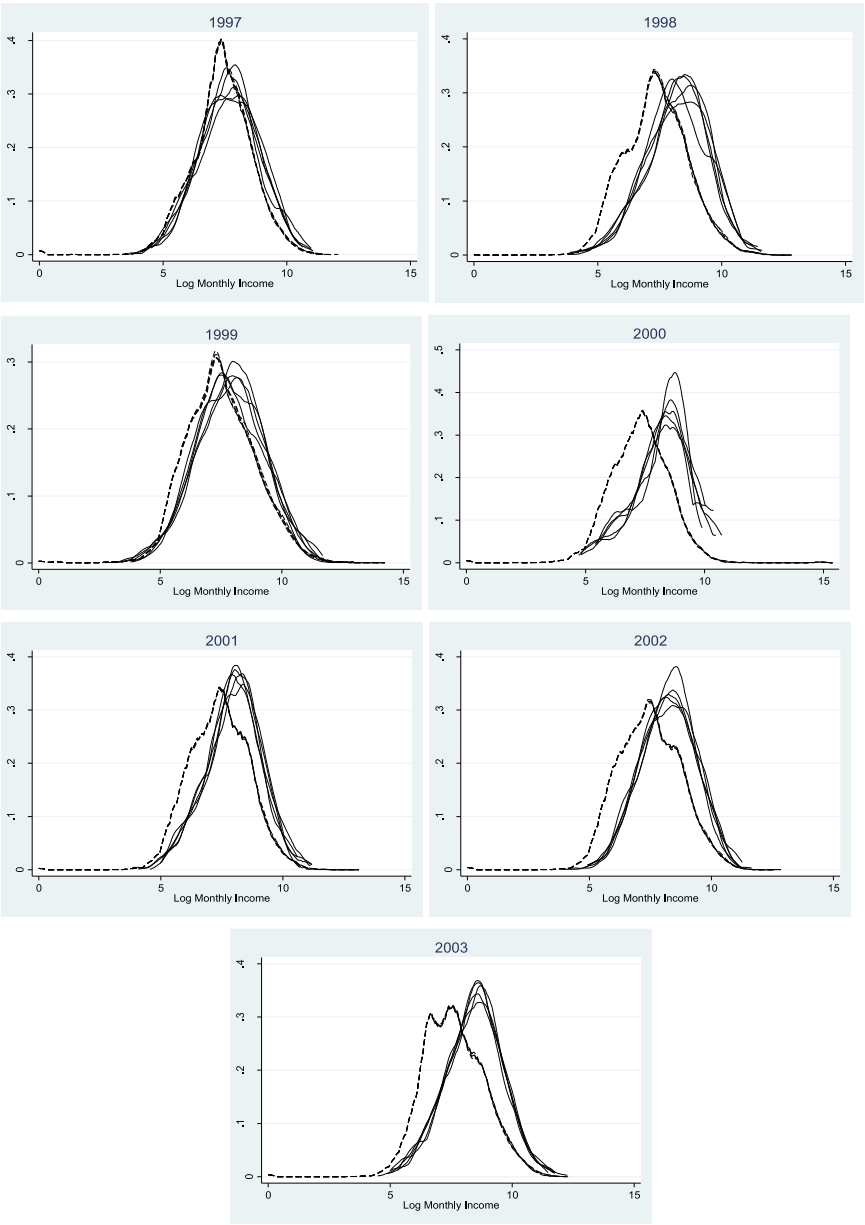

Fig. 4.2 Multiply imputed missing income (solid line) compared to observed (multiply imputed bracket and continuous—dashed line) income: 1997–2003

compare the four multiple imputation models against each other. In Fig. 4.3, the misspecified imputation method (model 1) is on the left hand side while the first-best imputation method (model 4) is on the right hand side.

It is evident from Fig. 4.3 that there is now a lot more variation between the imputed draws for each response group, and there are very different inferences about the distribution of don't know and refuse responses depending on which multiple imputation method is used. According to model one, the two groups are nearly indistinguishable, whereas in model four they are always very different. The densities of imputed income draws for refusals always lie to the right of the don't know responses. This shows a clear advantage of correctly specifying multiple imputation algorithms.

To evaluate the sensitivity of this finding, we now compare the results for multiple imputation models 2 and 3 against each other. Figure 4.4 presents the densities where refuse responses are the solid lines while don't know responses are the dashed lines.

We can see from Fig. 4.4 that regardless of whether the multiple imputation algorithm is specified with response propensity covariates only, or whether it is specified with earnings function covariates, the imputed draws for don't know and refuse subsets of the income distribution show very different distributions. The fact that both models predict this difference is unsurprising because some of the response propensity covariates were chosen precisely because they're correlated with income.

