Journal of Survey Statistics and Methodology (2021) ***00****, 1–20*

INTERVIEWER EFFECTS IN LIVE VIDEO AND PRERECORDED VIDEO INTERVIEWING

BRADY T. WEST*
AI RENE ONG
FREDERICK G. CONRAD
MICHAEL F. SCHOBER
KALLAN M. LARSEN
ANDREW L. HUPP

Live video (LV) communication tools (e.g., Zoom) have the potential to provide survey researchers with many of the benefits of in-person interviewing, while also greatly reducing data collection costs, given that interviewers do not need to travel and make in-person visits to sampled households. The COVID-19 pandemic has exposed the vulnerability of in-person data collection to public health crises, forcing survey researchers to explore remote data collection modes—such as LV interviewing—that seem likely to yield high-quality data without in-person interaction. