DataFirst Technical Papers

DataFirst

SALDRU

A Framework for Investigating Micro Data Quality, with Application to South African Labour Market Household Surveys

by

Reza C. Daniels

Technical Paper Series
Number 19

About the Author(s) and Acknowledgments

Reza C Daniels - School of Economics and SALDRU, University of Cape Town, reza.daniels@uct.ac.za

This document has been produced as a chapter of my doctoral dissertation. I am grateful to the African Economic Research Consortium and the National Research Foundation for funding. I am also grateful to my supervisers – Murray Leibbrandt and Martin Wittenberg – for their guidance. All errors and/or omissions are the responsibility of the author.

This is a joint DataFirst and SALDRU paper.

Recommended citation

Daniels, R. (2012). A Framework for Investigating Micro Data Quality, with Application to South African Labour Market Household Surveys. A DataFirst Technical Paper Number 19. Cape Town: DataFirst, University of Cape Town

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

A Framework for Investigating Micro Data Quality, with Application to South African Labour Market Household Surveys[*]

Reza C. Daniels[†]

Abstract

In this paper the Total Survey Error (TSE) paradigm is combined with detailed data quality indicators to develop a framework for investigating micro data quality. The TSE framework is widely used in the survey methodology literature to identify different components of error that arise in the survey process. Consequently, it provides a very useful typology for researchers to understand which data quality issues are relevant in applied work based on these surveys. In order to demonstrate how the framework sheds light on micro data quality, two labour market household surveys conducted by Statistics South Africa are reviewed, spanning a time-frame from 1995-2007. It is argued that efforts to improve data quality should involve a virtuous interaction between producers and consumers of micro data and should be considered an evolving process. For producers of data, the preparation and publication of detailed data quality frameworks is recommended, and two examples of these frameworks are reviewed. For consumers of data, judicious analyses of the univariate, bivariate and multivariate relationships in public-use versions of the datasets can help shed light on different components of survey error, and should be communicated back to survey organisations. Ultimately, improving data quality is about being more explicit about the limitations of data production at each stage of the process, which does not stop at initial public release.

Keywords: Data Quality Evaluation and Assessment, Total Survey Error

JEL Codes: C81, C83

[*] This document has been produced as a chapter of my doctoral dissertation. I am grateful to the African Economic Research Consortium and the National Research Foundation in South Africa for funding. I am also grateful to my supervisers – Murray Leibbrandt and Martin Wittenberg – for their guidance, as well as three external examiners for their comments. All errors and/or omissions are the responsibility of the author.

[†] School of Economics & SALDRU, University of Cape Town. reza.daniels@uct.ac.za

1 Introduction

This paper identifies a framework for investigating micro data 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 paper 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 policy makers from making public statements about data quality that are often ill-informed and frequently incorrect.

The choice of time-period to investigate micro data 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, 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 micro data quality in South Africa during one of the most fascinating periods in the country's history.

The rest of this paper 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 micro data quality. We then apply this framework to SA labour market household surveys from 1995-2007. Lastly, we discuss the generalisability of the framework and its scope for application to other surveys and countries.

2 Framing the Discourse on Data Quality

Micro data quality is an artifact of a data production process controlled by survey organisations with finite budget constraints. 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 micro data need to be considered when attempting to create a framework for investigating micro data quality.

In this section we locate the discourse of creating a framework for micro data 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.

2.1 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 micro data quality has also raised the primacy of this debate (see Flinn, Kulka, Moffitt and Wolpin, 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 (who 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.

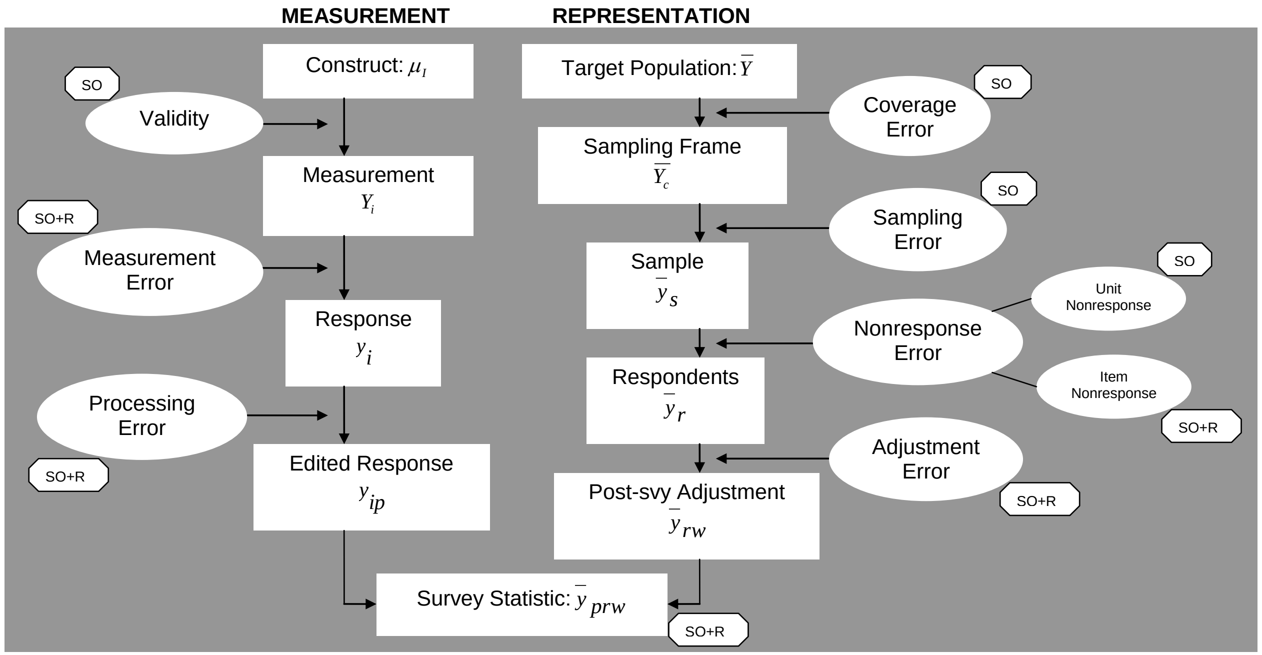
2.2 The Total Survey Error (TSE) Framework

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 themselves 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. Under the first theme, the possible sources of error include validity of the construct, measurement error and processing error. For the second theme, the sources of error include coverage error, sampling error, nonresponse error, and adjustment error.

Because researchers and survey organisations frame the concept of data quality differently, it is helpful to consider the agency of these two groups in the TSE framework. Figure 1 below presents a schematic overview of TSE.

Figure 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

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, etc.).

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 figure 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 variables 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).

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 provides 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 (2009, 2010). This is largely to anchor the relatively abstract discussion of Total Survey Error within the context of the practical realities faced by statistical organisations. 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".

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 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.

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 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 guideliness 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 edit 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.

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.

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 Statistical 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 Petterson, 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.

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, 2012). 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 catalogs the information needed to understand the trade-offs of decisions that affect the accuracy of the outcomes.

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 program, 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 allows the item response process to be understood using the same general methods for understanding unit response (see Daniels, 2012b for an example of this).

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 probabilities 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.

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 micro data 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 was 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.

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 policy makers 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.

4.1.1 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 are undertaking surveys every year, and employing 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.

4.1.2 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 (Pettersen, 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.

4.1.3 Nonresponse Error

There are two components of nonresponse, namely unit and item nonresponse. Our focus here is on unit nonresponse only. 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 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 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 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

The table 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).

4.1.4 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.

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.

4.2.1 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: 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.

4.2.2 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 interviewer surveys (Groves & 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 below.

Table 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

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 Vasquez-Alvarez (2003) and Melenberg, van Soest and Vasquez-Alvarez (2006) have demonstrated, the unfolding bracket design introduces anchoring bias. Anchor strategies are purposefully introduced into surveys to aid respondent recall (see Blair, Menon & Bickart, 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 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.

The table below 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.

Table 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

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.

4.2.3 Processing error

The impact of processing error on the survey is often difficult to detect for the income question specifically, and there are potentially significant implications of it. 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 (see Daniels, 2012a for a discussion of processing error in the income variable in SSA surveys). Other examples of processing error in the income question include:

- Incorrectly coding an income value, for example by inputing 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.

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 policy makers 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 orgnisation.

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 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.

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 on the one hand (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 into 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.