Consequently, despite the fact that the response process for the income question was explained in the previous chapter, where it was evident that refusals were not discernibly different to don't know responses on observable covariates, when we impute for refusals and don't knows there *are* discernible differences between these subsets of the income distribution. The former finding reinforces the fact that this was likely due to weak power associated with small sample sizes for the third stage response propensity models. However, when refusals and don't know responses are set to missing and imputed off observed incomes, discernible differences do exist between these groups.

4.4.5 Unspecified Responses as a Source of Error

In this section we isolate two survey years where unspecified responses represent a significant source of error, namely 1999 and 2000. Unspecified responses in 1999 are confounded with refusals; they consequently enter into the multiple imputations models discussed above. However, in 2000 unspecified responses represent a source of error only because don't know and refuse responses complete the nonresponse possibilities. Therefore, these responses are not imputed in models 1 through 4 above. However, in this section we conduct a new multiple imputation exercise for the LFS 2000 that is identical to model 4 above, but that does multiply impute values for unspecified responses. We then evaluate the densities of these unspecified responses compared to the other nonresponse subsets (Fig. 4.5).

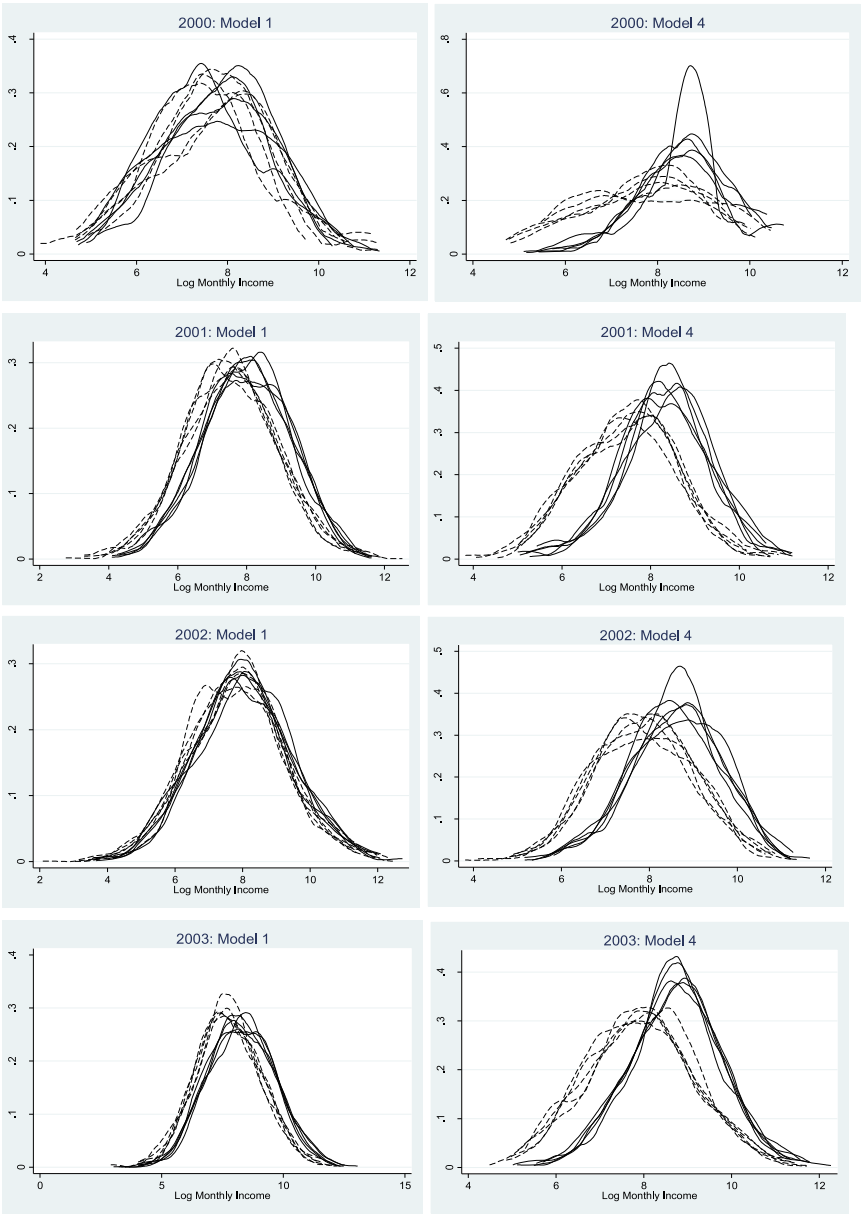

Fig. 4.3 Multiply imputed missing income: refusals (solid line) compared to don't know (dashed line): 2000–2003

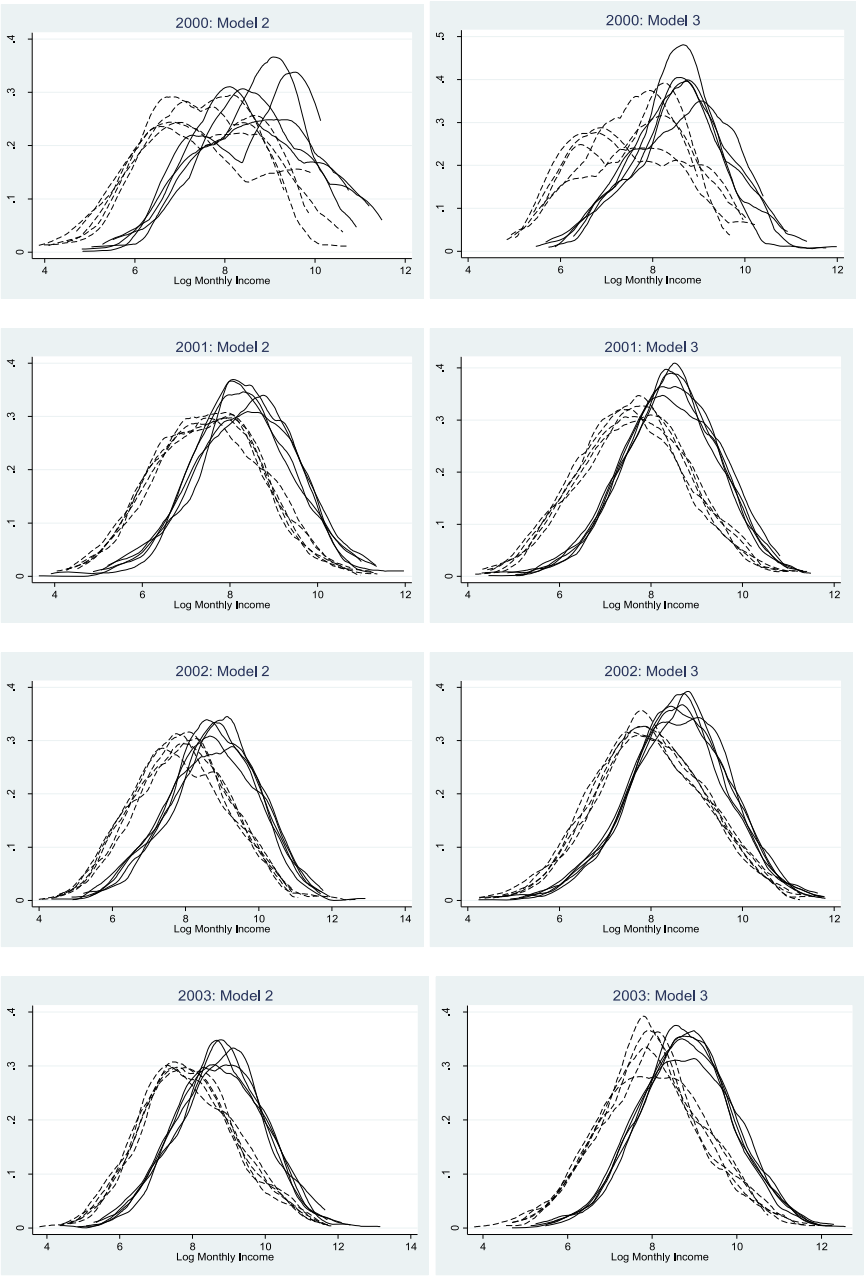

Fig. 4.4 Refusals (solid line) compared to don't know (dashed line): response propensity (model 2) and earnings function (model 3) imputations: 2000–2003

