10-13-2021

Examining Interviewers' Ratings of Respondents' Health: Does Location in the Survey Matter for Interviewers' Evaluations of Respondents?

Dana Garbarski
Loyola University Chicago, dgarbarski@luc.edu

Nora Cate Schaeffer
University of Wisconsin-Madison

Jennifer Dykema
University of Wisconsin-Madison

Author Manuscript

This is a pre-publication author manuscript of the final, published article.

Recommended Citation

Garbarski, Dana; Schaeffer, Nora Cate; and Dykema, Jennifer. Examining Interviewers' Ratings of Respondents' Health: Does Location in the Survey Matter for Interviewers' Evaluations of Respondents?. Journal of Survey Statistics and Methodology, 10, 2: 276–298, 2021. Retrieved from Loyola eCommons, Sociology: Faculty Publications and Other Works, http://dx.doi.org/10.1093/jssam/smab025

Examining Interviewers' Ratings of Respondents' Health:
Does Location in the Survey Matter for Interviewers' Evaluations of Respondents?

Dana Garbarski, PhD[1*]
Phone: 1(773) 508-3445, Fax: 1(773)508-7099, Email: dgarbarski@luc.edu

Nora Cate Schaeffer, PhD[2]
Phone: 1(608) 262-9051, Fax: 1(608)262-8400, Email: schaeffer@ssc.wisc.edu

Jennifer Dykema, PhD[3,4]
Phone: 1(608) 262-8385, Fax: 1(608)262-8400, Email: dykema@ssc.wisc.edu

*Corresponding author

[1]Associate Professor, Department of Sociology, Loyola University Chicago
1032 W. Sheridan Rd., Chicago, IL 60660 USA

[2]Sewell Bascom Professor of Sociology, Emerita, Department of Sociology, University of Wisconsin-Madison, 475 N. Charter Street Madison, WI 53706 USA

[3]Visiting Associate Professor, Department of Sociology, University of Wisconsin-Madison, 475 N. Charter Street Madison, WI 53706 USA

[4]Faculty Director, University of Wisconsin Survey Center, University of Wisconsin-Madison
475 N. Charter Street, Madison, WI 53706 USA

Word count: 6237

Keywords: interviewer-rated health; self-rated health; interviewer observations; interviewer evaluations; Innovation Panel

Acknowledgments

The Innovation Panel, part of the UK longitudinal study Understanding Society, fielded the experiment *Examining the Validity of Interviewers' Ratings of Respondents' Health*, in Wave 8 of the Innovation Panel. This work was supported by core funding to the Center for Demography and Ecology from the Eunice Kennedy Shriver National Institute of Child Health & Human Development (P2C HD047873) and core funding to the Center for Demography of Health and Aging from the National Institute on Aging (P30 AG017266) at the University of Wisconsin–Madison, and the Gannon Center for Women and Leadership at Loyola University Chicago. The study design and analysis were not preregistered. Earlier versions of the research were presented at Understanding Society Scientific Conference (2017) and the Midwest (2016) and American (2017) Association for Public Opinion Research. We thank participants in these conferences, JSSAM's Editor, Associate Editor, and the anonymous reviewers for their helpful comments. The opinions expressed herein are those of the authors.

Abstract

Interviewers' ratings of survey respondents' health (IRH) are a promising measure of health to include in surveys as a complementary measure to self-rated health (SRH). However, our understanding of the factors contributing to IRH remains incomplete. This is the first study to examine whether and how it matters when in the interview interviewers evaluate respondents' health in a face-to-face survey, in an experiment embedded in the UK Innovation Panel Study. We find that interviewers are more likely to rate the respondent's health as "excellent" when IRH is rated at the end of the interview compared to the beginning. Drawing from the continuum model of impression formation, we examined whether associations between IRH and relevant covariates vary depending on placement in interview. We find that across several characteristics of interviewers and respondents, only the number of interviews completed by interviewers varies by IRH assessment location in its effect on IRH. We also find evidence that interviewer variance is lower when IRH is assessed prior to compared to after the interview. Finally, the location of IRH assessment does not impact the concurrent or predictive validity of IRH. Overall, the results suggest that in a general population study with some health questions, there may be benefits to having interviewers rate respondents' health at the beginning of the interview (rather than at the end as in prior research) in terms of lower interviewer variance, particularly in the absence of interviewer training that mitigates the impact of within-study experience on IRH assessments.

Keywords: interviewer-rated health; self-rated health; interviewer observations; interviewer evaluations; Innovation Panel

Statement of Significance

Interviewers' ratings of survey respondents' health (IRH) are a promising measure of health to include in surveys as a complementary measure to self-rated health (SRH). However, our understanding of the factors contributing to IRH remains incomplete. This is the first study to examine whether and how it matters when in the interview interviewers evaluate respondents' health in a face-to-face survey. The results suggest that in a general population study with some health questions, there may be benefits to having interviewers rate respondents' health at the beginning of the interview (rather than at the end as in prior research) in terms of lower interviewer variance, particularly in the absence of interviewer training that mitigates the impact of within-study experience on IRH assessments.

Although the self-rated health (SRH) question – e.g., "would you say your health in general is excellent, very good, good, fair, or poor?" – is widely used to study health in surveys (Garbarski 2016; Idler and Benyamini 1997), interviewers' global health ratings (IRH) -- e.g., "Would you say the respondent's health in general is excellent, very good, good, fair, or poor?" -- have the potential to augment the measurement respondents' health. Previous research indicates that in face-to-face interviews in Taiwan, China, and the US, IRH provides supplementary information about respondents' health status: IRH and SRH are only moderately correlated (Brissette et al. 2003; Garbarski, Schaeffer, and Dykema 2019; Smith and Goldman 2011), IRH and SRH are associated with different correlates of health (Feng et al. 2016; Garbarski et al. 2019; Smith and Goldman 2011), IRH predicts mortality better than SRH (Brissette et al. 2003; Feng et al. 2016; Garbarski et al. 2019; Todd and Goldman 2013) and, in one study, better than physicians' ratings (Todd and Goldman 2013).

That IRH does not simply mirror SRH is unsurprising for at least two reasons. First, interviewers likely determine what is salient or problematic about the respondent's health differently than the respondent given the select information available (e.g., observing the respondent's physical functioning compared to the respondent's own assessment of their functioning) (Garbarski et al. 2019; Smith and Goldman 2011; Todd and Goldman 2013). Second, in the studies noted above, interviewers evaluated respondents' health at the end of long interviews containing many health-related questions. Thus, the better predictive validity of IRH compared to SRH may be due to how interviewers integrate the detailed health information that they learn and observe as respondents answer questions and perform tasks during the interview (Garbarski et al. 2019).

SRH is usually asked early in the interview because it can be influenced by context, both by the type of health questions preceding it and respondents' answers to those questions (Garbarski, Schaeffer, and Dykema 2015). In contrast, IRH is usually asked at the end of the interview, allowing interviewers to use the data they collected in ways that contribute to the predictive validity of IRH (at least with respect to mortality) (Garbarski et al. 2019). However, having more information to process could increase both the random and systematic error attributable to interviewers, reducing the reliability and validity of IRH. Thus, an issue unexplored by previous research concerns the optimal placement of IRH in the interview.

