Can Interviewers Effectively Rate the Likelihood of Cases to Cooperate?

Stephanie Eckman (corresponding author)

Institute for Employment Research

Weddigenstraße 20-22

Nuremberg, Germany 90478

+49 911 179 3165

stephanie.eckman@iab.de

Jennifer Sinibaldi

Institute for Employment Research

Weddigenstraße 20-22

Nuremberg, Germany 90478

+49 911 179 8594

jennifer.sinibaldi@iab.de

Aleksa Möntmann-Hertz

LINK Institute for Market and Social Research

Burgstraße 106

Frankfurt, Germany 60389

+49 69 945 40128

moentmann.aleksa@link-institut.de

Stephanie Eckman, Jennifer Sinibaldi, Aleksa Möntmann-Hertz, Can Interviewers Effectively Rate the Likelihood of Cases to Cooperate?, *Public Opinion Quarterly*, Volume 77, Issue 2, Summer 2013, Pages 561–573, https://doi.org/10.1093/poq/nft012

AUTHOR NOTE:

STEPHANIE ECKMAN is a Senior Researcher at the Institute for Employment Research (IAB) in Nuremberg, Germany. JENNIFER SINIBALDI is a researcher and GradAB scholar at the Institute for Employment Research (IAB) in Nuremberg, Germany. ALEKSA MÖNTMANN-HERTZ is a Senior Project Manager with the LINK Institute in Frankfurt, Germany. The authors would like to thank Susan Kenney McCollouch, Frauke Kreuter, and three reviewers for helpful comments.

*Address correspondence to Stephanie Eckman, Institute for Employment Research, Weddigenstraße 20-22, Nuremberg, Germany 90478, +49 911 179 3165, stephanie.eckman@iab.de.

Abstract

This paper explores how well interviewers can judge which cases are likely to cooperate with a survey request and which are unlikely. Interviewers in a telephone survey rated the response likelihood of each case after every call they placed, on a scale from zero to 100, where zero meant that the case would never cooperate and 100 meant that it certainly would. Analyses of the ratings reveal that they do correlate with the cooperation rate among the cases, and that the likelihood ratings are influenced in the expected directions by case and call level attributes. However, there are also strong interviewer effects. The ratings show promise for use in responsive design procedures.

Introduction

This study investigates the viability and usefulness of a new type of interviewer observation: interviewer ratings of how likely each case is to cooperate. Interviewers likely already form impressions about the response likelihood of the cases assigned to them (Campanelli, Sturgis, and Purdon 1997, 4–39). In this study, we capture and analyze these expectations.

There are at least two reasons why survey researchers might be interested in such observations. First, interviewer expectations of response likelihood might affect interviewer behavior during recruitment. There is evidence that interviewers who think recruiting cases for a given survey will be easy have higher response rates (Singer, Frankel, and Glassman 1983, Hox and de Leeuw 2002). Kennickell (2012) takes this line of reasoning one step further and argues that when interviewers have control over the recruitment process, their expectations about how likely each case is to respond can introduce nonresponse bias: Because interviewers are pressured to reach target response rates, they recruit cases that seem like cooperators, that is, that are similar to the cases that have already completed the survey. Thus the field should better understand how interviewers form expectations about which cases are more likely to cooperate. Second, if interviewer ratings of response likelihood are reasonably accurate, they may be useful for targeting cases in responsive design protocols that aim to reduce data collection costs and decrease bias (Groves and Heeringa 2006), especially in telephone surveys where few case-level characteristics are available for the calculation of response propensity models.

Only a few surveys have directly asked interviewers to estimate cases' response likelihoods. The 2004 Health and Retirement Survey asked interviewers to make weekly ratings using a three-point scale: low, medium, and high (Wagner and Guyer 2005). The 2010 Survey of Consumer Finances had interviewers rate their cases on a five-point scale of likelihood to participate (Kennickell 2012). Cycle 7 of the National Survey of Family Growth rated all incomplete cases before nonresponse subsampling in each quarter of data collection (J. Wagner, personal communication, January 7, 2013). The Dioxin Exposure study collected a four-point rating of resistance after each contact in a nonresponse follow-up study (Olson, Sinibaldi, and Lepkowski 2006). The current wave of the National Survey of Sexual Attitudes and Lifestyles in the U.K. asked interviewers to score each case on a five-point willingness scale after every contact, though no external documentation yet exists.

All of these prior studies that used interviewer ratings of response likelihood were face-to-face surveys. We have chosen to study interviewer ratings of response likelihood in a telephone survey, for two reasons. First, in the telephone survey discussed here, cases were randomly assigned to interviewers, and thus we can separate the effects of call, case and interviewer characteristics. Such interpenetrated designs are prohibitively expensive in face-to-face surveys. Second, many of the more traditional interviewer observations, such as the condition of the house and neighborhood, are not possible over the phone. Thus likelihood ratings are likely to be of greater practical use in telephone than in face-to-face surveys.

This study is the first, to our knowledge, to investigate such ratings in a telephone study. The paper first explores how accurate the ratings are—that is, whether they correlate in the aggregate with the cooperation 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 that 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.

Data

The data we use to address these questions come from a telephone survey conducted over 19 days in January 2012 by the LINK Institute in Germany. At the end of each contact attempt, the interviewer rated the likelihood of the selected target person at that household to complete the survey on a later call, using a scale from zero to 100. The text of this question, translated from German, 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 permitted to skip the question or to respond “don’t know.” All calls not resulting in cooperation (that is, an interview), except those handled entirely by the autodialer, were rated by the interviewers.

In total, 34 interviewers rated 11,251 calls. For each call, we also have the time, date, outcome, whether the selected respondent was reached, and the interviewer ID. However, in many of the rated calls, the interviewer reached a fax machine or an answering machine, or the person who answered the phone hung up without speaking to the interviewer. Because interviewers in these situations had little information with which to make the likelihood ratings, the analyses in this paper focus only on the 6,892 ratings resulting from calls with contact.[1] For a brief discussion of all the ratings and more details on the survey itself, please see Appendix 1.

Interviewers were not able to see the ratings assigned to the same case by other interviewers on previous calls, if any. While the interviewing software did display the number of prior calls, and whether or not a respondent had already been selected, this information was not visible at the time of the rating. All interviewers received general training as well as survey-specific training that included practice with this question.

In compliance with German law, refusal conversion in this study was limited. Those cases where a respondent was not yet selected but a contact person indicated that he or she was not interested in topic, had no time, was not in the mood, or simply hung up without saying anything, were called one more time after a few days. Cases where the selected respondent indicated that he or she had no time or were not in the mood were also called once more. Harder refusals were not recontacted. Of the 3,579 cases that refused at one point, 98 eventually cooperated.

In the LINK telephone studio, cases were assigned to whichever interviewer was available. There were no interviewers who specialized in refusal conversion or who worked more difficult cases. Thus we can assume that the assignment of cases to interviewers in this study was nearly random, setting aside differences in the shifts interviewers work and the kinds of cases called in each shift. Such an interpenetrated design is ideal for studying interviewer effects (Mahalanobis 1946).

In addition to the call record data, we have data about the 34 interviewers who participated in this study. The interviewer questionnaire contained questions about interviewing experience and opinions about refusal conversion. These items are used to explore how interviewers' characteristics correlate with the response likelihood ratings they give.

Analyses

[1] The dataset includes 43 calls to 43 cases that ended in hard refusals and were not rated. We have imputed 0 ratings to these calls and included them in our analyses.

To investigate the correlation of the likelihood ratings with cooperation rates,[2] we calculate the average rating for each case across all calls resulting in contact. We then categorize cases by their average rating into bins: 0, (0–10], (10–20], etc. Within each bin, we calculate the percent of cases that completed the survey by the end of the field period.

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. The calls are also grouped into cases, which are crossed with interviewers. However, because the analysis dataset contains only calls that 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).

We first estimate the empty model, containing no independent variables, to examine the effect of the clustering of calls within interviewers. The model is:

$$\text{Rating}_{ij} = \alpha_j + \varepsilon_{ij}$$

$$\alpha_j \sim N(\alpha, \sigma^2_{\text{int}})$$

where *i* indexes the calls and *j* the interviewers. The intraclass correlation coefficient (ICC) in this model is the ratio of the variance in the interviewer intercepts (σ^2_{int}) to the total variance in the ratings. An ICC that is significantly different than zero indicates interviewer effects in the ratings.[3]

The full model adds independent variables for call, case and interviewer characteristics. For each call, the model includes 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.). An interviewer's experience on the call placed just prior may also affect how he or she rates the current call; therefore indicators of these outcomes are included in the model.[4] Case characteristics, such as whether the case had previously refused to participate, and whether the case ever cooperated are also included. (A list of all variables considered as predictors is given in Appendix 2.)

[2] By cooperation rate, we simply mean the fraction of cases in our analysis dataset completing the study. Because our dataset includes only contacted cases which did not complete on the first call, we do not use an AAPOR cooperation rate.

[3] We evaluate the significance of the ICC using the method developed by Hedges, Hedberg, and Kuyper (2012).

[4] The inclusion of these lagged variables causes the case-base to decrease slightly from 6,892 to 6,883. We use this smaller dataset to estimate both models.

Interviewer characteristics come from the interviewer questionnaire, which collected demographic information and opinions. Eight questions captured interviewers' expectations about refusal conversion, all using a four-point response scale from "strongly agree" to "strongly disagree." Those who expect refusal conversion to be easier may give higher ratings to cases ending in refusal. To test this in the model, we include a factor score summarizing the responses to all eight questions. The factor ranges from −1.72 to 2.07, with a mean across the 34 interviewers of zero. See Appendix 3 for details on the development of the factor score. Because more successful interviewers may be more likely to give higher ratings, the models include the percent of all calls handled by an interviewer that resulted in cooperation. Also included is the number of months of interviewing experience, which ranged in this study from 13 to 300.[5]

Results

Before we begin to answer our research questions, we first explore the ratings overall. Figure 1 shows the distribution of the likelihood ratings assigned by the interviewers to the 6,892 calls resulting in contact. The mean rating was 24.8 and the median was 10. The modal rating is zero, and nearly all of the calls rated as zero ended with a refusal. We see quite a bit of rounding: 73% of all ratings are multiples of ten, and 88% are multiples of five.

[Figure 1 about here.]

Do Likelihood Ratings Match Cooperation Rates?

Figure 2 displays the relationship between the average likelihood rating by case, in bins by ten, and the cooperation rate in that bin. The first point on the left in this graph shows the cases with an average rating of zero. There are 961 such cases, and fewer than three percent completed the interview. The percent of cases within each bin which eventually cooperated increases steadily from the (0,10] to the (80,90] bin. The cooperation rate in the last bin is lower, but the group size is also quite small.

[Figure 2 about here.]

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, for

[5] We considered alternative models such as the poisson and negative binomial, which did not change the substantive conclusions. We also assssed model diagnostics and non-linear relationships between the predictor variables and the ratings (results not shown).

example, only 19.1% of the cases cooperated, a rate quite a bit lower than that suggested by the average of the interviewer assigned ratings.

Is There Evidence of Interviewer Effects in Ratings?

Interviewer variance may arise if different interviewers use different methods to make the ratings. We find evidence for an interviewer effect in the multilevel regression model without any independent variables. The ICC for interviewers is 9.2% and significant (SE=2.2 percentage points). Interviewer effects in this rating should not be surprising, given their ubiquity in other stages of the survey process such as coding, response collection, respondent recruitment, and frame creation (O'Muircheartaigh and Marckward 1980, Campanelli, Sturgis, and Purdon 1997, O'Muircheartaigh and Campanelli 1999, Schnell and Kreuter 2005, West and Olson 2010, Eckman 2013,).

What Factors Affect Willingness Ratings?

The model shown in Table 1 adds several explanatory variables to discern the call, case and interviewer characteristics that influenced the ratings that the interviewers assigned. The ICC in this model is still significant (ρ_{int}= 12.9%, SE = 3.0 percentage points), indicating that even after controlling for these characteristics, unexplained variance due to the interviewers remains.

The call outcome correlates as expected with the assigned rating: an appointment is associated with a 6.4-point higher rating, on a scale from zero to 100, and a refusal with a 40.3-point lower rating, relative to contacts with other outcomes. Additional calls, and reaching the selected target respondent, however, have no significant effect on the rating.