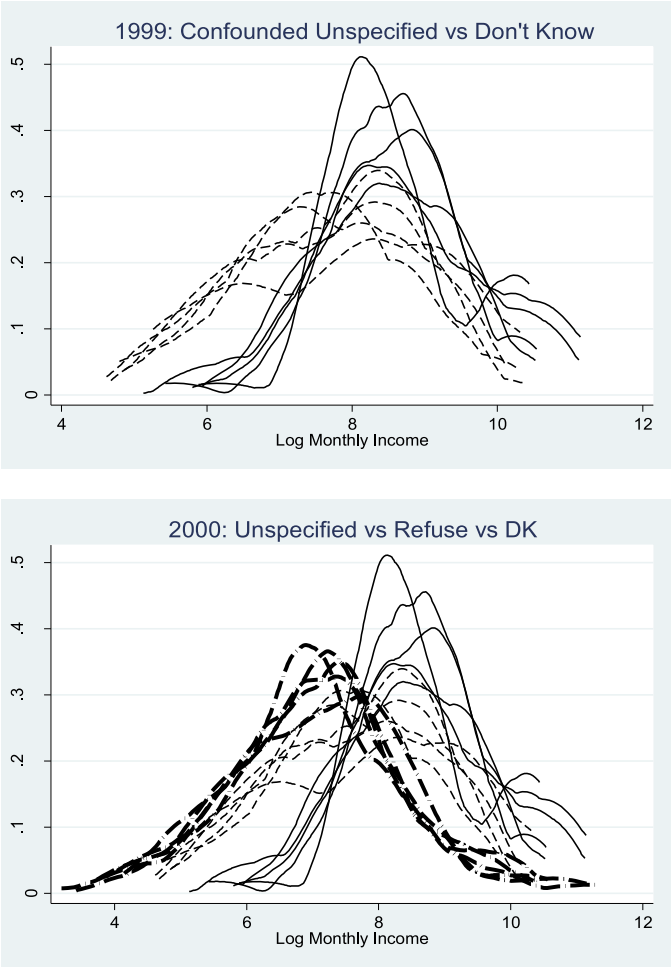

Fig. 4.5 Unspecified response error imputations: 1999 and 2000

Table 4.1 on page 126 presents the subsample sizes for unspecified responses. We now want to compare the multiply imputed draws for these responses against the imputed draws for don't know responses in 1999, and against both don't knows and refusals in 2000. Figure 4.5 presents the results. In 1999, the densities for unspecified income draws are the dashed lines, while the solid lines represent don't know responses. In 2000, the densities for unspecified income draws are the bold dashed lines, whereas refusals are the solid lines and don't know the narrower dashed lines.

From the figure it is clear that unspecified responses are substantially different to identified nonresponse groups in both 1999 and 2000. In 1999, if the unspecified responses were only refusals, then we would expect the distribution of these responses

to lie to the right of the imputed don't know densities, as they do for every survey year in Figs. 4.3 and 4.4. However, they are much more widely spread across the income distribution than refusals.

The same is true in 2000, when there is no longer confounding with refusals. Here, the densities for the imputed unspecified responses are spread across a much larger range than either the don't know or refuse imputations. This suggests that processing error is a completely different error mechanism to nonresponse on the income question, and should consequently not enter multiple imputation algorithms that do not explicitly account for the very different properties of this component of error.

4.4.6 Stability of Parameter Estimates as the Number of Multiple Imputations Increase

The final section of this paper evaluates the stability of parameter estimates of imputed income as the number of imputations increase from two to five to twenty. We conduct multiple imputations using the specification of model 4 only. A-priori, we know that there is not much variation in imputed draws below the median of monthly income from previous analysis (see Table 4.3 on page 138). However, above this level there is more scope for variation. In particular, the largest (open-ended) income bracket as well as the distribution for imputed refusals and don't know responses should be considered to be highly variable given the analysis above. We therefore need to establish the bounds of sensitivity due to the number of multiple imputations conducted. Tables 4.4 and 4.5 present the results of this exercise.

Parameter estimates in Table 4.4 are calculated as the mean of the two, five and twenty multiply imputed monthly income variables in the each respective datasets, as per Eq. 4.3 of Rubin's Rules. Evident from the table is that quantile estimates are almost identical below the median. For the mean of monthly income, the estimates are also very close across the two, five and twenty imputations for each survey year. In fact, this observation holds for every quantile including the maximum in every survey year. Even when we sum up all of the observations for monthly income to create a population-based estimate of the total monthly income earned by employees in South Africa, we can see that estimates do not differ substantially.

The coefficient of variation of these estimates is presented in Table 4.5. Given that the means of parameter estimates are stable over two, five and twenty imputations–as presented in Table 4.4–the coefficient of variation is informative about the magnitude of the inflation in the variance observed as the number of imputations increase.

We can see from the table that the ratio of the standard deviation to the mean is very small across every quantile and moment as the number of imputations increase. The largest values for the coefficient of variation are all found in the maximum column,

