DataFirst Technical Papers

Questionnaire Design and Response Propensities for Employee Income Micro Data

by
Reza C. Daniels

Technical Paper Series
Number 18

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.C. (2012). Questionnaire Design and Response Propensities for Employee Income Micro Data. A DataFirst Technical Paper Number 18. 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

Questionnaire Design and Response Propensities for Employee Income Micro Data[*]

Reza C. Daniels[†]

Abstract

The design of the income question in household surveys usually includes response options for actual income, bounded values, "Don't Know" and "Refuse" responses. This paper conducts an analysis of these response types using sequential response models. We analyse the employee income question in Statistics South Africa's October Household Surveys (1997-1999) and Labour Force Surveys (2000-2003). The choice of survey years coincides with a period of development of the income question during which additional response options were steadily introduced to the questionnaire. An analysis of this sort sheds light on the underlying response process, which is useful for survey planning purposes and to researchers concerned with diagnosing the item missing and partial response mechanisms for variables of interest. It was found that the presentation of follow-up brackets overturned initial refusals to the income question, and that these respondents were higher income earners. In the sequential response models, initial nonresponse was therefore clearly correlated with predictors of income, but after the presentation of the bracket showcards, this correlate of income effect was removed. This suggested that final nonresponse was no longer a function of income. This has important implications for ignorability determination and (single or multiple) imputation strategies.

Key Words: Questionnaire Design, Response Propensity Models, Ignorability, Employee Income

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

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 paper 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 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 implications for single and multiple imputation strategies for missing data.

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 paper is to investigate the bounded response and item nonresponse processes analgously.

The paper 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.

2 Questionnaire Design and the Income Question

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 & 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, Eltinge, Groves and Little, 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, underreporting and measurement error. Hurd, Juster and Smith (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, Kahneman and Mochona, 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 original 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.

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 and Marquis, 2001; Press, 2004).

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, Juster and Smith (2003) find that there is little evidence of anchoring bias in the Health and Retirement Study (HRS) in the USA, but Juster, Smith and Stafford (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, Cao, Couper, Hill, Hurd, Lutpon, Perry and Smith, 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 estimates. 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.

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

[1] Note that our treatment of "Don't Know" responses as a form of nonresponse takes its precedence from Rubin, Stern and Vehovar (1995). However, this definition imposes

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 non-participation 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.

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, O-Rourke, Burris and Owens (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.

no constraints on the analysis, and later we consider "Don't Know" as a partial form of response because it reveals at least some information about income, as opposed to refusals.

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.

2.4 Questionnaire Design Changes in SA Labour Market Household Surveys

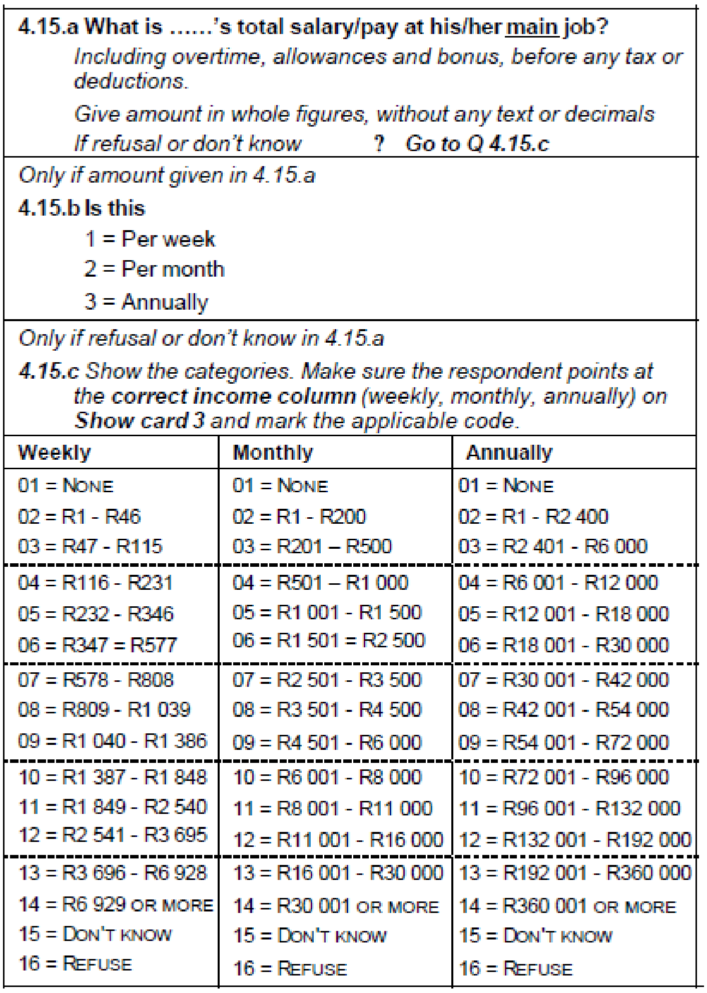
We evaluate employee income in South Africa's two major household interview labour market surveys: the October Household Surveys (OHS; 1997-1999), and the Labour Force Surveys (LFS; 2000-2003 September waves only). The OHS was a repeated cross-sectional survey, while the LFS was a biannual rotating panel 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 was 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 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 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, Menon, and Bickart, 1991). The intuition behind this is simple: a proxy reporter is less likely to know the