Contacts with cases that refused on a previous call were given lower ratings ($\hat{\beta}= -13.6$). As shown in Figure 2, contacts with those cases that did eventually cooperate were given ratings that were 5.6 points higher, indicating a relationship—though not a strong one—between the true response likelihood and the interviewers' ratings. These results are sensible and in the expected directions.

Each additional call rated by the interviewers has a very small, but significant, negative effect on the rating ($\hat{\beta}= -0.0044$), as if interviewers become more pessimistic with each additional call that they place. Note that the model 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. The outcome of the previous call also correlates with the rating—

interviewers whose previous call ended in a refusal give ratings that are 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.

[Table 1 about here.]

In the next section of the table are the independent variables relating to the interviewers. The higher an interviewer's score on the factor relating to optimism about refusal conversion, the lower the rating she assigns to calls with non-refusal outcomes ($\hat{\beta}= -6.9$), but the higher the rating she assigns to calls with refusal outcomes ($\hat{\beta}= 6.6$). While a refusal outcome on a call has a large negative influence on all interviewers, those who believe refusal conversion is more appropriate tend to give higher ratings to refusal calls. Interviewers with higher success rates (the fraction of all calls leading to an interview) do not give different ratings, and neither do those with more experience. Finally, interviewer demographics were not significant, indicating that age, gender, income and other attributes do not correlate with the ratings.

Discussion

There is much more information passed between interviewers and respondents than is contained in the final case dispositions and call record data. In a telephone survey, interviewers can listen to a respondent's tone, or hear something in the background, to judge whether "I'm too busy" or "Call another time" means that another call might be successful or is really a polite way to indicate a refusal. This study has captured and summarized such information by asking interviewers to rate how likely each case is to eventually cooperate. Interviewers' ratings are correlated with cooperation. Call, case and interviewer characteristics correlate with the ratings in largely expected directions. However, the ratings are also subject to random variation across interviewers. Additional experience in making the ratings, or further training, may be able to reduce this variability. Statistical methods could also be used to reduce the influence of interviewer variability on the ratings. Random effects models, such as the model we fit above, can produce empirical best linear unbiased predictions of individual interviewers' random effects on the ratings (West, Welch, and Galecki 2007, 45) and these could be removed from the ratings before they are used in non-response adjustments.

The investigations presented here are exploratory and more research is needed. To improve the quality of the ratings, we also suggest that future surveys test alternative scales, such as reducing the scale to 5 or 10 points, and conduct debriefings with the interviewers to better understand how they make the ratings. We also encourage testing whether interviewer perceptions of response likelihood can influence interviewer behavior during recruitment, as Kennickell (2012) argues, and whether the process of making the observations changes interviewers' perceptions, as suggested by some of the findings in our model. Finally, future studies should explore whether these ratings can play a role in directing field work efforts, along the lines suggested by the responsive design framework (Groves and Heeringa 2006). For example, effort could be directed towards cases with high average (or recent) likelihood ratings to reduce data collection costs. We encourage future studies of likelihood ratings in both telephone and face-to-face modes.

Appendix 1: Survey Details

The topic of the survey was the use of computers, mobile phones and other devices for reading online and digital text. The gross sample of 9,175 German landline telephone numbers was nationally representative and geographically stratified. The target population was all adults living in Germany. Mobile phone numbers were excluded from the frame due to the costs of dialing such numbers, and thus adults without landline telephones were undercovered in this study. After contacting a household, the interviewer used a Kish algorithm to select one person to participate in the survey (Kish 1965, Section 11.3B). The interview lasted just over seven minutes, and 1,002 interviews were completed. The AAPOR RR1 response rate was 13% (American Association for Public Opinion Research 2011).

Over 28,000 calls were placed to the selected numbers and 11,251 calls were rated. The analysis in the body of the paper uses only likelihood ratings assigned on the 6,892 calls resulting in contact, because the ratings are more meaningful if the interviewer spoke to someone. For completeness, Figure 3 shows all of the ratings assigned in this study (n=11,251).

[Figure 3 about here.]

The mean rating was 29.4 and the median was 20. The modal rating was zero, 25.2% of all ratings. Many of these calls resulted in fax machines (27.9% of all calls with zero ratings), followed by calls where the respondent hung up during the introduction (17.5% of these calls). Only 3.4% of these calls resulted in contact with a selected respondent. The second most commonly used rating in the study was 50: 23.6% of all ratings have this score. Interviewers seem to use this score to indicate “don’t know,” as previous research with probability scales has shown (Fischhoff and Bruine de Bruin 1999). The majority of the calls given ratings of 50 (65.6%) ended with answering machines. Only 347 calls (3.1% of all calls rated) received likelihood ratings of 100. (Remember that calls resulting in completed interviews were not rated.) Again, most of these were answering machines (62.0%). Hard appointments were also common in this group: 11.2% of all calls with ratings of 100.

Appendix 2: Available Variables for Random Effects Model

Below is a complete list of the variables we considered as predictors in the model shown in Table 1. The final model retains the variables that were significant as well as those that we felt were interesting because they were not significant.

Call level: Call ended in other contact; Call ended in appointment; Call ended in refusal; Cumulative number of calls to case; Contact with target person; Case had previous refusal; Interviewer's prior call ended in a refusal; Interviewer's prior call ended in a complete; Interviewer's prior call ended in an appointment; Interviewer's prior call ended in other contact; Day of week of call; Time of call; Whether interviewer had called this case before.

Case level: Case cooperated; Case ever refused; Case ended as refusal.

Interviewer level: Factor 1 score; Factor 2 score; Refusal interacted with Factor 1 score; Refusal interacted with Factor 2 score; Fraction of calls leading to interview; Months of interviewing experience; Cumulative number of calls rated by interviewer; Interviewer gender; Interviewer age; Interviewer nationality (German vs. non-German); Big 5 neuroticism; Big 5 openness; Big 5 extroversion; Big 5 agreeableness; Big 5 conscientiousness; Interviewer's family receives income support; Interviewer is currently a student; Importance of pay as interviewer; Satisfaction with pay as interviewer.

Appendix 3: Factor Analysis

The interviewer questionnaire contained a section of eight items relating to interviewer beliefs and opinions regarding respondent recruitment and refusal conversion. The statements were introduced with the statement:

> Sample persons have different reactions to the request to participate in a study: Some agree easily, others hesitate or refuse immediately. In the following statements, please tell us your opinion as an experienced interviewer.

The wording of the eight statements is given in Table 2.[6] The response scale for each statement ranged from 1 (strongly agree) to 4 (strongly disagree). Two interviewers responded “don’t know” to one of the eight statements, for a total of two missing values. Modal imputation was used to fill in these responses.

The eight variables could not all be included in the regression model due to multicollinearity. Instead, the model contains the first factor from a factor analysis of all eight responses. The factor analysis was run in Stata using principal factor extraction and orthogonal rotation (StataCorp 2011a,b). Based on the eigenvalues from the factor analysis, we retained two factors. The first captures 55% of the total variance in the eight items and the second an additional 31%. Table 2 gives the mean response given by the interviewers and the loadings on the two factors. Cronbach’s alpha for all 8 items is 0.53. For the three items loading mostly strongly on factor 1, it is 0.68, and for those 4 items loading on factor 2, it is 0.49.

Factor 1 captures interviewers’ optimistic attitudes towards refusal conversion efforts: Questions 1, 2 and 3 load most strongly on this factor. Factor 2 is less clear but seems to represent respect for the respondent: questions 4, 5, 6 and 7 load on this factor. The eighth question does not load onto either factor. Only the first factor was significant in our model predicting the likelihood ratings.

[Table 2 about here.]

[6] Translation from German into English by Blom and Korbmacher (2013) and the authors. The text of the introduction and statements was based on the interviewer questionnaire developed by attendees of the International Workshop on Household Survey Nonresponse in Nuremberg, Germany in 2010.

References

American Association for Public Opinion Research. 2011. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* 7th ed. http://www.aapor.org/Resources.htm.

Blom, Annelies and Julie Korbmacher. 2013. “Measuring Interviewer Characteristics Pertinent to Social Surveys: A Conceptual Framework.” *Survey Methods: Insights from the Field*. http://surveyinsights.org/?p=817

Campanelli, Pamela C., Patricia Sturgis, and Susan Purdon. 1997. *Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates*. London: National Centre for Social Research.

Campanelli, Pamela C., Katarina Thomson, Nick Moon, and Tessa Staples. 1997. “The Quality of Occupational Coding in the UK.” In *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin, 437–457. New York: Wiley.