Table 4.4 Quantile estimates of imputed income as number of imputations increase

Yr & # Imps	p10	p25	p50	mean	p75	p90	p95	p99	max	wgt.sum	Est.N
97 m = 2	348	800	1 652	3 291	3 550	7 285	11 487	26 409	124 035	25 097 196 736	22 805
97 m = 5	348	800	1 656	3 287	3 516	7 278	11 457	26 572	127 069	25 067 172 581	22 805
97 m = 20	348	800	1 661	3 310	3 523	7 271	11 493	26 856	157 552	25 241 903 829	22 805
98 m = 2	300	652	1 586	3 766	3 834	8 394	13 808	32 589	370 000	22 605 113 545	11 356
98 m = 5	300	652	1 586	3 756	3 803	8 270	13 741	33 601	370 000	22 547 061 243	11 356
98 m = 20	300	652	1 592	3 809	3 826	8 435	13 990	34 417	370 000	22 864 444 500	11 356
99 m = 2	300	674	1 704	5 651	4 712	11 998	20 982	57 871	1 522 138	43 505 765 737	19 562
99 m = 5	300	678	1 702	5 697	4 738	12 137	21 297	56 636	1 522 138	43 867 855 872	19 562
99 m = 20	300	674	1 702	5 650	4 703	12 084	21 026	55 297	1 522 138	43 499 371 526	19 562
00 m = 2	300	652	1 500	5 683	3 350	6 654	10 076	22 779	4 726 242	50 081 776 261	20 538
00 m = 5	300	652	1 500	5 678	3 358	6 611	10 157	23 446	4 726 242	50 044 395 951	20 538
00 m = 20	300	652	1 500	5 686	3 349	6 635	10 103	23 158	4 726 242	50 112 869 667	20 538
01 m = 2	350	700	1 700	3 481	4 000	7 936	12 086	27 635	500 000	28 759 747 602	20 156
01 m = 5	350	700	1 700	3 471	4 000	7 855	11 972	28 095	500 000	28 683 413 421	20 156
01 m = 20	350	700	1 704	3 489	4 000	7 951	12 019	28 640	500 000	28 826 536 243	20 156
02 m = 2	350	700	1 800	4 161	4 591	9 837	15 897	34 629	380 000	35 123 620 901	19 549
02 m = 5	350	701	1 800	4 122	4 580	9 618	15 558	34 388	380 000	34 800 753 362	19 549
02 m = 20	350	704	1 800	4 153	4 582	9 662	15 494	34 306	380 000	35 060 187 137	19 549
03 m = 2	471	828	2 000	4 685	5 000	11 119	18 175	39 574	145 035	42 606 474 187	19 359
03 m = 5	472	818	2 000	4 697	5 000	11 027	17 980	40 299	212 935	42 717 106 246	19 359
03 m = 20	470	813	2 000	4 732	5 001	11 215	18 300	40 466	225 885	43 033 850 802	19 359

Table 4.5 Coefficient of variation of quantiles and moments as number of imputations increase

Yr & # Imputations	p10	p25	p50	Mean	p75	p90	p95	p99	Max	Sum	N
97 m = 2	0.0000	0.0000	0.0107	0.0054	0.0062	0.0029	0.0047	0.0100	0.0344	0.0052	22805
97 m = 5	0.0026	0.0000	0.0063	0.0137	0.0045	0.0159	0.0229	0.0516	0.2556	0.0137	22805
97 m = 20	0.0013	0.0000	0.0077	0.0116	0.0076	0.0129	0.0171	0.0382	0.3134	0.0116	22805
98 m = 2	0.0000	0.0000	0.0125	0.0272	0.0245	0.0393	0.0553	0.0000	0.0000	0.0272	11356
98 m = 5	0.0000	0.0000	0.0090	0.0104	0.0101	0.0160	0.0295	0.0280	0.0000	0.0104	11356
98 m = 20	0.0000	0.0000	0.0072	0.0228	0.0136	0.0275	0.0401	0.0615	0.0000	0.0228	11356
99 m = 2	0.0000	0.0000	0.0033	0.0229	0.0036	0.0071	0.0204	0.0585	0.0000	0.0229	19562
99 m = 5	0.0000	0.0073	0.0136	0.0229	0.0096	0.0180	0.0263	0.0744	0.0000	0.0229	19562
99 m = 20	0.0000	0.0145	0.0044	0.0179	0.0131	0.0157	0.0242	0.0469	0.0000	0.0179	19562
00 m = 2	0.0000	0.0000	0.0000	0.0061	0.0211	0.0114	0.0107	0.0484	0.0000	0.0062	20538
00 m = 5	0.0000	0.0000	0.0000	0.0068	0.0082	0.0075	0.0205	0.0480	0.0000	0.0068	20538
00 m = 20	0.0000	0.0000	0.0000	0.0059	0.0138	0.0099	0.0185	0.0357	0.0000	0.0059	20538
01 m = 2	0.0000	0.0000	0.0000	0.0087	0.0000	0.0101	0.0100	0.0547	0.0000	0.0087	20156
01 m = 5	0.0000	0.0000	0.0000	0.0123	0.0000	0.0138	0.0041	0.0558	0.0000	0.0122	20156
01 m = 20	0.0000	0.0000	0.0044	0.0091	0.0000	0.0087	0.0099	0.0344	0.0000	0.0091	20156
02 m = 2	0.0000	0.0000	0.0000	0.0046	0.0029	0.0094	0.0179	0.0152	0.0000	0.0046	19549
02 m = 5	0.0000	0.0025	0.0000	0.0095	0.0090	0.0140	0.0114	0.0325	0.0000	0.0095	19549
02 m = 20	0.0000	0.0070	0.0000	0.0107	0.0121	0.0127	0.0154	0.0293	0.0000	0.0107	19549
03 m = 2	0.0286	0.0214	0.0000	0.0006	0.0000	0.0067	0.0055	0.0081	0.0360	0.0006	19359
03 m = 5	0.0185	0.0037	0.0000	0.0125	0.0000	0.0103	0.0016	0.0540	0.0652	0.0125	19359
03 m = 20	0.0162	0.0099	0.0000	0.0113	0.0005	0.0165	0.0167	0.0376	0.3968	0.0113	19359

for the survey years 1997 and 2003. Even here though, the numbers are less than 0.5. Aside from these larger values, every other estimate of the coefficient of variation is always below 0.1.