Figure 1: The Income Question: Labour Force Survey 2000 September



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

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 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.

3.1 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].

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

- 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.

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.

3.2 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.

3.3 Estimation, Specification and Testing

3.3.1 Estimation

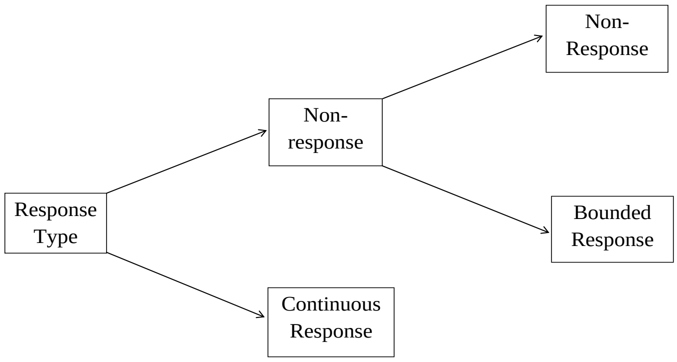
We can think of response propensity models for employee income as modelling a latent variable for the *unwillingness* to disclose income. This variable is not directly observed, but we do observe the response type for the income question, which gives us information about the level of information disclosure 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].

[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 figure 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.

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 2 and 3 depict this.

Figure 2: The Employee Income Response Process in OHS 1997 and 1998

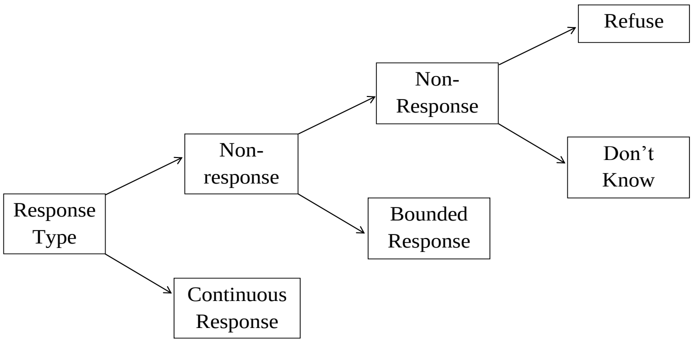


From figure 2, we see that the respondent can first provide an exact income value or state that they don't know or refuse (collectively grouped as "nonresponse" in the figure). The interviewer then prompts the respondent to answer again, this time with a bounded response follow-up question presented 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 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 below presents this sequential structure.

Figure 3: The Employee Income Response Process in LFS 2000-2003



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 Figure 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:

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

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 figure 3 and equation 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 equation 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).

3.3.2 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 section 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 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.

Table 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-cultural influence to disclose income, though effects ambiguous	Willingness to disclose (WD)
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: (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	Correlate of Income (CI)
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 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. Naturally, a question arises about the relationship between nonresponse on income and nonresponse on expenditure, which we explore in the analysis below.

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, 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 paper 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.

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.

4.1 A Descriptive Analysis of Employee Income Response Type

Table 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.

Table 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

The percentage of exact responses in each survey year ranges from 87 percent in 2000 to 54 percent 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 percent of the sample in 2000 to 37 percent 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 percent in 2000 to about 7 percent 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 is steadily increasing 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 exact subset) are converted into bounded ranges before the probability is calculated. Table 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 percent 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 percent 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.

Table 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

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, where we see it is lowest in 2000 at 10 percent and highest in 1998 at 38 percent. 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.

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 figure 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 figure 3 and equation 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.

4.2.1 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 4 and the second-stage results are reported in table 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 figure 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 1 for a recap of the variable groups).

Table 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 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 1997-1999 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.

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 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 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 percent in 1997 and 0.33 percent 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 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.

[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 paper, 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 subsection 3.3.1 above.

Table 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***
>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%

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 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 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 percent, while a self-reporter reduces the odds of nonresponse ten-fold. Since 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 percent, while an adult (aged 16-64) would reduce the odds of nonresponse by 32 percent.

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 percent. 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 percent. Men have odds that are 63 percent 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.