This study examines whether the distribution of IRH, the association of IRH with covariates, interviewer variance, and the concurrent and predictive validity of IRH, vary depending on the location of IRH: before the first substantive survey question, so that IRH is primarily based on the interviewer's initial impressions of the respondent's appearance and functioning, versus after the interview is completed, when the interviewer has the opportunity to summarize and integrate information from more extensive physical cues and answers to survey questions.

Background

Interviewers' observations and evaluations of the interview context have long been collected for a variety of administrative and analytic purposes, such as nonresponse and data quality adjustments (Olson and Parkhurst 2013). One set of interviewers' observations focuses on more objective features of the interviewing environment: features of the household (e.g., type of living quarters, noise, presence of others) or surrounding neighborhood (e.g., characteristics of buildings, streets, or yards) (Casas-Cordero et al. 2013; Sinibaldi, Durrant, and Kreuter 2013; West 2013; West and Kreuter 2013, 2015; West, Kreuter, and Trappmann 2014; West et al.

2020). Another set of observations is more subjective, focusing on interviewers' evaluations of respondents' behaviors: assessments of respondents' engagement or performance (e.g., how difficult answering questions seemed, how likely respondents would be to participate in the future) and interpersonal affiliation (e.g., the respondent's friendliness) (Freedman et al. 2012; Holbrook et al. 2014; Hurtado 1994; Kirchner, Olson, and Smyth 2017; Olson and Peytchev 2007; Tarnai and Paxson 2005; West et al. 2020).

The type of information assessed by IRH occupies a space between the more objective information interviewers observe about the environment and the more subjective assessments they make about respondents. Interviewers' ratings of respondents' health are comprised of interrelated factors such as 1) respondents' characteristics that interviewers ascertain from respondents' answers to survey questions, living environments, appearance, and functioning (physical, cognitive, and social); 2) interviewers' perceptions of some of what they have observed; and 3) interviewers' characteristics that influence how they interpret and integrate information to form an assessment, including their sociodemographic characteristics (e.g., gender, age, education, health) and past interviewing experience (Garbarski et al. 2019; West and Blom 2017).

In social psychology, dual process theories of social-information processing hold that actors process information along a continuum with more automatic and heuristic thinking at one end versus more integrative and deliberative thinking at the other (Chaiken and Trope 1999). This framework can be applied to understanding how interviewers might make observations about respondents. In particular, the continuum model of impression formation suggests that interviewers might form impressions about respondents' health using various levels of processing, ranging from category-based processing (based on stereotypes associated with immediately salient categories, such as gender, race/ethnicity, age, and body size) to individuating processing (or piecemeal integration, attribute by attribute, to form an overall impression) (Fiske and Neuberg 1990; Fiske, Lin, and Neuberg 1999). When making assessments of IRH at the end of the interview, as they do in prior studies, interviewers have the opportunity for piecemeal integration of information about the respondent's health based on the respondent's answers to survey questions and their own observations about the respondent's appearance, environment, and physical, psychological, and social functioning during the interview (Garbarski et al. 2019; Kirchner, Olson, and Smyth 2017).

In contrast, if interviewers assess respondents' health before the interview, their impressions will be based on the first moments in which they observe respondents. The continuum model of impression formation suggests that during these initial observations, interviewers might categorize respondents based on observable characteristics, then formulate impressions based on assumptions or stereotypes about groups with those characteristics (e.g., someone who is overweight assessed as being in poor health). In other words, with less information about the respondent before compared to the end of an interview, interviewers might be more likely to form impressions heuristically and with little effortful processing. Indeed, research suggests that under conditions of greater uncertainty, people rely more on stereotypes when making judgments about others (Kunda and Thagard 1996). Even if an interviewer does not rely on assumptions or stereotypes about the respondent, they have less information about the respondent at the beginning of the interview than the end.

A related issue is how the location of IRH—rated before or after the interview—affects concurrent and predictive validity. When a criterion is available to judge the accuracy of observations made in the first moments of an interaction, the influence of heuristics, memory retrieval biases, or other cognitive biases leads to errors compared to more intentional processes (Tversky and Kahneman 1974; Kahneman 2011). However, other work shows the opposite to be true, that more integrative or deliberative processing leads to errors (Ambady 2010; Ambady, Hallahan, and Conner 1999; Patterson and Stockbridge 1998). Psychologists have noted that people make attributions about others' personality characteristics from their facial features with consensus—although not necessarily accuracy—with implications for outcomes such as voting and criminal sentencing (Todorov et al. 2015). Although previous research indicates that part of the predictive validity of IRH when assessed at the end of the interview is because it incorporates respondents' health information ascertained from their answers to survey questions, having interviewers rate respondents' health at the end of the interview may also increase systematic error, particularly if the survey questions provide information that does not reflect health beyond what an interviewer observes in their first moments of interaction.

As with other types of more subjective interviewer evaluations, measures of IRH are also comprised of interviewer-specific variance. Previous research indicates intra-interviewer correlations around 0.08 for IRH assessed at the end of the interview (Brissette et al. 2003; Garbarski et al. 2019). In terms of whether interviewer variance in IRH is lower when IRH is assessed at the beginning or end of the interview, we posit two competing hypotheses. First, interviewer variance could be lower when IRH is assessed before the interview because this placement focuses interviewers' processing on a limited set of characteristics in a similar way: initial observations of visible sociodemographic characteristics, appearance, physical functioning, living environment, and so forth. The competing hypothesis is that at the end of the interview, interviewers will be able to integrate a more expansive set of health-relevant information to rate respondents' health in a similar way across respondents—answers to survey questions as well as observations of physical, cognitive, and social functioning from throughout the interview. The issue of interviewer variance in rating respondents' health becomes one of interviewers consistently considering a smaller set of information versus a more expansive set of health-relevant information. In other words, is the additional information available by the end of the interview applied differently by interviewers in a way that increases interviewer variance? Or does the health-relevance of the information available in the interview give interviewers a framework with which to more consistently assess IRH?

The current study experimentally varies whether the interviewer rates the respondent's health before or after the substantive interview questions and explores how the measurement of IRH differs based on its location. We examine the following research questions:

1. Do associations between IRH and relevant covariates vary based on the location of IRH? Drawing from the continuum model of impression formation, we pose two complementary hypotheses:

   Hypothesis 1a: Respondents' and interviewers' sociodemographic characteristics and interviewers' observations about the household (such as neighborhood and household appearance that occur prior to contact with members of the household) each have stronger relationships with IRH when IRH is assessed prior to the interview, since the interviewer has less health-specific information to draw on and is constrained to forming their assessments based on this limited information.

   Hypothesis 1b: Similarly, the association between health-relevant covariates (that the interviewer asks about during the interview) and IRH is stronger when IRH is assessed at the end of the interview.

2. Does the concurrent and predictive validity of IRH vary based on the location of IRH? To the best of our knowledge, this is the first study to experimentally manipulate the location of IRH in the interview. Past studies of impression formation with other measures are mixed, with some studies showing better validity with initial observations, and others showing the opposite. Thus, we examine two competing hypotheses:

   Hypothesis 2a: IRH shows better concurrent and predictive validity when it is assessed at the beginning of the interview.

   Hypothesis 2b: IRH shows better concurrent and predictive validity when it is assessed at the end of the interview.

3. Does interviewer variance in IRH vary based on the location of IRH? We do not have a single a priori hypothesis based on prior research and instead propose two competing hypotheses:

   Hypothesis 3a: Interviewer variance is lower when IRH is assessed at the beginning of the interview.

   Hypothesis 3b: Interviewer variance is lower when IRH is assessed after the end of the interview.

Methods

Data

Data are from the UK Innovation Panel Study (IPS), part of the UK Understanding Society study (University of Essex 2019; https://www.ukdataservice.ac.uk/get-data/how-to-access.aspx). The IPS is an omnibus study that embeds multiple experimental studies in a longitudinal design. Members are drawn from a random sample of households in England, Scotland, and Wales. The random sample of households is periodically refreshed as households attrite from the sample. The current study uses panel data from Wave 8 (2015), in which 2,378 individuals were interviewed either in person or online. The Wave 8 AAPOR RR2 is 84.9%; this response rate is modified in the sense that it is computed for households that are fielded in the original Wave 1 sample and the refresher samples in Waves 4 and 7, as ineligible households are not issued fieldwork (personal communication with Understanding Society User Support, September 2020). We restrict our analyses to the 1,439 respondents who were randomly assigned to be interviewed in person in their residence by 110 interviewers, who completed between 1 and 67 interviews (mean=13, standard deviation=9.65, median=12 interviews). The in-person interviews consisted of both interviewer-administered questions followed by a computer-assisted self-interview (CASI). Our analytic samples vary in size depending on the model due to small amounts of missing data for respondent and interviewer characteristics.

Interviewers are assigned to respondents based on geography and shift. Thus, the design of the study is not interpenetrated. The sample is issued in monthly batches in order to maximize contact with households. All respondents in the household are to be interviewed. There is a re-issue period later in the field period where non-responding cases may be assigned to another interviewer (https://www.understandingsociety.ac.uk/documentation/mainstage/technical-reports). Thirty-nine percent of respondents were interviewed by the same interviewer in Wave 8 as in Wave 7 (see Sensitivity Analyses in the Results section).

The main experimental manipulation was the random assignment of the interviewer's rating of the respondent's health (IRH) either before the interview (i.e., before any substantive questions) or after the interview but before other interviewer assessments. Random assignment occurred at the level of the respondent rather than stratified by interviewer such that a given interviewer completed assessments before the interview for some respondents and after the interview for others. (We confirmed that this randomization scheme did not result in any interviewers completing all of their assignments in only one experimental condition [not shown].) Although interviewers were not given explicit instructions on how to rate respondents' health, they were told not to change any of their observations about the household or respondent made prior to the interview if they learned additional information during the interview (Understanding Society User Support, May 2017). We do not have access to paradata to confirm that interviewers followed these instructions.

Measures

Table 1 shows the descriptive statistics for the measures used in the study. The dependent variable, IRH, is described above. Respondents' characteristics include observations about the household made by interviewers prior to household contact, sociodemographic characteristics, and health information from survey questions asked during the interview. *Interviewers' observations about the household* and surrounding area were made before the interview and include: an indicator for whether there were any boarded-up houses, trash or junk on road, or heavy traffic; a rating of the condition of the household as "good" or "fair" compared to "bad"; and a rating of the condition of the household compared to others in the neighborhood as "better" or "about the same" compared to "worse." We also include "missing" as a category for each of these variables, as we expect the data are missing not at random given the large percentage missing (9%) relative to other variables in the dataset for Wave 8. Alternatives were to drop the cases by listwise deletion or to use multiple imputation to replace the missing data, which is justifiable when data are missing at random but potentially problematic when data are missing not at random.

Respondents' sociodemographic characteristics include their gender, age, rural or urban location, and race/ethnicity (person of color or white as denoted by the racial/ethnic groups in the Innovation Panel: "British/English/Scottish/Welsh/Northern Irish," "Irish," "other white" as "white" compared to all other groups). We again include "missing" as a category for race/ethnicity given the large amount of missing data (12%).

Health questions from the interview include questions about health conditions, health behaviors, and healthcare visits. All respondents were asked whether they have any chronic condition ("Do you have any long-standing physical or mental impairment, illness or disability?" in which longstanding is then defined as at least 12 months). We also generated a measure of the number of reported health conditions: respondents in their first IPS interview reported whether a doctor has ever told them that they have a particular health condition and continuing respondents reported about health conditions since the last interview. (Note that this measure could not be used to estimate the incidence or prevalence of health conditions in the population, given the varying reference periods—ever vs. since last interview—and does not account for whether and when the condition may have resolved nor its severity.) Respondents also reported whether they currently smoked cigarettes and the number of days they engaged in three types of activity for at least 10 minutes: vigorous (activities that make breathing harder than normal), moderate (activities that make breathing somewhat harder than normal), and walking. Finally, respondents provided the number of visits in the past year to a general practitioner, number of outpatient hospital visits, and whether they spent time as an inpatient in the hospital not due to childbirth.

Interviewers' characteristics include gender, age, number of years of prior interviewing experience with the current field work agency, and number of interviews the interviewer completed prior to the current interview (there was no significant interaction between prior and within-survey experience in predicting IRH [Kirchner and Olson 2017; Olson and Peytchev 2007]). Interviewers received their caseload all at once so that cases completed at the end of the field period could have more contacts. Thus, we also control for the number of times a respondent was contacted by an interviewer to further isolate the impact of the number of interviews completed by the interviewer. We did not include interviewers' race/ethnicity as a covariate because only two interviewers were people of color and two did not have information on their race.

We examine the bivariate association between independent health measures and IRH to assess the validity of the two locations of IRH (described in Analytic Strategy below). We use health measures from the CASI portion of the interviews to assess *concurrent validity*. Measures include self-rated health (SRH), health satisfaction (summed so that a higher scores indicates more satisfaction), subjective well-being using the general health questionnaire (GHQ) scale (higher scores indicate more distress), and the SF-12 physical and mental component scores (higher scores indicate better physical and mental health, respectively). To examine the *predictive validity* of IRH, we use measures from the interviewer- and self-administered portions of the Wave 9 interview. These include whether the respondent participated in Wave 9 (main survey), any reported chronic health condition, any reported disability (this was not asked of everyone in Wave 8), the number of health conditions (following the same procedure described for Wave 8), and health satisfaction rating.