Despite the small magnitude of these coefficients, an important observation is the fact that they do not simply reduce in size as the number of imputations increase. This prevents any strong conclusions about the relationship between the number of imputations and its impact on inference. Two contributing factors to this finding are that (1) the percentage of missing observations is small (at between 3 and 5% for each survey year), and (2) the range of the bounded subset of observations is restricted through the imputation algorithm to lie within the lower and upper bound of each income bracket, thereby formulaically reducing the variance for imputed draws for all but the highest, open-ended income bracket.

For the highest, open-ended income bracket, we saw that specification of the prediction equation in the imputation algorithm is important for reducing the right skewness of the upper tail. Since parameter estimates in Tables 4.4 and 4.5 use both response propensity and earnings function covariates in the model, the variance even in the upper tail of the distribution is relatively low.

The overall conclusion from this analysis is that stability in the point estimates of parameters of multiply imputed income is achieved with as little as two multiple imputations.

4.5 Conclusion

This chapter conducted univariate multiple imputation for coarse subsets of the employee income distribution in South African household surveys from 1997 to 2003. During this time, the employee income question itself evolved, shedding greater light on the coarse response mechanism. The coarse data framework was very useful in guiding the approach not only to the imputation algorithm, where an important implication was restricting the range of the imputed draws to lie within each income bracket, but also to the treatment of unspecified responses when they were identified as a source of survey error. This is one of the major advantages of the coarse data framework: it encourages an explicit approach to the characterisation of the response mechanism, which then leads to clear rules about what can and cannot be accommodated in the imputation step.

For processing error, the fact that two variables are released in the public-use dataset—one for actual income responses and one for bracketed responses—implies that there is a non-zero chance of error between these variables that needs to be addressed when it exists. We identified two types of survey errors: one where duplicate income responses were identified for the same individual, and another where unspecified responses were present in the data even when response options that complete the missing data subset were present in the questionnaire (i.e. don't know and refusals). The solution to the first type of error was to create a new variable for income that overwrites the duplicate records of bounded income with the actual income values. However, for the second type of error, there was no simple solution because the problem ought not to exist for the subsample of interest (employed economically active individuals). Hence these observations were not imputed in the main analysis and analysed separately instead.

An important relationship that repeatedly presented itself in each section of this chapter was that of the relationship between questionnaire design and the resulting data structure. This made the analytical task iterative to an extent more than complex, for it required careful data checks and question wording and sequencing checks that mandated a fastidious and detail-oriented approach to the problems and interpretation of the results. An overall lesson learnt from this chapter is that it is incumbent upon researchers to be absolutely meticulous in their data preparation, imputation, estimation and analysis tasks when working with micro datasets.

The univariate approach to multiple imputation utilised here allowed for very specific sensitivity analyses to be performed. Four different specifications of the imputation models provided the basis for sensitivity analysis to mis-specification in the imputation algorithm. We used four different models for this purpose: a mis-specified algorithm (model 1), one that explained the response process only (model 2),

one that explained income itself (model 3), and a final one that combined covariates from model 2 and 3. It was this fourth model that was chosen as the first-best model, given the recommendations for covariate selection of Van Buuren et al. (1999). The main limitation with this model was a reduction in the estimation sample size due to the greater prevalence of covariate missing data compared to the other models.

The advantage of incorporating predictors for the response process in the imputation algorithm as well as earnings covariates was that it evidently reduced the right-skewness of the imputed monthly income values. The plausibility of imputed draws for the highest, open-ended income bracket, the refusals, don't know and unspecified response groups, was clearly affected by covariate selection in the imputation process. The fact that the first-best model reduced these values relative to the other three specifications suggests there is considerable merit to paying close attention to the response process in multiple imputation algorithms and not simply to predictors of the outcome variable.

This has important implications for more sophisticated multiple imputation exercises that seek to impute for covariate coarse data too, for it suggests that each variable with coarse observations needs: (1) a model of the coarse data mechanism for that variable (this would include checks for additional forms of survey error); (2) an analysis of the factors explaining the response process for that variable; and (3) appropriate prediction equations for that variable, which include covariates that explain both the response process and the outcome variable of interest.

References

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods & Research, 28,* 301–309.

Ardington, C., Lam, D., Leibbrandt, M., & Welch, M. (2006). The sensitivity of estimates of changes in post-Apartheid poverty and inequality to key data imputations. *Economic Modelling, 23,* 822–835.