Eckman, Stephanie. 2013. “Do Different Listers Make the Same Housing Unit Frame? Variability in Housing Unit Listing.” Forthcoming in *Journal of Official Statistics*.

Fischhoff, Baruch and Wändi Bruine de Bruin. 1999. “Fifty-Fifty = 50%?” *Journal of Behavioral Decision Making* 12:149–163.

Groves, Robert, and Steven 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)* 169:439–457.

Hedges, Larry V., Eric C. Hedberg, and Arend M. Kuyper. 2012. “The Variance of Intraclass Correlations in Three- and Four-Level Models.” *Educational and Psychological Measurement* 72:893–909.

Hox, Joop. 1998. “Multilevel Modeling: When and Why,” in *Classification, Data Analysis, and Data Highways*, edited by Ingo Balderjahn, Rudolph Mathar, and Martin Schader, 147–154. New York: Springer-Verlag.

Hox, Joop, and Edith de Leeuw. 2002. “The Influence of Interviewers’ Attitude and Behavior on Household Survey Nonresponse: An International Comparison.” In *Survey Nonresponse*, edited by Robert Groves, Don Dillman, John Eltinge, and Roderick Little, 103–118. New York: Wiley.

Kennickell, Arthur B. 2012. What’s the Chance? Interviewers’ Expectations of Response in the 2010 SCF. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association.

Kish, Leslie. 1965. *Survey Sampling.* New York: Wiley.

Mahalanobis, Prasanta. 1946. “Recent Experiments in Statistical Sampling in the Indian Statistical Institute.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 109:325–370.

Olson, Kristen, Jennifer Sinibaldi, and James Lepkowski. 2006. “Analysis of a New Form of Contact Observations: How Does it Compare to the Traditional?” Paper presented at the Midwestern Association for Public Opinion Research Conference, Chicago, IL.

O’Muircheartaigh, Colm and Pamela 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)* 162:437–446.

O’Muircheartaigh, Colm and Albert Marckward. 1980. “An Assessment of the Reliability of World Fertility Study Data.” *Proceedings of the World Fertility Survey Conference Record of Proceedings* 3:305–379.

Schnell, Rainer and Frauke Kreuter. 2005. “Separating Interviewer and Sampling-Point Effects.” *Journal of Official Statistics* 21:389–410.

Singer, Eleanor, Martin Frankel, and Marc Glassman. 1983. “The Effect of Interviewer Characteristics and Expectations on Response.” *Public Opinion Quarterly* 47:68–83.

StataCorp. 2011a. *Stata 12 Multivariate Statistics Reference Manual.* College Station, TX: StataCorp LP.

StataCorp. 2011b. *Stata Statistical Software: Release 12.* College Station, TX: StataCorp LP.

Wagner, James and Heidi Guyer. 2005. “Statistical Propensity Models to Predict Likelihood of Survey Response Compared to Interviewer Judgments of Likelihood of Response.” Paper presented at the Midwestern Association for Public Opinion Research Conference, Chicago, IL.

West, Brady, and Kristen Olson. 2010. "How Much of Interviewer Variance is Really Nonresponse Error Variance?" *Public Opinion Quarterly* 74:1027–1045.

West, Brady, Kathleen Welch, and Andrzej Galecki. 2007. *Linear Mixed Models*. Boca Raton: Chapman & Hall/CRC.

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	6.42* (0.677)
Call ended in refusal	−40.32* (0.557)
Cumulative number of calls to case[a]	−0.047 (0.127)
Contact with another person	*reference*
Contact with target person	0.67 (1.081)
Case had previous refusal	−13.59* (0.683)
Case ended as complete	5.60* (0.775)
Cumulative number of calls rated by interviewer	−0.0044* (0.00182)
Interviewer's prior call ended in other contact[a]	*reference*
Interviewer's prior call ended in refusal[a]	−1.42* (0.443)
Interviewer's prior call ended in complete[a]	−1.33 (0.847)
Interviewer's prior call ended in appointment[a]	−0.34 (0.771)
Interviewer Characteristics	
Factor 1 Score: Positive attitude towards refusal conversion	−6.94* (1.238)
Refusal interacted with Factor 1 score	6.62* (0.456)
Fraction of calls leading to interview[a]	−0.51 (0.757)
Months of interviewing experience	0.0053 (0.0177)
Random Effect: σ Interviewers	5.82* (0.785)
ρ Interviewers	0.129 (0.030)
N	6883
N cases	4544
N interviewers	34
Pseudo-R^2	0.120

Standard errors in parentheses

* $p < 0.05$

Estimates of constant not displayedz

[a] Calculated on larger dataset of all rated calls (n=11,251)

Table 2: Factor Analysis of Items Relating to Interviewers' Attitudes Towards Cooperation and Refusal Conversion

			Loadings	
	Statement[a]	Mean[b]	Factor 1	Factor 2
1	Reluctant respondents should always be persuaded to participate	2.03	0.7962	−0.0160
2	With enough effort, even the most reluctant respondent can be persuaded to participate	2.46	0.8022	0.0621
3	It does not make sense to contact reluctant target persons repeatedly	2.40	−0.4055	0.0203
4	An interviewer should respect the privacy of the respondent	1.51	−0.1605	0.4787
5	If a respondent is reluctant, a refusal should be accepted	2.14	−0.1201	0.3491
6	One should always emphasize the voluntary nature of participation	2.00	0.2012	0.4957
7	If you catch them at the right time, most people will agree to participate[c]	1.89	0.1093	0.5324
8	Respondents who were persuaded after great effort do not provide reliable answers[c]	2.94	−0.2034	−0.0462

[a] Translation from German into English by Blom and Korbmacher (2013) and the authors

[b] Response scale from 1 to 4, where 1 meant strongly agree and 4 meant strongly disagree

[c] One missing value prior to imputation

Order of questions in the table does not reflect the order in the questionnaire

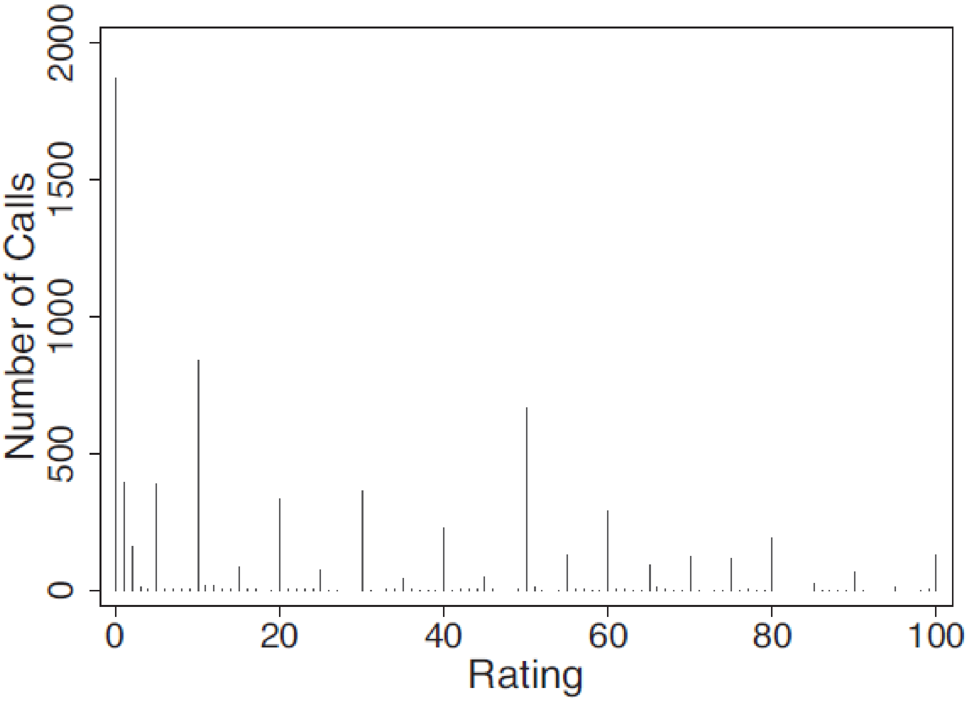

Figure 1: Distribution of Ratings on Calls Ending in Contact.

