Investigation of a New Interviewer Observation: Interviewer Assessments of Response Propensity

Jennifer Sinibaldi and Stephanie Eckman
Institute for Employment Research
Aleksa Möntmann-Hertz, LINK Institute

1 Introduction

Interviewer observations of respondent and neighborhood characteristics have attracted the attention of survey researchers, because such observations are rather inexpensive to collect and can be made on both responding and nonresponding cases. This study collected interviewer ratings of response propensity in a telephone survey, which has two advantages. First, cases can be more easily randomly assigned to interviewers, and thus we can separate out the effects of call, case and interviewer characteristics. Second, many of the more traditional interviewer observations, such as the condition of the house and neighborhood, are not possible over the phone. Thus developing a new interviewer observation for CATI would contribute to the small amount of paradata available for this mode.

The paper first explores how accurate the ratings are: that is, whether they correlate in the aggregate with the true completion rate of the cases. It then investigates whether interviewers differ systematically in how they rate cases. Finally a regression model identifies the call, case, and interviewer characteristics which influence the assigned ratings. The paper ends with thoughts about how interviewer ratings of response likelihood can be used in future surveys, and suggests areas for additional research with this interviewer observation.

2 Data

The data we use to address these questions come from a general population telephone survey conducted over nineteen days in January 2012 by the LINK Institute. At the end of each call where nonresponse was obtained, the interviewer rated the likelihood of the case to ever complete the survey on a scale from zero to 100, with no option to skip the question or answer "don't know." The text of this question, translated from German by the authors, was:

> How likely is it that this case will complete the interview at a later contact attempt? Please give the probability in percent, from 0 to 100.

Interviewers were not able to see the ratings assigned to the same case by other interviewers on previous calls, if any. The only information interviewers had about previous work on the case was the number of attempted calls, and whether or not a respondent had already been selected, and this information was not visible at the time of the rating. Interviewers received training on making these ratings as part of the usual project training.

All calls not resulting in cooperation, except those handled entirely by the autodialer (in which no contact with a live person was made - e.g., busy signals, answering machines, etc.), were rated by the interviewers. Within the 11,208 rated calls, this analysis focuses on the 6,892 ratings resulting from calls with contact, as interviewers are in better position

to make likelihood ratings when they have spoken with someone. We also have data on the time, date and outcome of each call placed on the survey and information about the 34 interviewers who participated in this study from an interviewer questionnaire.

3 Methods

To address our research questions about interviewer effects and the impact of call, case and interviewer characteristics, we use multilevel regression models to predict the response likelihood ratings. The first level in these models is the calls themselves, and the second is the interviewers. If the ratings demonstrate an interviewer effect, we can detect this as a random effect at the second level that is significantly different from zero. The calls are also grouped into cases, which are crossed with interviewers: cases are worked by
more than one interviewer, and interviewers work more than one case. However, because the analysis dataset contains only calls which resulted in contact but not cooperation, a majority of the cases (58%) appear only once in the analysis dataset, and thus estimation of a case-level random effect is inappropriate (Hox 1998).

The full model adds several independent variables: call level characteristics such as the sequence number of the call to the case (first call, second call, etc.) and the outcome of the call (refusal, appointment, etc.); case level characteristics such as whether the case had previously refused to participate, and whether the case later cooperated; and interviewer level characteristics.

4 Results

Before we begin to answer our research questions, we first explore the ratings overall to get a feel for the data. Figure 1 shows the distribution of the likelihood ratings assigned by the interviewers to the 6,892 calls resulting in contact.

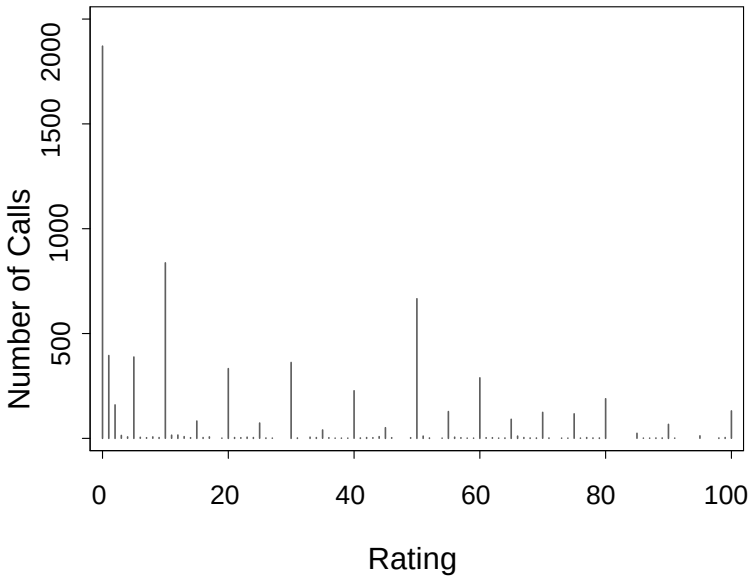

Figure 1: Distribution of Ratings

The modal rating is 0, and nearly all of these calls ended with a refusal. We see quite a bit of rounding: 73% of all ratings are on the tens, and 88% are on the tens and fives.

Do Likelihood Ratings Match Cooperation Rates?

To investigate the correlation of the likelihood ratings with the true cooperation rates, we calculate the average rating for each case across all calls resulting in contact, and group this average into bins by ten. Within each bin, we calculate the percent of cases that completed the survey by the end of the field period. Figure 2 shows the results. As discussed above, the cases with an average rating of 0 were largely unproductive: fewer than three percent completed the interview. The percent of cases within each grouping that responded to the interview increases steadily from the (0,10] group to the (80,90] group.

Figure 2 shows broad agreement between the ratings and the observed rate of completion. However, it also demonstrates that the ratings cannot be interpreted literally. In the (50,60] group, only 19.1% of the cases cooperated, a rate quite a bit lower than that suggested by the average rating.

Is There Evidence of Interviewer Effects in Ratings?

Because of the near-random assignment of cases to interviewers, all interviewers should give the same ratings, on average. However, in the multilevel regression model without any independent variables, the intracluster correlation coefficient for interviewers is greater than zero (ρ int = 9.2%) and the random effect is significant (σ^2= 71.9, SE = 18.6). This result suggests that different interviewers do give different ratings.

Interviewer effects in this rating should not be surprising, given the ubiquity of such effects in other stages of the survey process, such as coding (Campanelli et al. 1997), respondent recruitment (O'Muircheartaigh and Campanelli 1999, West and Olson 2010), response collection (O'Muircheartaigh and Marckward 1980, Schnell and Kreuter 2005), and frame creation (Eckman 2012).

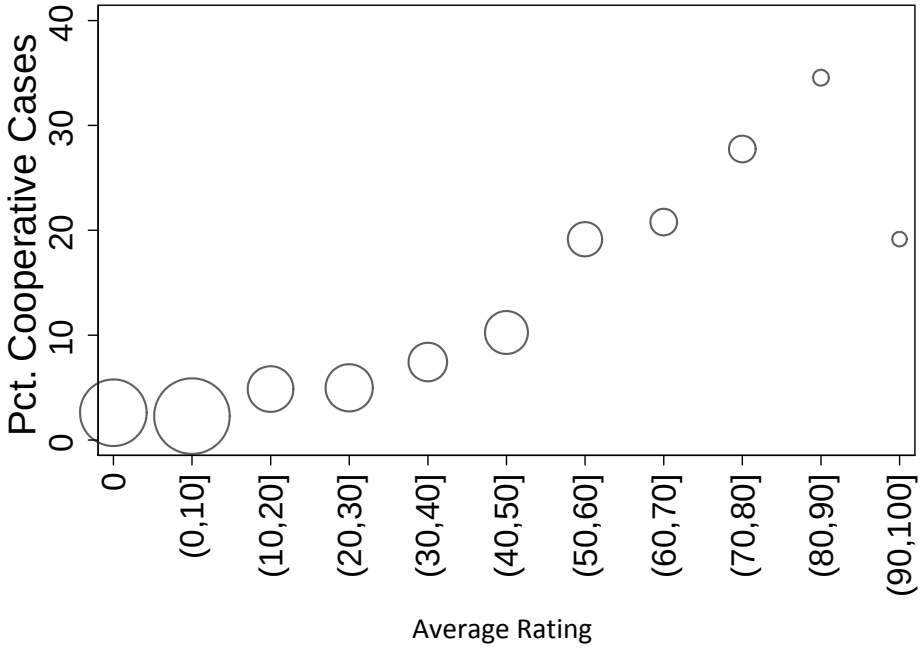

Figure 2: Completion Rates by Average Rating, Size of Circle Proportional to Number of Cases

The finding of variability due to interviewers in the likelihood rating raises a concern for those who wish to use such ratings in nonresponse adjustment. The random error in these ratings is not routinely captured in the adjusted weights. We note, however, that other methods of response propensity estimation are subject to their own variability, such as modeling error, which is routinely ignored.

What Factors Affect Willingness Ratings?

The model shown in Table 1 adds several explanatory variables. The call outcome correlates as expected with the assigned rating: an appointment is associated with a 5.1 point higher rating, on a scale from 0 to 100, and a refusal with a 40.6 point lower rating, relative to calls with other outcomes. Additional calls to a case, and reaching the selected target respondent, however, have no significant effect on the rating. Cases which had a prior refusal on a previous call were given lower ratings ($\hat{\beta} = -14.2$), and those that did eventually cooperate were given ratings that were 5.6 points higher, indicating, as in Figure 2, a weak relationship between the true response likelihood and the interviewers' ratings. These results are sensible and in the expected directions: interviewers should use these call and case characteristics when assigning likelihood ratings.