Carpenter, J. R., Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: A weighting approach. *Statistical Methods in Medical Research, 16,* 259–275.

Daniels, R. C. (2008). *The income distribution with coarse data* (Working Paper Number 82). Cape Town: Economic Research Southern Africa.

Ghosh-Dastidar, B., & Schafer, J. L. (2003). Multiple edit / multiple imputation for multivariate continuous data. *Journal of the American Statistical Association, 98*(464), 807–817.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Preventative Science, 8,* 206–213.

Heeringa, S. G. (1995). Application of generalized iterative Bayesian simulation methods to estimation and inference for coarsened household income and asset data. In *The Proceedings of the Section on Survey Methods* (pp. 42–51). American Statistical Association.

Heeringa, S. G., Little, R. J. A., & Raghunathan, T. E. (2002). Multivariate imputation of coarsened survey data on household wealth. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse*. New Jersey: Wiley.

Heitjan, D. F., & Basu, S. (1996). Distinguishing "Missing at Random" and "Missing Completely at Random". *The American Statistician,50*(3), 207–213.

Heitjan, D. F. (1994). Ignorability in general incomplete data models. *Biometrika, 81,* 701–708.
Heitjan, D. F., & Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association, 85*(410), 304–314.
Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics, 19*(4), 2244–2253.
Kenward, M. G., & Carpenter, J. (2007). Multiple imputation: Current perspectives. *Statistical Methods in Medical Research, 16,* 199–218.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New Jersey: Wiley.
Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85–95.
Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association, 102*(480), 1462–1471.
Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal, 4*(3), 227–241.
Royston, P. (2005). Multiple imputation of missing values: Update. *The Stata Journal, 5*(2), 188–201.
Royston, P. (2007). Multiple imputation of missing values: Further update of ICE, with an emphasis on interval censoring. *Stata Journal, 7*(4), 445–464.
Royston, P. (2009). Multiple imputation of missing values: Further update of ICE, with an emphasis on categorical variables. *Stata Journal, 9*(3), 466–477.
Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association,91*(434), 473–489.
Rubin, D. B. (1976). Inference and missing data. *Biometrica, 63,* 581–592.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8,* 3–15.
Schwartz, L. & Paulin, G. (2000). Improving response rates to income questions. In *American Statistical Association (ASA) Section on Survey Research Methods, Proceedings* (pp. 965–970).
StataCorp. (2011). *Stata multiple imputation reference manual: Release 12*. College Station: StataCorp LP.
Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival models. *Statistics in Medicine, 18,* 681–694.
Vermaak, C. (2010). *The impact of multiple imputation of coarsened data on estimates of the working poor in South Africa* (World Institute for Development Economics Research (WIDER) Working Paper No. 2010/86). Helsinki: WIDER.
White, I. R., Wood, A., & Royston, P. (Eds). (2007). Editorial: Multiple imputation in practise. *Statistical Methods in Medical Research,16*, 195–197.
White, I. R., Royston, P., & Wood, A. (2011). Multiple imputation using chained equations: Issues and guidance for practise. *Statistics in Medicine, 30,* 377–399.
Wittenberg, M. (2008). *Nonparametric estimation when income is reported in bands and at points* (Working Paper Number 94). Cape Town: Economic Research Southern Africa.

Chapter 5
Conclusion: How Data Quality Affects Our Understanding of the Earnings Distribution

Household survey data are subject to multiple forms of survey error that can have a direct bearing on data quality, influencing end-user estimates of parameters of interest in unpredictable ways. This book has focussed specifically on employee income, but the insights are generalisable to any component of income.

Chapter Two developed a framework for investigating microdata quality that was based largely on the total survey error (TSE) paradigm, but that also included specific data quality control elements. The TSE framework decomposes survey error into coverage error, sampling error, nonresponse error, adjustment error, processing error, measurement error and validity. We focussed on adapting the total survey error framework to shed light on which aspects of data quality researchers can observe and do something about. This framework then served as the basis for evaluating the evolution of data quality in Statistics South Africa's labour market household surveys from the early 1990s to 2007.

It was argued that efforts to improve data quality should involve a virtuous interaction between producers and consumers of microdata and should be considered an evolving process. For producers of data, the preparation and publication of detailed data quality frameworks was emphasised, and two such examples were reviewed (the Statistics Canada and SSA Data Quality Frameworks). These frameworks are also excellent documents to inform users about issues of relevance to survey organisations and how these may affect the overall quality of the public-release data. For example, the late 1990s would have been an excellent time for the national statistics office (SSA) to inform users to expect variation over the repeated cross-sections of survey data due to non-sampling errors, and to explain that process in some detail. However, data quality frameworks were not in use by SSA at that time.