As part of a sensitivity analysis discussed in the Results section, we included the *length of time* of the Wave 8 interview in minutes as an indicator of the amount of the fatigue interviewers might have at the end of the interview.

Table 1. Descriptive Statistics, Innovation Panel Waves 8 and 9 (2015-6), United Kingdom

	Mean or Percentage	Std. Dev.	Min	Max	Missing %
Respondents' characteristics					
Female (vs. Male)	54.55 %				
Age	50.37	18.80	16	97	0.28%
Person of color (vs. white)	7.16 %				12.16%
Urban (vs. rural) location	78.53 %				0.07%
Interviewers' pre-interview household observations, Wave 8					
Vicinity has boarded houses, trash, traffic (vs. not)	15.77 %				8.76%
Household condition is fair/bad (vs. good)	1.74 %				9.31%
Household is in worse condition (vs. better/same) than others in area	4.66 %				9.45%
Health measures in interviewer-administered questionnaire, Wave 8					
Any chronic condition (vs. none)	40.31 %				0.14%
Number of health conditions reported					0.28%
0	84.92 %				
1	12.16 %				
2 or more	2.64 %				
General practice visits in the past year					0.42%
None	19.18 %				
1 to 2	40.31 %				
3 to 5	21.61 %				
6 to 10	9.94 %				
10 or more	8.55 %				
Hospital outpatient visits in the past year					0.28%
None	53.72 %				
1 to 2	26.41 %				
3 or more	19.60 %				
Any time spent hospital inpatient not due to childbirth (vs. none)	8.96 %				0.14%
Current smoker (vs. not)	18.07 %				0.07%
Number of days of vigorous activity					0.35%
0	56.01 %				
1 to 2	18.35 %				
3 to 4	12.79 %				
5 to 7	12.51 %				
Number of days of moderate activity					0.56%
0	46.49 %				
1 to 2	19.60 %				
3 to 4	13.20 %				
5 to 7	20.15 %				

Table 1. Descriptive Statistics, Innovation Panel Waves 8 and 9 (2015-6), United Kingdom

	Mean or Percentage	Std. Dev.	Min	Max	Missing %
	Mean or Percentage	Std. Dev.	Min	Max	Missing %
Number of days of walking					0.42%
0	15.77	%			
1 to 3	25.64	%			
4 to 6	19.25	%			
7	38.92	%			
Number of contacts for respondent	4.30	2.77	1	20	.21%
Wave 8 interview time in minutes	39.58	13.30	9.22	148.70	.83%
Health measures from self-administered questionnaire, Wave 8					
Self-rated health					4.86%
Poor	5.98	%			
Fair	18.07	%			
Good	33.63	%			
Very good	27.03	%			
Excellent	10.42	%			
Satisfaction with health	4.96	1.67	1	7	4.93%
Subjective well being	10.81	5.30	0	36	5.84%
SF-12 Physical Component Score	48.26	11.43	7.02	68.18	5.84%
SF-12 Mental Component Score	49.69	10.06	9.78	69.63	5.84%
Wave 9 measures					
Participated in Wave 9 (vs. did not)	83.74	%			
Any chronic condition (vs. none)	33.91	%			16.40%
Any disability reported	24.39	%			16.89%
Any health condition reported	11.61	%			17.37%
Satisfaction with health	4.95	1.62	1	7	19.81%
Interviewers' characteristics (N=110)					
Female (vs. Male)	39.09	%			1.82%
Person of color (vs. white)	1.82	%			1.82%
Interviewers' age	57.19	11.18	25	80	1.82%
Years of experience with organization	6.61	5.94	0	25	1.82%
Total number of interviews completed for Wave 8	12.99	9.65	1	67	

Notes. N=1,439 respondents who were interviewed in person in Wave 8 of the Innovation Panel. No significant differences were found in the distribution of covariates across experimental treatment.

Analytic Strategy

We conducted analyses in Stata Version 16.1, using a chi-square test that adjusts for clustering of respondents within interviews (clchi2) and ordinal logistic mixed-effects regressions. The mixed-effects regressions account for the nesting of respondents within households and households within interviewers with a random intercept for households nested within interviewers, a random intercept for interviewers, a random slope for the placement of IRH, and a covariance between the random slope and intercept to account for the fact that interviewers are assessing respondents' health in both locations across the study period and may vary in how they treat the question depending on its placement. We restricted our analysis to instances in which interviewers conducted at least 2 interviews in each experimental condition to avoid estimation issues in the mixed-effects model. The Stata code for the analyses is available in Online Appendix A.

The parameterization of the models for Hypotheses 1 and 2 follows a dummy parameterization or contrast coding (Loosveldt and Buellens 2014). We denote IRH rated at the beginning of the interview with *Beg*=1 and *Beg*=0 for IRH rated at the end of the interview. The model predicting IRH for respondent *i*, household *j*, and interviewer *k* (Rabe-Heskreth and Skrondal 2012) is

Model 1: $y^*_{ijk} = \beta_1 Beg_{ijk} + \zeta_{jk} + \zeta_{1k} + \zeta_{2k} Beg_{ijk} + \varepsilon_{ijk}$

with

$$\begin{bmatrix}\zeta_{1k}\\ \zeta_{2k}\end{bmatrix} \sim N(0,\psi) \text{ where } \psi = \begin{bmatrix}\sigma^2_{\zeta_{1k}} & \sigma_{(\zeta_{1k},\zeta_{2k})}\\ \sigma_{(\zeta_{2k},\zeta_{1k})} & \sigma^2_{\zeta_{2k}}\end{bmatrix}$$

and

$$[\zeta_{jk}] \sim N\left(0, \sigma^2_{\zeta_{jk}}\right)$$

In this model, β_1 is the fixed effect for IRH being rated at the beginning of the interview compared to the end, across all interviewers and households. ζ_{1k} is the random intercept for interviewers, which covaries with the random slope for where in the interview IRH is assessed (ζ_{2k}). ζ_{jk} is the random intercept for households, which are nested within interviewers.

Hypotheses 1 and 2 explicate interaction effects between the location of IRH assessment and covariates (Hypothesis 1) and measures of concurrent and predictive validity (Hypothesis 2). Thus, Model 2 extends Model 1 to include covariates Z_{ijk} and an interaction between Z_{ijk} and Beg_{ijk}

Model 2: $y^*_{ijk} = \beta_1 Beg_{ijk} + \beta_2 Z_{ijk} + \beta_3 Beg_{ijk} Z_{ijk} + \zeta_{jk} + \zeta_{1k} + \zeta_{2k} Beg_{ijk} + \varepsilon_{ijk}$

To examine whether the interviewer variance varies across IRH placement (Hypothesis 3), we estimate an alternative mixed-effects regression model that omits the random intercept for interviewers and instead includes two random slopes, one for each group of respondents (having their IRH rated before the interview and after). (The random intercept for households nested within interviewers is omitted as well to allow for estimation of the two random slopes.) This separate coding model (Herzing 2018; Jones and Subramanian 2013) allows for direct estimation of both variance components and their covariance by including in the random part of the model all dummy variables with no reference category omitted (*Beg*=1 when IRH is rated prior to the interview and 0 at the end, and *End*=1 for IRH rated at the end of the interview and 0 at the beginning) rather than the contrast coding (only *Beg* included) described in Models 1 and 2. The dummy variables in this model represent a direct estimate of the interviewer variance for each group of respondents (IRH assessed before or after the interview).

