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 for it implies that the missingness mechanism is likely missing at random.

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. This paper discusses the design of the employee income question and evaluates the characteristics of respondents who report their incomes as exact values, bracketed values, or nonresponse (refusals, don't know).

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 survey years with "don't know" and "refuse" options only introduced to the questionnaires in 2000. The introduction of

*This document is a summary of 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. For the full paper, please email the author.

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

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.

2 Methodology

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 in Stats SA's questionnaires, 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.

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, define the outcome variable Y to have four possible alternatives:

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

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

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

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

2.1 Results

A three-stage response propensity model is now estimated 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. Note

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.

Table 1: 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 1 shows that for variables associated with the cognitive burden of answering the income question, 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.

For personal characteristics variables, 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 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, the finding for Indian people can be interpreted as new evidence of a socio-cultural effect to survey responding in South Africa.

We now turn to the second stage of the sequential response model, which evaluates nonresponse that combines refusals with don't know responses and compares it to bracketed responses. Table 2 presents the results.

Evident from the table is that variables associated with cognitive burden of response 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.

Table 2: 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 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 noteworthy for it suggests that age plays a diminished or non-existent role in explaining further nonresponse beyond the first stage of income reporting.

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.

The third stage of the model is not reported here (contact the author). In this stage of the model we start seeing very large effect sizes for certain variables, indicative of small cell sizes in this stage of the response model, leading to near perfect prediction of the outcome. 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.

3 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.

The sequential logistic response model proved to be a suitable estimator for response propensities to employee income. The overall results from the first stage of the sequential response models was that initial nonresponse was strongly associated with variables correlated with income. 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 the results were unstable across the years except for self-reporting, which was large and significant in every survey year except 2000.

The finding that final nonresponse was not consistently predicted by the same covariates suggests an apparent randomness to income missing data for employees. This is an important finding because much of the literature in econometrics on income nonresponse has argued that, on a-priori grounds, it should be treated as non-random. The apparently random nature of nonresponse was found to be largely due to the existence of the bracketed response follow-up question, which attenuates bias in the employee income distribution.

References

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

[2] Statistics South Africa (SSA), 1999, *"October Household Survey"*, Pretoria: SSA

[3] Statistics South Africa (SSA), 2000-2003, *"Labour Force Survey"*, Pretoria: SSA