13. August 2010

Recent developments in interviewer training at Statistics Iceland: Minimizing interviewer effects and reducing refusal rates.

Anton Örn Karlsson
anton.karlsson@statice.is
Statistics Iceland
Surveys

Abstract: In 2009, Statistics Iceland started using a new training progam for CATI-interviewers. The training program is mainly aimed at two factors: 1) Minimizing the number of refusals and, 2) Standardizing data collection to reduce the effect each interviewer has on the data he or she collects. Data from EU-SILC and ICT show that while some progress has been made, there still is room for improvement, especially when it comes to the effects of interviewers on collected data.

1. Introduction

Types of survey errors

Survey errors can be divided into four categories (see, for example, Groves, 1989): Coverage error, non response error, sampling error and measurement error. Much of survey work is aimed at reducing the effects of these errors on the results of the survey. For example a specific sampling design can be employed to increase accuracy of survey estimates and thus reducing the effects of sampling error. Another possibility is to combine a specific register with other references to be able to make a more accurate sampling frame and reduce coverage error.

Reducing errors with training

Survey managers can also use interviewer training to reduce the effects of certain types of survey error. Specifically, interviewer training can be used to increase response rate[1] through refusal aversion training (Groves & McGonagle, 2001; Mayer & O'Brien, 2001). Also, interviewer related error can be reduced by training standardized interviewer techniques (Fowler & Mangione, 1990; Fowler, 1991). From 2009, interviewers working at Statistics Iceland (SI) have been trained using both of these methods.

Interviewer training for ICT and SILC

In total, SI employs 13 CATI interviewers and 8 field interviewers on a permanent basis. In addition, about 40 interviewers are hired and trained by SI in January every year to collect data for two surveys: Survey on Information and Communication Technology usage in households and by individuals (ICT) and The European Union Survey on Income and Living Conditions (SILC). In 2009 the training program for interviewers hired on a temporary basis was changed with two main goals in mind: 1) Reducing interviewer related error in the data and 2) Minimizing refusal rates. In reaching the first goal, a large part of the training program was dedicated to the four basic techniques of standardized interviewing (Fowler & Mangione, 1990; Gwartney, 2007): Reading questions as worded, probing inadequate or incomplete answers, recording answers directly and being neutral with regards

[1] Even though the relationship between non response bias and response rates are not perfect, high response rates can reduce bias in survey estimators (Groves & Peytcheva, 2008).

to the respondent. This training was both in the form of short lectures and practical assignments for the interviewers. Similar methods were used in reaching the goal of minimizing refusal rates; short lectures with practical assignments, both based on the ideas of Groves and McGonagle (2001) about how the interviewer tailors the participation request (see also Morton-Williams, 1993).

Assessment of the new training program.

Intraclass correlation coefficient

One way of assessing if the two goals have been reached is to compare the performance of interviewers the year before the new training program was first used with the performance of interviewers in the first year the program was used[2]. Since the main goals of the new training program were to reduce refusal rates and interviewer effects, it is natural to examine refusal rates of both surveys over both years and also the Intraclass Correlation Coefficient (ICC). According to Fowler and Mangione (1990) the ICC indicates how much effect interviewers have on the data they collect. Perfectly standardized interviewers should not affect the answers they collect and therefore the ICC would be 0.0. If the ICC is something else than 0.0 then the interviewers have an effect on the answers. The higher the ICC, the more effects the interviewers have on the data.

The purpose of the paper

The objective of this paper was to compare refusal rates for both ICT and SILC for the years 2008 and 2009 and compute, and compare, ICC for selected questions in the questionnaires from both years. The main hypothesis is that refusal rates and ICC's should be lower in 2009 than in 2008.

2. Method

2.a. Data collection

Data collection

In 2008, data was collected from the beginning of January to the end of May, beginning with the ICT, which was followed by a travel survey and finally SILC was conducted. In 2009 data was collected from the beginning of February to the beginning of May, starting with the SILC. Data for both surveys in both years was collected exclusively via telephone.