The base model predicting IRH for respondent *i* and interviewer *j* is

Model 3: $y^*_{ij} = \beta_1 Beg_{ij} + \zeta_{1j} Beg_{ij} + \zeta_{2j} End_{ij} + \varepsilon_{ij}$

With

$$\begin{bmatrix}\zeta_{1j}\\ \zeta_{2j}\end{bmatrix} \sim N(0,\psi) \text{ where } \psi = \begin{bmatrix}\sigma^2_{\zeta_{1j}} & \sigma_{(\zeta_{1j},\zeta_{2j})}\\ \sigma_{(\zeta_{2j},\zeta_{1j})} & \sigma^2_{\zeta_{2j}}\end{bmatrix}$$

We then conduct a likelihood ratio chi-square test comparing this model to one in which the random coefficients are constrained to be equal (West and Elliott 2014). We also calculate the intraclass correlations for Model 3 as $\rho_{\text{beg}}=\sigma^2_{\zeta 1_j}/\ (\sigma^2_{\zeta_{1j}}+\pi^2/3)$ and $\rho_{\text{end}}=\sigma^2_{\zeta 2_j}/\ (\sigma^2_{\zeta_{2j}}+\pi^2/3)$, in which $\sigma^2_{\zeta_{1j}}$ and $\sigma^2_{\zeta_{2j}}$ are the interviewer-level variances and $\pi^2/3$ is the error variance for logistic regression models (Hedeker 2003).

The lack of random assignment of respondents to interviewers means that the variance component for interviewers is likely overestimated in that it conflates interviewer effects with geographic and other clustering since interviewer assignments are often based on geography. We do not have data for geographic clusters available for this analysis. Thus, estimates of interviewer variance may be inflated, but the difference in interviewer variance across the experimental treatments is not biased given that random assignment occurs at the level of the respondent.

Results

The overall distribution of IRH varies depending on its placement in the interview (cluster chi-square= 11.092, df=4, p=.026) (Table 2). Descriptively, a greater share of answers are "excellent" and fewer are "good" when IRH is assessed at the end of the interview compared to before: at the end of the interview, over half of respondents are rated as having "excellent" health, compared to 38% of respondents rated as "excellent" when their health is rated by the interviewer at the beginning of the interview.

Table 2. Distribution of interviewers' ratings of respondents' health (IRH) before or after interview, Innovation Panel Wave 8 (2015), United Kingdom

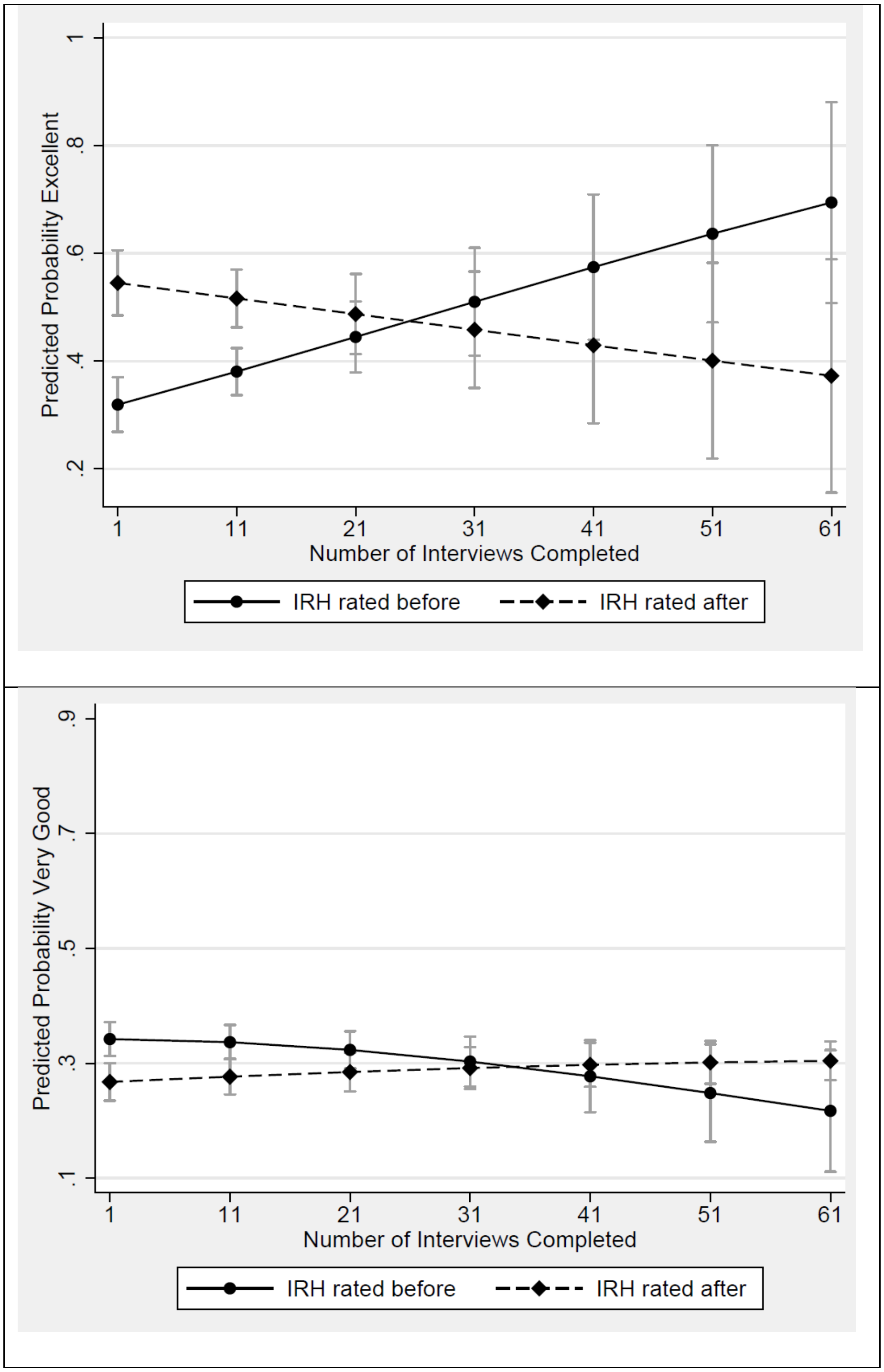
	IRH Before		IRH After	
Excellent	38	%	52	%
Very good	31	%	28	%
Good	19	%	10	%
Fair	8	%	6	%
Poor	3	%	4	%
Total	99	%	100	%
N	708		731	

Notes. Columns may not sum to 100% due to rounding. "Before" and "after" indicate random assignment to have the interviewer rate the respondent's health prior to or after the interview. Distribution is significantly different across groups (cluster adjusted chi-square= 11.09, df=4, p<.026). Question text: Would you say the respondent's health in general is excellent, very good, good, fair, poor?