For consumers of data, judicious analyses of the univariate, bivariate and multivariate relationships in public-use versions of the datasets shed light on different components of survey error in variables of interest. Any problems associated there-

R. C. Daniels, *How Data Quality Affects our Understanding of the Earnings Distribution*,
https://doi.org/10.1007/978-981-19-3639-5_5

with should be communicated back to survey organisations. However, this does not make the analysis task any easier, and comparisons of repeated cross-sections of income data are particularly vulnerable to components of survey error directly under the control of the survey organisation. Ultimately, it was noted that improving data quality for income in particular is about improving data quality for household surveys in general.

Chapter Three isolated questionnaire design and item nonresponse for the employee income question in two South African labour market surveys: the October Household Survey (1997–1999) and the Labour Force Survey (2000–2003). The choice of time period isolated a period of changing questionnaire design for the employee income question. Between 1997 and 2000, the income question gradually included new response options for the respondent to state that they don't know or refuse to answer the question. We used sequential logistic response models to evaluate how improvements to the income question improved the capacity to understand the nonresponse and bounded response mechanisms. The use of these models represents an important contribution to the literature, for they can be used to evaluate the response process regardless of whether the bounded response question is in the form of a showcard, an unfolding bracket or a respondent generated interval.

It was found that the probability of initial nonresponse to the exact income question was correlated with income, but when the second follow-up bounded income question was presented to respondents, final nonresponse was no longer repeatedly associated with predictors of income. This suggested that the bounded income question overturned initial nonresponse to the exact income question and included more high income earners in the observed response subset. The addition of refuse and don't know response options to the employee income question played a very important role in improving the understanding of the nonresponse process, but in the final analysis of this chapter at least, respondents who refused to answer the employee income question were no longer significantly different to those who stated that they didn't know their income, at least as far as predictors of income were concerned. Rather, correlates of the knowledge of income became significant, with self-reporters and those cohabiting with romantic partners having the most consistently higher odds of refusing over time.

Chapter Four was concerned with conducting univariate multiple imputations for coarse response subsets of the employee income question. An analysis of the interrelationship between the exact income and bounded income variables released in the public-use data revealed a non-trivial degree of processing and/or measurement error for certain survey years between 1997–2003. We identified two forms of error that had to be dealt with effectively before multiple imputation could be performed. We also noted an idiosyncratic feature of the bounded employee income question in all of SSA's household surveys, namely the existence of a zero bracket. This was left in the data and not imputed because it was deemed to be a reasonable response value to the income question given the fact that employees could state they were not working due to being ill.

Once these features of the public-use data were effectively treated, we then conducted multiple imputations for coarse income observations using four differ-

ently specified models to test the sensitivity of imputed draws of income to misspecification in the imputation algorithm. It was found that a combination of response propensity and Mincerian earnings function covariates led to imputed draws that were the least likely to be extreme values in the income distribution, relative to alternative specifications. This has very important implications for more complex multiple imputation algorithms that seek to simultaneously impute income and covariate coarse data, an exercise that will require much initial data preparation and analysis before the integrity of the algorithm can be validated.

We then also evaluated the point estimates of quantiles and moments of the multiply imputed income distributions as the number of imputations increased, where it was found that stability in the estimates and inferences was achieved after only two imputations. This was likely a product of both the low percentage of item missing data and the restricted ranges of plausible imputed draws for the bounded income respondents. However, despite the low percentage of item missing data, it was found that imputed draws for refusals always had higher values than don't know respondents. This was a very important finding that was not discernible in Chapter Three, where predictors of the refuse subset no longer seemed to be different to the don't know subset on variables correlated with income.

The coarse data framework proved to be very useful in Chapter Four in guiding the approach to multiple imputation, not only because it informed the use of an interval censored regression algorithm, but also because it led to the decision rule to exclude unspecified responses in the LFS from being imputed in the primary analysis. When we then conducted a separate imputation process for these unspecified responses in 1999 and 2000, it was found that imputed draws were very differently distributed compared to imputed draws for don't know and refuse responses. This suggested that unspecified responses was an altogether different error process to item nonresponse on the employee income question, and should be treated as such.

Taken in combination, Chapters Three and Four show the necessary steps that researchers need to take when preparing the data for final estimates of univariate parameters of the earnings distribution. Post-imputation, poverty and inequality estimates can only then be thought of as accurate to the maximum degree possible given the data. This is true regardless of the country for which data is collected, which makes the methodology generalisable to any context. Limitations could still exist though, to the extent that unobservable components of survey error, such as frame error and sampling error, remain material.

In summary then, the presence of multiple sources of survey error in microdata need not impose undue constraints to the reliable estimation of parameters of the income distribution. What is required is that each source of survey error's potential impact on that distribution is known, even though nothing can be done about some of those components of error after public release of the data. For those components of error that can be observed, statistically rigorous methodology has to inform the approach to univariate and multivariate analyses, and researchers need to be explicit about their treatment of each relevant component of error.