2.b. Calculations

Calculations of results

Refusal rates were calculated as number of refusals divided by the number of eligible sampling units, with ineligible units being those living abroad, persons living in institutions and deceased sampling units. The definitions and non response codes were the same for both surveys.
For calculations of ICC a small number of questions were selected beforehand from each survey: In SILC, all questions about amounts regarding the habitat of the household, debts and other economic issues of the household were selected. From the ICT, yes/no questions (recoded into 0/1 dichotomous variables) about the technical equipment in the household were assessed.

Criteria for calculations

Three types of criteria were used before calculating the ICC: 1) Each question had to have at least 100 answers, 2) Each interviewer used in the calculations had to have conducted at least 10 interviews in the survey[3], 3) The wording of the questions had to be the same for both years.

[2] The results from such a comparison should only be used for guidance because many variables will not be controlled in a nonexperimental design like this.

[3] SI employs a polish speaking interviewer to conduct interviews with polish sampling units. Since all interviews in polish are diverted to her, she is the only interviewer who is not assigned to interviews by the CATI system and

A specific function was written in the statistical software R in order to calculate the ICC according to the formula found in Kish (1965; see also Fowler & Mangione, 1990; Groves, 1989). Significance between two correlation coefficients was assessed by converting the coefficients into z-scores by Fisher's *z*-score transformation, dividing the outcome by the coefficient's standard error and then by comparing the two scores using a one tailed *t*-distribution. A similar method as used by Sayles, Belli & Serrano (2010) in determining if ICC's were significantly different from 0.

3. Results

Table 1. Number of interviewers and mean number of interviews.

Survey	Number of interviewers	Interviews per interviewer
ICT2008	29	55
ICT2009	37	48
SILC2008	34	76
SILC2009	44	65

Number of interviewers and mean number of interviews

The number of interviewers and average number of interviews per interviewer is presented in table 1. Fewer interviewers conducted interviews in 2008 than in 2009, perhaps reflecting the increase in unemployment at the same time in Iceland (Statistics Iceland, 2010) which meant that more people were available for working as interviewers in 2009 than in 2008. Also, the mean number of interviews conducted by each interviewer was lower in 2009; the average interviewer conducted 7 interviews less in the ICT in 2009 and 10 less interviews in the SILC than the year before.

3.a. Interviewer variability

Table 2. ICC for SILC2008 and 2009.[4]

Question	2008	*n*	2009	*n*
Rent	-0.037	349	**0.003**	409
Rent support	**-0.063**	114	0.128	134
House fund	-0.086	1640	***0.074**	1690
Maintenance fund	**0.013**	2449	0.020	2431
Rent idea	0.110	2448	**#0.040**	2430
Lowest income	0.084	2824	**#0.009**	2871
Lowest income guess	-0.024	306	**-0.012**	395
Alimony received	**-0.032**	310	-0.075	318
Alimony paid	0.262	255	**#0.014**	238
Financial support	**0.060**	210	0.144	266
Weighted average	0.040		**0.034**	

Lower numbers for each question are written in boldface.
** Significant, α = 0,01, one-tailed.
*Significant, α = 0,05, one-tailed.
$ Significant, α = 0,1, one-tailed.
#Marginally significant, α = 0,15, one-tailed.

ICC for SILC

In table 2 are the ICC's for selected questions from SILC2008 and 2009. The majority of the coefficients are lower in 2009 than in 2008 with six out of ten coefficients being lower. Only in one instance was the difference between the coefficients significant, in a question about amount paid for the households' house fund. In three instances the 2009 coefficients were mar-

thus violates the assumption that each interviewer must be assigned to cases randomly (Fowler & Mangione, 1990; McGraw & Wong, 1996). Her interviews were therefore not included in the calculations of the ICC.

[4] A translation of the questions can be found in appendix 1.

ginally significantly lower than in 2008. The weighted average of the ICC's for the selected variables was lower in 2009 than in 2008, or 0.034 compared to 0.04.

Table 3. ICC for SILC2008 and 2009.

Question	2008	*n*	2009	*n*
TV	0.152	1597	***-0.033**	1635
VCR	***0.327**	1597	0.468	1635
Games console	***0.017**	1597	-0.147	1635
Desktop	**0.073**	1597	-0.106	1635
Laptop	**0.028**	1597	0.124	1635
Palmtop	0.062	1597	**-0.045**	1635
DVD player	****0.111**	1597	0.452	1635
MP3 player	***0.129**	1597	0.286	1635
TV record	0.163	1597	**0.139**	1635
Flatscreen TV	**$0.097**	1597	-0.203	1635
Theater system	0.346	1597	****0.024**	1635
n of desktops	**0.029**	1597	0.086	1635
n of laptops	-0.032	1597	**0.004**	1635
Satellite dish	0.052	1597	**0.048**	1635
n of TV	**-0.046**	1597	0.049	1635
TV	0.152	1597	***-0.033**	1635
VCR	***0.327**	1597	0.468	1635
Weighted average	0.101		**0.076**	

Lower numbers for each question are written in boldface.
** Significant, α = 0,01, one-tailed.
*Significant, α = 0,05, one-tailed.
$ Significant, α = 0,1, one-tailed.

ICC for ICT

In table 3 are ICC's for selected questions from ICT2008 and 2009. The results are a bit different for the ICT then SILC as can be seen in that the majority of the coefficients are *higher* in 2009 than in 2008. Ten out of seventeen coefficients were actually lower in 2008 than 2009, which is the opposite pattern of results than was expected. Also, in six cases were the coefficients significantly higher in 2009 than in 2008, and in three cases was the difference significant in the other direction. The average ICC was lower in 2009 than in 2008, or 0.076 compared to 0.101.

3.b. Refusal rates

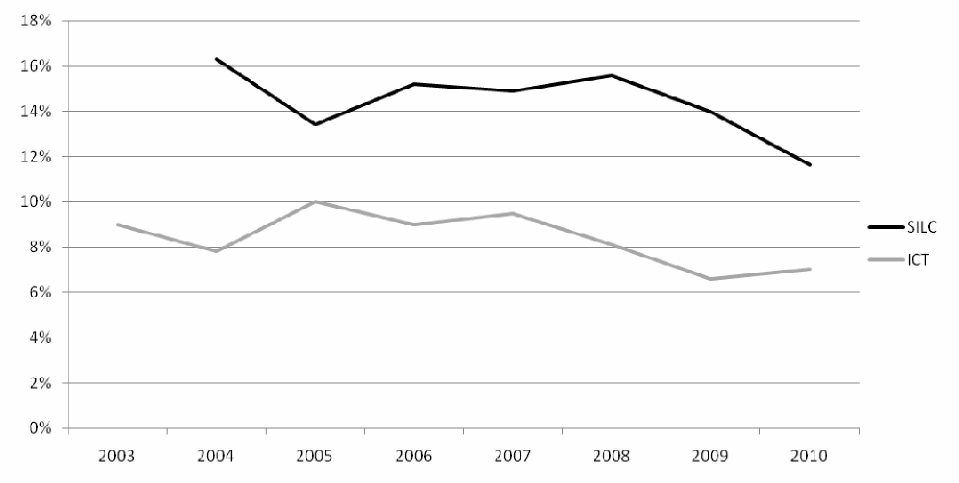

Figure 1. Refusal rates for ICT and SILC from 2003 – 2010

Refusal rates for both surveys from 2003 - 2010

In figure 1 the refusal rates for ICT and SILC can be seen, from 2003 for ICT and from 2004 for SILC. The refusal rate for SILC has dropped for two years in a row (13.9% in 2009 and 11.6% 2010) after reaching the second highest percentage of refusal in 2008 (15.5%) since the beginning of the survey in 2004 when the refusal rate was 16.3%.

In ICT, the refusal rate was highest in 2005, when 10% of the eligible sampling units refused to take part in the survey. From 2006-2008 the refusal rate was between 8 – 10%. In 2009, however, the rate dropped below 7% for the first time, with a final refusal rate of 6.6% and in 2010 the rate was 7%. The rates for these two years are the lowest refusal rates for the ICT ever, which is in accordance with the hypothesis of this paper.

4. Discussion

4.a. Interviewer variability

High ICC's

According to Groves and Magilavys (1986; see also Collins, 1980) overview of studies on ICC, the mean coefficients reported in this paper are unusually high, since three out of four were higher than 0.4. A sign that the training program is partially successful is that the mean of the coefficients was lower in 2008 than in 2009, a result that was in accordance with the working hypothesis of this paper. On the other hand, in the ICT the majority of coefficients were larger in 2009 than 2008. According to research by Bradburn and Sudman from 1979 (cited by Fowler & Mangione, 1990), experienced interviewers seem to use less standardized methods than those who are inexperienced. For example, experienced interviewers tend to ask questions without following the exact wording and, sometimes, leave out alternatives for the respondent to choose from. This could explain the results of the comparison of the ICC's from individual questions in the ICT between 2008 and 2009 because in 2008 the data for ICT was collected before data was collected for the SILC. In 2009 the sequence was reversed and the interviewers started by working on SILC. This means that in 2008 the interviewers were already experienced when SILC started after collecting data for the ICT, but in 2009, it was the other way around, they were experienced when the ICT started and could possibly have shown a tendency to use less standardized methods when collecting the data.

The effects of interviewer experience on standardization

How to deal with high ICC's

Therefore, it could be that the effects of the training scheme were only to be found in the first survey but not in the second survey since by then the interviewers had become experienced and they may have decided that it was not important to follow the questionnaire word for word. If that is the case, it is necessary for SI to follow the interviewers' behaviour when they are conducting interviews, and intervene if they move away from standardized interviewing. This is very important because, like Groves and Magilavy (1986) point out, a workload of 50 interviews per interviewer and a small ICC of 0.01 can multiply the variance of an estimate by a factor of 1.5, not including other sources of variance inflation. Therefore it is very important for SI to reduce the interviewer effects dramatically, and by a greater amount than shown in this paper (a 0.006 reduction in SILC; 0.025 in ICT). To be able to reach that goal, two steps must be taken: 1) Start to monitor the interviewers on a regular basis, since that would be a way to give them feedback about the how they conduct interviews and standardize the way they read the questions and probe for answers (Fowler & Mangione, 1990); 2) Some questions in the questionnaires will have to be rewritten in order to ensure that all respondents understand them in the same way and the interviewers do not feel they have to change them to suit the respondent or probe much to get an answer (see, for example, Converse & Presser, 1986). Groves and Magilavy (1986) also point out that specifying standard phrases for interviewers to use in the interview seems to reduce the effects of interviewers on the data.

4.b. Refusal rates

Effects of training on refusal rates

It seems that the special refusal training has had at least some positive effects in both surveys. In SILC, the refusal rate has dropped by four percentage points from 2008 to 2010. In the ICT, the refusal rates for 2009 and 2010 are the lowest for the survey. SI will continue to use, and expand upon, the same methods for refusal aversion training as it has done for 2009 and 2010. Also the interviewers have had access to their own personal refusal rate as well as other interviewer's rates in order to see what is acceptable and when their performance is substandard.

4.c. General discussion

Next steps based on the results of the paper

Taken together, the results of this paper show that the new training program for interviewers at Statistics Iceland has shown some promising results on the interviewer's performance in the ICT and SILC, especially in reducing refusals. It is clear, though, that some changes and refinements have to be made, like monitoring interviewers, being more thorough in writing standardized questions and possibly suggesting standard phrases for the interviewers to use while conducting interviews.

5. References

Collins, M. (1980). Interviewer variability: A review of the problem. *Journal of the Market Research Society, 22*, 77 – 95.

Converse, J.M. & Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire.* London, England: Sage.

Fowler, F.J, Jr. (1991). Reducing interviewer related error through interviewer training, supervision, and other means. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman (eds.), *Measurement Error in Surveys.* New York, NY: Wiley.

Fowler, F.J, Jr. & Mangione, T.W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error.* London, England: Sage.

Groves, R.M. (1989). *Survey error and Survey Cost.* New York, NY: Wiley.

Groves, R.M. & Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public opinion quarterly, 50*, 251 – 266.

Groves, R.M. & McGonagle, K.A. (2001). A theory guided training protocol regarding survey participation. *Journal of offical statistics, 17,* 249 – 265.

Groves, R.M. & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public opinion quarterly, 72*, 167 – 189.

Gwartney, P.A. (2007). *The Telephone Interviewer's Handbook: How to Conduct Standardized Conversations.* San Fransisco, Ca :Jossey-Bass

Kish, L. (1965). *Survey Sampling*. New York, NY: Wiley.

Mayer, T.S. & O'Brien, E. (2001, August). *Interviewer Refusal Aversion Training to Increase Survey Participation*. Paper presented at the annual meeting of the American Statistical Association. Retrieved 11. May, 2010, from: http://www.fcsm.gov/committees/ihsng/ASA2001.pdf

McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some Intraclass Correlation Coefficients. *Psychological Methods, 1*, 30 – 46.

Morton-Williams, J. (1993). *Interviewer Approaches.* Aldershot, England: Dartmouth.

Statistics Iceland (2010). *Labour market statistics 1991 – 2009.* Retrieved 12. May, 2010, from: *https://hagstofa.is/lisalib/getfile.aspx?ItemID=11081*

Sayles, H., Belli, R.F. & Serrano, E. (2010). Interviewer variance between event history calendar and conventional questionnaire interviews. *Public Opinion Quarterly, 74,* 140 – 153.

6. Appendix 1: Text of selected questions.

Question	English version
	SILC questions
Rent	How much was paid in rent for last month?
Rent support	How much government support to pay the rent did you get per month?
House fund	In addition to what is paid into house fund, how much was paid into maintenance fund for the last month?
Maintenance fund	In addition to what was paid into maintenance fund or house fund, how much money was used in maintenance or improvements of the house during the last 12 months in addition to what was paid from house fund or maintenance fund?
Rent idea	What is your idea about what would be paid in rent for a house like yours on the open market?
Lowest income	What is according to you the lowest income, after taxes that you need to have per month to make ends meet?
Lowest income guess	But if you had to give your best guess?
Alimony received	How much alimony did household members receive a month per child?
Alimony paid	How much alimony or maintenance did household members pay a month per child?
Financial support	How much regular financial support did the household give to someone in another home over the last year?
	ICT questions
TV	Does the household or anyone within the household have a TV
VCR	Does the household or anyone within the household have a VCR
Games console	Does the household or anyone within the household have a game console
Desktop	Does the household or anyone within the household have a desktop computer
Laptop	Does the household or anyone within the household have a laptop
Palmtop	Does the household or anyone within the household have a hand-held computer
DVD player	Does the household or anyone within the household have a DVD player, which is not within a computer or a game console.
MP3 player	Does the household or anyone within the household have an MP3 player, iPod or other digital player
TV record	Does the household or anyone within the household have a device that can record a television program and save it in digital format, for example a DVD player with a hard drive
Flatscreen TV	Does the household or anyone within the household have a flatscreen TV
Theater system	Does the household or anyone within the household have a home theater system
n of desktops	How many desktop computers are currently in use in the home?
n of laptops	How many laptop computers are currently in use in the home?
Satellite dish	Does the household have access to A Satellite dish
n of TV	How many television sets are there in your household?