We examined whether the association between IRH and the covariates of interest that interviewers can ascertain through observation or answers to survey questions vary depending on when interviewers rate respondents' health. In a model that includes the interaction of when in the interview IRH is assessed (before vs. after) with each of these respondent and interviewer characteristics, only the interaction between IRH location and the number of interviews completed by the interviewer was significant (Online Appendix B). Thus, neither Hypothesis 1a or 1b is supported, as the association between interviewer characteristics, respondent characteristics, and health-relevant covariates with IRH largely does not vary by whether IRH is assessed at the beginning or end of the interview. In this model, the variance of the random slope is positive and significant, indicating that the effect of location of IRH assessment on IRH varies across interviewers. The covariance between the random intercept and slope is not significant, indicating that the effect of IRH location is not associated with interviewers' mean ratings of respondents' health. The variance of the random intercept for household is significant, indicating that that overall level of IRH varies between households (Online Appendix B).

Figure 1 illustrates the interaction between the number of interviews completed and IRH location in their effect on IRH in the metric of predicted probabilities. The top part of the figure shows the probability that interviewers rate respondents in “excellent” health is lower when this rating occurs before the interview and increases with more interviews completed, and is higher when this rating occurs after the interview and decreases with more interviews completed. The bottom of the figure shows the probability that interviewers rate respondents in “very good” health follows the opposite pattern: higher when the rating occurs before the interview and decreases with more interviews completed, and lower when rated after the interview and increases with more interviews completed; the pattern for “good,” “fair,” and “poor” ratings mirrors that of “very good.”

Figure 1. Predicted probabilities of interviewers' ratings of respondents' health (IRH) by location of IRH assessment and interviewers' number of interviews completed, Innovation Panel Wave 8 (2015), United Kingdom.



Note. Spikes are 95% confidence intervals.

We next examine whether the concurrent and predictive validity of IRH varies by IRH location. The results listed in Online Appendix C show that the association of IRH with each of the measures of concurrent and predictive validity is not significantly different between IRH locations. Thus, we find no support for Hypotheses 2a or 2b.

We then examine whether interviewers' unique contribution to variation in IRH depends on IRH location, with an ordinal logistic mixed-effects regression of IRH on location of assessment and using separate coding for the random part of the model (Model 3 from the Analytic Strategy section). The interviewer variance components and resulting intraclass correlations for IRH are larger when IRH is assessed after the interview (ρ_{end}=.30) compared to before (ρ_{beg}=.17) ($\sigma^2_{\zeta_{1j}}$=0.66 [95% CI: 0.39, 1.11], $\sigma^2_{\zeta_{2j}}$1.40 [95% CI: 0.86, 2.28], $\sigma_{(\zeta_{1j},\zeta_{2j})}$0.82 [95% CI: 0.45, 1.19]). The likelihood ratio test comparing this model to the model in which the variances are constrained to be equal is LR $\chi^2_{(1)}$=6.34, which is .01<*p*<.05. Thus, there is evidence in support of Hypothesis 3a, that the interviewer variance is lower when IRH is assessed prior to the interview, which we posited is because interviewers focus on a limited set of characteristics in a similar way.

Sensitivity analyses

To ensure differences across experimental treatments were not due to the uneven distribution of cases across the field period, we examined whether the distribution of cases across the experimental treatments varies across interview date (days since January 1, 1960), finding no evidence for this difference (t=1.23, *p*=.22). Thus, although the composition of cases may change across the field period, the difference between the experimental treatment groups appears to remain constant.

As an explanation for the greater proportion of "excellent" ratings when IRH is assessed at the end of the interview compared to before, we examined whether the difference in the distribution of IRH across assessment location could be due to fatigue on the part of the interviewer, leading interviewers to satisfice or disproportionately select the first response option that seems acceptable (Krosnick 1991). If interviewers are fatigued and rushing through the task of rating respondents' health at the end of the interview, then longer interview times should be associated with better ratings of respondents' health (since the response options start with the positive end of the scale). However, the interaction between interview time length and IRH assessment location in predicting IRH was not statistically significant (Online Appendix D).

We also examined whether the results varied by the study's panel design, in that the same interviewer may interview the respondent at Wave 7 and 8, thus minimizing differences in the effect of placement on IRH if the interviewer recalls information about the respondent from Wave 7 during the Wave 8 interview. Thirty nine percent of respondents were interviewed by the same interviewer in Waves 7 and 8. We examined three-way interactions of interviewer status in Wave 8 (same or different interviewer), assessment location, and 1) respondent and interviewer characteristics and 2) measures of concurrent and predictive validity. We uncovered no significant effects (results available upon request).

Discussion

The promise of having interviewers rate respondents' health in face-to-face surveys has been documented in previous research. However, it is unclear when in the interview these assessments should occur. Interviewers could rate the health of respondents at the end of the interview as they have in prior studies. This strategy affords interviewers the opportunity to summarize across physical cues and respondents' answers to the survey questions. Alternatively, interviewers could rate respondents’ health before the first substantive interview question, constraining their ratings to their initial impressions of respondents’ appearance and functioning.

We find that the overall distribution of IRH varies between locations: IRH is more likely to be “excellent” when interviewers rate respondents at the end of the interview than the beginning. We examined but found no evidence to support that the difference in IRH rating by location was due to fatigue (proxied by the length of the interview) or respondents’ answers to health-relevant survey questions. We speculate that the larger proportion of “excellent” ratings at the end of the interview might be due to norms of reciprocity or politeness: at the end of an interview in which the respondent has answered survey questions, interviewers may be less willing to rate respondents’ health as something other than “excellent.” Furthermore, the health-relevant information respondents report during the interview is couched within several other modules that vary in their health-relevance, such that other aspects of the interview may interfere with the health information about respondents that interviewers are encoding and storing in their working memory. This may lead interviewers to draw more on the interaction and rapport from during the interview when rating respondents’ health at the end of the interview (Garbarski, Schaeffer, and Dykema 2016; Kirchner, Olson, and Smyth 2017). Each of these hypotheses should be examined by future research, for example using cognitive interviews of interviewers asking how they rate respondents’ health.

Previous research shows that IRH is associated with a range of interviewer characteristics and respondent characteristics that interviewers learn and observe about the respondent during face-to face interviews (Garbarski, Schaeffer, and Dykema 2019). The current study demonstrates that the association of IRH with interviewers’ and respondents’ characteristics largely does not vary by the location of IRH. This is particularly intriguing since interviewers do not know respondents' answers to the health-relevant survey questions when they are assessing IRH prior to the interview. Thus, information that is (thus far) unmeasured in the model is informing interviewers' assessments of respondents' health in ways that lead to different distributions of IRH by location. These could be other observations of respondents' living environments; appearance; and physical, cognitive, and social functioning from the current the interview or prior interviews. Future research should continue to examine the factors that inform IRH, for example, information from prior interviews.

The finding that the association between IRH and the number of interviews completed by the interviewer varied by the location of IRH indicates that interviewers formulate their assessments of respondents' health based in part on their cumulative experience interviewing respondents in this survey (Kirchner and Olson 2017; Olson and Peytchev 2007). The effects observed at the beginning of interviewers' within-study experience (probability of "excellent" higher when rated at end of interview and lower at beginning of interview) may be because interviewers 1) are more likely to be influenced by norms of reciprocity or interference in working memory and rapport at the end of the interview and 2) underestimate better health at the beginning of the interview because of limited information, and each of these processes diminish with more experience and a more representative reference group (the study population). In other words, because interviewers rate respondents' health at both locations across the field period (as the random assignment of location of the rating is at the level of the respondent), their implicit models of how to rate respondents' health continually revise rather than remain constant across location of assessment. Future research should examine whether this effect is replicated in other contexts—e.g., with more detailed health information in the survey, a similar general population study in another country, telephone or virtual modes of administration, and interviewers' evaluations about respondents that are not about health—to examine why IRH varies by the number of completed interviews and where IRH is assessed.

We were able to include several interviewer characteristics in the analysis (gender, age, years of experience, and number of interviews completed), yet the variance of the random slope is positive and significant, indicating that the effect of location on IRH varies across interviewers in ways that are unexplained by the current model (Online Appendix B). Future research with a more comprehensive set of interviewer characteristics should be used to explain the between-interviewer variance in the location effects of IRH (and other interviewer evaluations), such as previous experience in certain types of studies (e.g., studies with other interviewer observations and evaluations); workload; participation in general and study-specific trainings, debriefings, and quality control check-ins; and measures of interviewers' attitudes, beliefs, and personality characteristics (Olson et al. 2020).

We proposed two competing hypotheses on the validity of IRH depending on its location in the survey instrument, based on previous studies of IRH and social psychological studies that show mixed effects with regard to validity and information processing. We find no difference in validity based on IRH location. However, we note that the criteria (which do not include the gold standard criterion of mortality) or time lag (one year) may not allow for a holistic assessment of the predictive validity of IRH or differences in its predictive validity across location, an issue future research could explore.

The results of this study have implications for the implementation of interviewer observations and evaluations in survey research. First, this study adds evidence for the utility of an interviewer observation—IRH—that could easily be incorporated more broadly into other interviewer-administered surveys and is already included in several publicly available data sources (with minimum data use agreements) with populations from Taiwan (Social Environment and Biomarkers of Aging Study), China (Chinese Longitudinal Healthy Longevity Survey), the UK (Understanding Society Innovation Panel) and the US (General Social Survey, Wisconsin Longitudinal Study). In addition, researchers could consider having interviewers conduct evaluations that are not dependent on engagement with the respondent or the respondent's performance in the interview—such as IRH and the respondent's appearance—at the beginning of the interview when they are making more objective observations about the household. This is supported the fact that interviewer variance is higher when IRH is assessed after compared to before the interview. Importantly, this would also serve to break up the potential common method variance presumed to exist among the interviewers' more subjective evaluations that occur at the end of the interview. Indeed, IRH shows stronger relationships with the other interviewer evaluations about respondents (cooperation, interest, and so forth) that occur at the end of the interview when IRH is also administered at the end of the interview compared to at the beginning (not shown), although the placement of IRH and other interviewer observations needs to be fully crossed in an experimental design to distinguish whether this is due to correlated measurement error or better measurement. Second, in terms of training interviewers on how to do make assessments or evaluations about respondents: like most studies of IRH, interviewers were not explicitly trained on how to make ratings or observations about respondents or households (Understanding Society User Support Feedback, May 2017). Yet the fact that the association between the number of interviews completed and IRH varies across location of IRH assessment indicates that training interviewers on how to complete these assessments needs to be considered.

This study contains limitations. First, the focus of the study and the population differs from that of prior studies of IRH. While prior studies tend to focus on older populations and survey conditions primarily focused on health and functioning, the current study is based on a general population survey that included a more limited set of health questions than the prior studies of IRH. We argue that it is important to examine IRH in this sort of context, as we seek to study health across a range of populations; however, it does limit the direct comparability to previous studies of IRH. Future studies would do well to make the direct comparison of the performance of IRH in studies that cross 1) the population (general population vs. older adults), 2) the survey conditions (more limited vs. considerable health focus), and 3) mode (phone interviews have not been studied with respect to IRH and provide a different context for information for interviewers) to uncover the boundaries of the validity of the IRH measure across a range of survey conditions. In addition, because it is a general population survey, some social characteristics, such as the respondents' and interviewers' race and ethnicity, are not well represented, precluding our ability to examine their potential role. Finally, interviewers' observations and evaluations are prone to measurement error, which may attenuate some of the associations examined here (West 2013; West and Kreuter 2013).

References

Ambady, N. (2010), "The Perils of Pondering: Intuition and Thin Slice Judgments. *Psychological Inquiry,* 21(4), 271-278.

Ambady, N., Hallahan, M., and Conner, B. (1999), "Accuracy of Judgments of Sexual Orientation from Thin Slices of Behavior," *Journal of Personality and Social Psychology,* 77(3), 538-547.

Brissette, I., Leventhal, H., and Leventhal, E. A. (2003), "Observer Ratings of Health and Sickness: Can Other People Tell Us Anything about Our Health That We Don't Already Know?" *Health Psychology,* 22(5), 471-478.

Casas-Cordero, C., Kreuter, F., Wang, Y., and Babey, S. (2013), "Assessing the Measurement Error Properties of Interviewer Observations of Neighbourhood Characteristics," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 227-249.

Chaiken, S., and Trope, Y. (1999), *Dual-Process Theories in Social Psychology.* New York: Guilford Press.

Feng, Q., Zhu, H., Zhen, Z., and Gu, D. (2016), "Self-Rated Health, Interviewer-Rated Health, and Their Predictive Powers on Mortality in Old Age," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences,* 71(3), 538-550.

Fiske, S. T., Lin, M., and Neuberg, S. L. (1999), "The Continuum Model: Ten Years Later," In S. Chaiken & Y. Trope (Eds.), *Dual-Process Theories in Social Psychology* (pp. 231-254), New York, NY, US: Guilford Press.

Fiske, S. T., and Neuberg, S. L. (1990), "A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology, 23*, (pp. 1-74), Elsevier.

Freedman, V., Stafford, F.P., Conrad, F.G., Schwarz, N., and Cornman, J. (2012), "Assessing Time Diary Quality for Older Couples: An Analysis of the Panel Study of Income Dynamics' Disability and Use of Time (DUST) Supplement," *Annals of Economics and Statistics,* 105/106, 271–89.

Garbarski, D. (2016), "Research in and Prospects for the Measurement of Health Using Self-Rated Health," *Public Opinion Quarterly,* 80(4), 977-997.