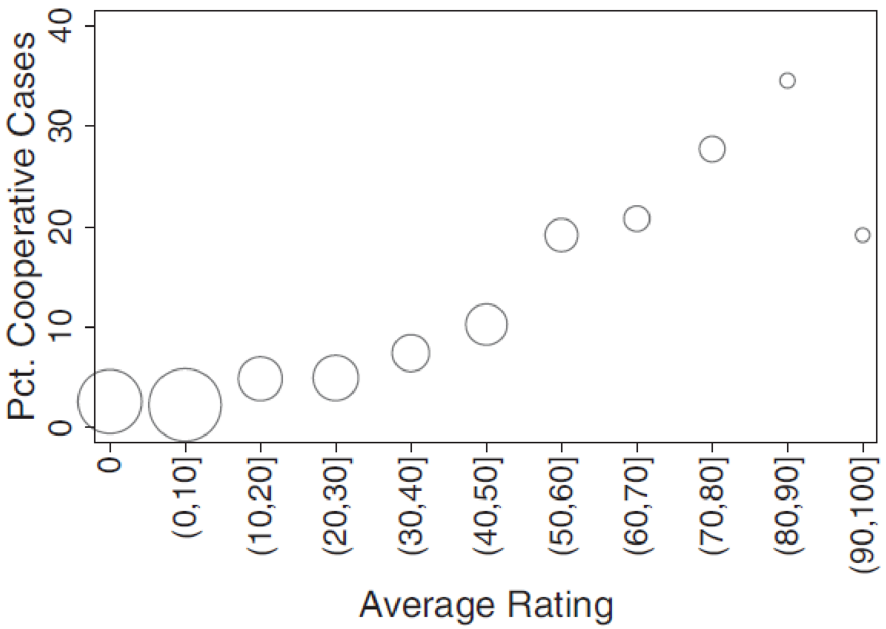

Figure 2: Cooperation Rates by Average Rating, Size of Circle Proportinal to Number of Cases

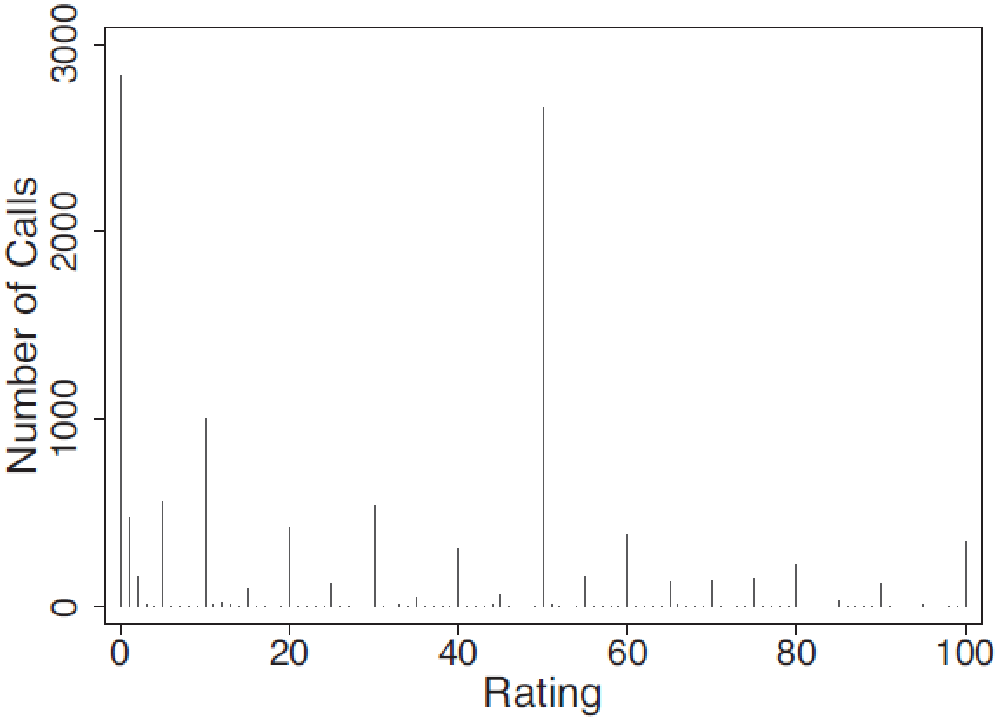

Figure 3: Distribution of All Ratings