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 paper has been to adapt the TSE framework into one that recognised the limited agency of researchers to assess data quality. This was distinct from a discussion of how survey organisations shape data quality and survey errors given their human resource and budget constraints. This helped create a framework for investigating micro data quality that was sensitive to the capacity of agents to diagnose data quality in the first place.

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 policy makers 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

[1] Bhorat, H., 1999, “The October Household Survey, unemployment and the informal sector: A note”, *South African Journal of Economics*, 67:2, 320-326

[2] 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., *Measurement error in surveys*, New York: Wiley

[3] Brackstone, G., 1999, “Managing Data Quality in a Statistical Agency”, *Survey Methodology*, 25:2, 139-149

[4] 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”, *Southern Africa Labour and Development Policy Research Unit (SALDRU) Working Paper Series Number 38*, SALDRU and Data First, Cape Town

[5] Branson, N. and 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

[6] Branson, N. and 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 Papers 54, Cape Town: Southern Africa Labour and Development Research Unit (SALDRU), University of Cape Town

[7] Casale, D. and 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

[8] Daniels, R.C., 2012a, “Univariate multiple imputation for coarse employee income data”, Southern Africa Labour & Development Research Unit (SALDRU) Working Paper Number 88, Cape Town: SALDRU

[9] Daniels, R.C., 2012b, “Questionnaire design and response propensities for employee income micro data”, Southern Africa Labour & Development Research Unit (SALDRU) Working Paper Number 89, Cape Town: SALDRU

[10] Daniels, R.C. and Rospabe, S., 2005, “Estimating an Earnings Function from Coarsened Data using an Interval Censored Regression Procedure”, *Studies in Economics and Econometrics*, 29:1

[11] Daniels, R.C. and 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

[12] Flinn, C.J., Kulka, R., Moffitt, R., and Wolpin, K.I., 2001, “Introduction to the Journal of Human Resources Special Issue on Data Quality”, Journal of Human Resources, 36:3, 413-415

[13] 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, United Nations Statistics Division, New York

[14] Groves, R.M., 1991, “Measurement Error Across the Disciplines”, in Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. and Sudman, S. (eds), *“Measurement Errors in Surveys”*, Wiley, New Work

[15] Groves, R.M., 2004, *“Survey Errors and Survey Costs”*, Wiley Press, New York

[16] Groves, R.M. and Couper, M.P., 1998, *“Nonresponse in Household Interview Surveys”*, Wiley Press, New York

[17] Groves, R.M., Fowler Jr, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R., 2004, *“Survey Methodology”*, Wiley Press, New York

[18] Heeringa, S.G. and 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

[19] Kerr, A. and 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

[20] Kish, L., 1965, “Survey sampling”, New York: Wiley

[21] Krotki, K.P., 2008, “Sampling error”, in Lavrakas, P.J. (Ed.), Encyclopedia of Survey Research Methods, Volume 2 . Thousand Oaks: Sage Publications

[22] Luxembourg Income Study (LIS), 2011, www.lisproject.org

[23] McCutcheon, A.L., 2012, “Sampling bias”, in Lavrakas, P.J. (Ed.), Encyclopedia of Survey Research Methods, Volume 2 . Thousand Oaks: Sage Publications

[24] 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

[25] 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, United Nations Statistics Division, New York

[26] Rubin, D.B., 1976, “Inference and missing data”, *Biometrica*, 63, 581-592

[27] Rubin, D.B., 1987, *Multiple imputation for nonresponse in surveys*, New York: Wiley

[28] Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2008, Wave 1 [dataset]. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2010. Cape Town: DataFirst [distributor], 2010

[29] Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2010-2011, Wave 2 [dataset]. Version 1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2012. Cape Town: DataFirst [distributor], 2012

[30] Statistics Canada, 2000, *Policy on Informing Users of Data Quality and Methodology*, Ottawa: Statistics Canada

[31] Statistics Canada, 2003, *Statistics Canada Quality Guidelines. Fourth Edition*, Ottawa: Statistics Canada

[32] Statistics Canada, 2009, *Statistics Canada Quality Guidelines. Fifth Edition*, Ottawa: Statistics Canada

[33] Statistics South Africa (SSA), *October Household Survey 1995 Metadata*, Pretoria: SSA