In the next section of the table are the independent variables relating to the interviewers. If all interviewers assign the same ratings, none of these variables should be significantly related to the rating. Instead, we see that interviewers who think that persuasion of respondents who initially refuse is a good idea give ratings that are 5.1 points higher, and those who think it is not sensible to recontact refusers give ratings that are four points lower. These two expectation variables show that interviewers with more optimism about refusal conversion give higher ratings.

The interviewers' yield rate (the fraction of all calls leading to an interview) has no significant effect on the rating, and neither does the number of months of interviewing experience. Each additional call rated by the interviewers has a very small, but significant, negative effect on the rating ($\hat{\beta} = -0.005$), as if interviewers become more pessimistic with each attempted call. Note that the regression also controls for the number of calls to the case, so this estimated coefficient should capture only the effect of repeated ratings on the interviewer herself, not the changing case base over the course of data collection.

Surprisingly, the outcome of the previous call does correlate with the rating – interviewers whose previous call ended in a refusal or a complete, give ratings that are 1.5 and 1.4 points lower than those whose previous call ended with another outcome. It seems as if a refusal on the prior call creates some pessimism about the current case. But a complete on prior call, followed by some sort of non-complete on the current call also creates pessimism.

5 Discussion

Interviewers form expectations about the response likelihood of cases assigned to them, based whatever information they have at hand. In a telephone survey, they might review the notes from previous calls to tailor their introduction. In a face-to-face survey, they can observe much more about the case from the condition of the housing unit and neighborhood. These expectations are correlated with the true response likelihood of the cases, but are also subject to random variation across interviewers and to the effects of interviewer attitudes and prior experiences, and for this reason may not be suitable for nonresponse adjustment.

This paper has begun to explore how interviewers make these judgments of response likelihood. This research is exploratory and more investigation is needed. Future research should investigate whether they can play a roll in directing field work effort, along the lines suggested by the responsive design framework (Groves and Heeringa 2006). Because the ratings are almost costless to do, we hope that more studies in both telephone and

face-to-face modes will collect and analyze them.

References

Campanelli, P. C., K. Thomson, N. Moon, and T. Staples (1997). The Quality of Occupational Coding in the UK. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, and C. Dippo (Eds.), *Survey Measurement and Process Quality*, pp. 437–457. Wiley-Interscience.

Eckman, S. (2012). Do Different Listers Make the Same Housing Unit Frame? Variability in Housing Unit Listing. *Under Review*.

Groves, R. M. and S. G. Heeringa (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(3), 439–457.

Hox, J. J. (1998). Multilevel Modeling: When and Why. In I. Balderjahn, R. Mathar, and M. Schader (Eds.), *Classification, Data Analysis, and Data Highways*, pp. 147–154. New York: Springer-Verlag.

O'Muircheartaigh, C. and P. Campanelli (1999). A Multilevel Exploration of the Role of Interviewers in Survey Non-Response. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 437–446.

O'Muircheartaigh, C. and A. M. Marckward (1980). An Assessment of the Reliability of World Fertility Study Data. *Proceeedings of the World Fertility Survey Conference 3*, 305–379.

Schnell, R. and F. Kreuter (2005). Separating Interviewer and Sampling-Point Effects. *Journal of Official Statistics 21*(3), 389.

West, B. and K. Olson (2010). How Much of Interviewer Variance is Really Nonresponse Error Variance? *Public Opinion Quarterly 74*(5), 1027–1045.

Table 1: Influences of Call, Case and Interviewer Characteristics on Interviewer Ratings of Response Likelihood

Dependent Variable: Likelihood Rating [0-100]	$\hat{\beta}$ (SE)
Call & Case Characteristics	
Call ended in other contact	*reference*
Call ended in appointment	5.072*
	(0.68041)
Call ended in refusal	-40.60*
	(0.56403)
Cumulative number of calls to case[a]	-0.0832
	(0.12927)
Contact with another person	*reference*
Contact with target person	0.754
	(1.09705)
Case had previous refusal	-14.23*
	(0.68987)
Case ended as complete	5.626*
	(0.78654)
Interviewer Characteristics	
Refusals should always be persuaded	5.079*
	(2.39705)
Even hardest refusals can be persuaded	-0.887
	(2.24336)
All respondents can be persuaded	-4.056*
	(2.04920)
Fraction of calls leading to interview	-1.034
	(0.76215)
Months of interviewing experience	0.0103
	(0.01755)
Cumulative number of rated calls, with contact	-0.00470*
	(0.00185)
Interviewer's prior call was a refusal	-1.453*
	(0.44297)
Interviewer's prior call was a complete	-1.440+
	(0.85649)
Random Effect: σ^2 Interviewers	32.92*
	(9.15)
N	6883
N cases	4544
N interviewers	34
ρ Interviewers	0.124

Standard errors in parentheses

+ $p < 0.10$, * $p < 0.05$

Estimates of constant not displayed

[a] Calculated on larger dataset of all rated calls (n=11,208)