Garbarski, D., Schaeffer, N., and Dykema, J. (2015), "The Effects of Response Option Order and Question Order on Self-Rated Health. *Quality of Life Research,* 24(6), 1443-1453.

Garbarski, D., N.C. Schaeffer, and Dykema, J. (2016), "Interviewing Practices, Conversational Practices, and Rapport: Responsiveness and Engagement in the Standardized Survey Interview," *Sociological Methodology,* 46, 1-38.

Garbarski, D., Schaeffer, N. C., and Dykema, J. (2019), "Interviewers' Ratings of Respondents' Health: Predictors and Association with Mortality," *The Journals of Gerontology: Series B,* 74(7), 1213–1221.

Hedeker, D. (2003), "A Mixed-Effects Multinomial Logistic Regression Model," *Statistics in Medicine,* 22(9), 1433-1446.

Herzing, J. M. E. (2018), *The Impact of Technological Change on Survey Nonresponse and Measurement* [Doctoral dissertation, University of Mannheim]. Mannheim: University Library Mannheim.

Holbrook, A. L., Anand, S., Johnson, T. P., Cho, Y. I., Shavitt, S., Chávez, N., and Weiner, S. (2014), "Response Heaping in Interviewer-Administered Surveys: Is it Really a Form of Satisficing?" *Public Opinion Quarterly*, 78(3), 591-633.

Hurtado, A. (1994), "Does Similarity Breed Respect: Interviewer Evaluations of Mexican Descent Respondents in a Bilingual Survey," *Public Opinion Quarterly,* 58,77–95.

Jones, K and Subramanian, S.V. (2013), *Developing Multilevel Models for Analysing Contextuality, Heterogeneity and Change using mlwin 2.2,* Volume 1, Bristol: Centre for Multilevel Modelling.

Jylhä, M. (2009), "What is Self-Rated Health and Why Does it Predict Mortality? Towards a Unified Conceptual Model," *Social Science & Medicine*, 69(3), 307-316.

Kahneman, D. (2011), *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kirchner, A. and Olson, K. (2017), "Examining Changes of Interview Length Over the Course of the Field Period," *Journal of Survey Statistics and Methodology* 5:84-108.

Kirchner, A., Olson, K., and Smyth, J. D. (2017), "Do Interviewer Postsurvey Evaluations of Respondents' Engagement Measure Who Respondents Are or What They Do? A Behavior Coding Study," *Public Opinion Quarterly,* 81(4), 817-846.

Krosnick, J. A. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology,* 5(3), 213-236.

Kunda, Z., and Thagard, P. (1996), "Forming Impressions from Stereotypes, Traits, and Behaviors: A Parallel-Constraint-Satisfaction Theory," *Psychological Review,* 103(2), 284-308.

Loosveldt, G. and Beullens, K. (2014), "A Procedure to Assess Interviewer Effects on Nonresponse Bias," *SAGE Open,* 4, 1–12.

Lynn, P. (2019), "Applying Prospect Theory to Participation in a CAPI/Web Panel Survey," *Public Opinion Quarterly* 83(3), 559-67.

Idler, E. L., and Benyamini, Y. (1997), "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies," *Journal of Health and Social Behavior,* 38(1), 21-37.

Olson, K., and Parkhurst, B. (2013), "Collecting Paradata for Measurement Error Evaluations," In F. Kreuter (ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information,* John Wiley and Sons, 43-72.

Olson, K., and Peytchev, A. (2007), "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes," *Public Opinion Quarterly,* 71(2), 273-286.

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., and West, B. T. (2020). "The Past, Present, and Future of Research on Interviewer Effects," In (K. Olson et al., Ed.) *Interviewer Effects from a Total Survey Error Perspective*, Boca Raton: CRC Press, 3-16.

Patterson, M. L., and Stockbridge, E. (1998), "Effects of Cognitive Demand and Judgment Strategy on Person Perception Accuracy," *Journal of Nonverbal Behavior,* 22(4), 253-263.

Sinibaldi, J., Durrant, G. B., and Kreuter, F. (2013), "Evaluating the Measurement Error of Interviewer Observed Paradata," *Public Opinion Quarterly*, 77(S1), 173-193.

Smith, K. V., and Goldman, N. (2011), "Measuring Health Status: Self-, Interviewer, and Physician Reports of Overall Health," *Journal of Aging and Health,* 23(2), 242-266.

Tarnai, J., and Paxson, M.C. (2005), "Interviewer Judgments about the Quality of Telephone Interviews," In *American Statistical Association, Proceedings of the Survey Research Methods Section,* 3988–94.

Todd, M. A., and Goldman, N. (2013), "Do Interviewer and Physician Health Ratings Predict Mortality?: A Comparison with Self-Rated Health," *Epidemiology,* 24(6), 913-920.

Todorov, A., Olivola, C. Y., Dotsch, R., and Mende-Siedlecki, P. (2015), "Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance," *Annual Review of Psychology, 66,* 519-545.

Tversky, A., and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases," *Science,* 185(4157), 1124-1131.

University of Essex, Institute for Social and Economic Research, NatCen Social Research, Kantar Public. (2019), Understanding Society: Waves 1-9, 2009-2018 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 12th Edition. UK Data Service. SN: 6614.

Verbeke, G. and Molenberghs, G. (2000), "Inference for the Marginal Model," in *Linear Mixed Models for Longitudinal Data,* New York, NY: Springer, pp. 55–76.

West, B. T. (2013), "An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 211-225.

West, B. T., and Blom, A. G. (2017), "Explaining Interviewer Effects: A Research Synthesis," *Journal of Survey Statistics and Methodology*, 5(2), 175-211.

West, B. T. and Elliott, M. R. (2014), "Frequentist and Bayesian Approaches for Comparing Interviewer Variance Components in Two Groups of Survey Interviewers," S*urvey Methodology*, 40, 163-188.

West, B. T., and Kreuter, F. (2013), "Factors Affecting the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth," *Public Opinion Quarterly*, 77(2), 522-548.

West, B. T., and Kreuter, F. (2015), "A Practical Technique for Improving the Accuracy of Interviewer Observations of Respondent Characteristics," *Field Methods*, 27(2), 144-162.

West, B. T., Kreuter, F., and Trappmann, M. (2014), "Is the Collection of Interviewer Observations Worthwhile in an Economic Panel Survey? New Evidence from the German Labor Market and Social Security (PASS) Study," *Journal of Survey Statistics and Methodology*, 2(2), 159-181.

West, B. T., and Li, D. (2019), "Sources of Variance in the Accuracy of Interviewer Observations," *Sociological Methods & Research,* 48(3), 485-533.

West, B. T., Yan, T., Kreuter, F., Josten, M., and Schroeder, H. (2020). "Examining the Utility of Interviewer Observations on the Survey Response Process," In (K. Olson et al., Ed.) *Interviewer Effects from a Total Survey Error Perspective*, Boca Raton: CRC Press, 107-22.