[34] SSA, *October Household Survey 1996 Metadata*, Pretoria: SSA

[35] SSA, *October Household Survey 1997 Metadata*, Pretoria: SSA

[36] SSA, *October Household Survey 1998 Metadata*, Pretoria: SSA

[37] SSA, *October Household Survey 1999 Metadata*, Pretoria: SSA

[38] SSA, *Labour Force Survey 2000 February Metadata*, Pretoria: SSA

[39] SSA, *Labour Force Survey 2000 September Metadata*, Pretoria: SSA

[40] SSA, *Labour Force Survey 2001 February Metadata*, Pretoria: SSA

[41] SSA, *Labour Force Survey 2001 September Metadata*, Pretoria: SSA

[42] SSA, *Labour Force Survey 2002 February Metadata*, Pretoria: SSA

[43] SSA, *Labour Force Survey 2002 September Metadata*, Pretoria: SSA

[44] SSA, *Labour Force Survey 2003 February Metadata*, Pretoria: SSA

[45] SSA, *Labour Force Survey 2003 September Metadata*, Pretoria: SSA

[46] SSA, *Labour Force Survey 2004 February Metadata*, Pretoria: SSA

[47] SSA, *Labour Force Survey 2004 September Metadata*, Pretoria: SSA

[48] SSA, *Labour Force Survey 2005 February Metadata*, Pretoria: SSA

[49] SSA, *Labour Force Survey 2005 September Metadata*, Pretoria: SSA

[50] SSA, *Labour Force Survey 2006 February Metadata*, Pretoria: SSA

[51] SSA, *Labour Force Survey 2006 September Metadata*, Pretoria: SSA

[52] SSA, *Labour Force Survey 2007 February Metadata*, Pretoria: SSA

[53] SSA, *Labour Force Survey 2007 September Metadata*, Pretoria: SSA

[54] SSA, 2006a, *Draft Data Quality Framework 001: South African Statistical Quality Assessment Framework*, Pretoria: SSA

[55] SSA, 2006b, *Data Quality Policy 001: Policy on Informing Users of Data Quality*, Pretoria: SSA

[56] SSA, 2009, *South African Statistical Quality Assessment Framework*, Pretoria: SSA

[57] SSA, 2010, *South African Statistical Quality Assessment Framework: Second Edition*, Pretoria: SSA

[58] United Nations Statistics Devision (UNSD), 2011, *"Development of National Statistics Systems"*, accessed on the web at: http://unstats.un.org/unsd/dnss/,

[59] Van Der Berg, S. and Louw, M., 2004, "Changing Patterns of South African Income Distribution: Towards Time Series Estimates of Distribution and Poverty", *South African Journal of Economics*, 72:3, 546-572

[60] Vazquez-Alvarez, R., 2003, "Anchoring bias and covariate nonresponse", Mimeo, Version: July, 2006, St Gallen University, St Gallen

[61] Whitford, D.C. and 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*, Statistics Division, Department of Economics and Social Affairs, United Nations Secretariat, New York

[62] 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

[63] 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

[64] World Institute for Development Economics Research (WIDER), 2012, www.wider.unu.edu/research/Database

[65] Yansaneh, I.S., 2005, "Introduction", in *Household Sample Surveys in Developing and Transition Countries*, Mimeo: Department of Economic and Social Affairs, United Nations Statistics Division, New York

[66] Yu, D., 2007, "The Comparability of the Statistics South Africa October Household Surveys and Labour Force Surveys", *Stellenbosch Economic Working Papers: 17/07*, University of Stellenbsoch, Stellenbosch

[67] Yu, D., 2009, "The Comparability of Labour Force Survey and Quarterly Labour Force Survey", *Stellenbosch Economic Working Papers: 08/09*, University of Stellenbsoch, Stellenbosch

About DatatFirst

DataFirst is a research unit at the University of Cape Town engaged in promoting the long term preservation and reuse of data from African Socioeconomic surveys. This includes:

- the development and use of appropriate software for data curation to support the use of data for purposes beyond those of initial survey projects
- liaison with data producers - governments and research institutions - for the provision of data for reanalysis
- research to improve the quality of African survey data
- training of African data managers for better data curation on the continent
- training of data users to advance quantitative skills in the region.

The above strategies support a well-resourced research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708