4.2.2 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 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***
>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%

Table 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 (aged 16-64 years) 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 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 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 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 5 above, suggesting that it plays a diminished or non-existent role in explaining further nonresponse beyond the first stage of income reporting.

Table 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***
>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 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 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 below.

Table 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**
>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%

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 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.

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.

4.3.1 Model Fit

In this section we discuss model fit for the sequential logistic response models above 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 equation 1 above. The H-L test results in Table tab:rp10 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 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 percent 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 the subsample size is sufficiently large for stable results.

Table 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

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 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.

4.3.2 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 mitigated 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 10 presents the percentages of joint nonresponse for each survey year and the denominator subsample size in the percentage calculations.

Table 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

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 percent 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 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 percent of all nonresponse on income in the LFS, but rising to a very high 47 percent 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. 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 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 section 4.2.2, omitting the expenditure variables from each year. Table 11 presents the results for 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 percent 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 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 8	1631	336	724	648	542

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

[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.)

Table 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 8 shows somewhat similar findings, but given that the main finding in table 8 was that there were no stable findings across the years, this is not particularly informative. One identical effect in table 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 8 (i.e. 1999, 2001-2003).

In the two years when the expenditure category is always significant in Table 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 11 is for education in 2003, where the coefficients have now nearly doubled in magnitude and become significant (compared to table 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.

5 Conclusion

The main objective of this paper 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 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

[1] Beatty, P. and Herrmann, D., 2002, "To answer or not to answer: Decision processes related to survey item nonresponse", in Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (eds), *Survey nonresponse*, New Jersey: Wiley

[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 Jersey: Wiley

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

[4] Buis, M., 2012, "seqlogit: *Stata* module to fit a sequential logit model", Version number 1.1.15, `http://maartenbuis.nl/software/seqlogit.html`

[5] Cantwell, P.J., 2008, "Rotating Panel Design", in Lavrakas, P.J. (ed), *Encyclopedia of Survey Research Methods*, Thousand Oaks: Sage Publications

[6] 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

[7] 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

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

[9] Dillman, D.A., Eltinge, J.L., Groves, R.M. and Little, R.J.A., 2002, “Survey nonresponse in design, data collection, and analysis”, in Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (eds), *Survey nonresponse*, New Jersey: Wiley

[10] Frederick, S., Kahneman, D., Mochona, D., 2010, “Elaborating a Simple Theory of Anchoring” *Journal of Consumer Psychology*, Vol 20(1), 17-19

[11] Groves, R.M. and Couper, M.P., 1998, *Nonresponse in household interview surveys*, New Jersey: Wiley

[12] 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*, Vol 38, 758-772

[13] Jacowitz, K.E. and Kahneman, D., 1995, “Measures of Anchoring in Estimation Tasks”, *Personality and Social Psychology Bulletin*, 21(11), 1161-1166

[14] Johnson, T.P., O’Rourke, D., Burris, J., and Owens, L., 2002, “Culture and survey nonresponse”, in Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (eds), *Survey nonresponse*, New Jersey: Wiley

[15] Juster, T.F. and Smith, J.P., 1997, “Improving the Quality of Economic Data: Lessons from the HRS and AHEAD”, *Journal of the American Statistical Association*, Vol 92(440), 1268-1278

[16] Juster, F.T., Smith, J.P., Stafford, F., 1999, “The Measurement and Structure of Household Wealth”, *Labour Economics*, Vol 6, 253-275

[17] 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

[18] Maddala, G.S., 1983, *“Limited dependent and qualitative variables in econometrics”*, Cambridge: Cambridge University Press

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

[20] Press, S.J. and 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

[21] Press, S.J. and Tanur, J.M., 2004, "Relating Respondent-Generated Interval Questionnaire Design to Survey Accuracy and Response Rate", *Journal of Official Statistics*, Vol 20(2), 265-287

[22] Press, S.J. and Tanur, J.M., 2005, "An Overview of the Respondent-Generated Intervals (RGI) Approach to Sample Surveys", *American Statistical Association (ASA) Section on Survey Research Methods*, Proceedings, 3487-3493

[23] Rubin, D.B, Stern, H.S., and 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

[24] Schwarz, N. and Hippler, H-J, 1991, "Response alternatives: The impact of their choice and presentation order", in Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S., *Measurement error in surveys*, New Jersey: Wiley

[25] Schwartz, L. and Paulin, G., 2000, "Improving Response Rates to Income Questions", *American Statistical Association (ASA) Section on Survey Research Methods*, Proceedings, 965-970

[26] Vazquez-Alvarez, R., 2003, *"Anchoring bias and covariate nonresponse"*, Mimeo, Version: August, 2003, St Gallen University, St Gallen

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

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

SPES BONA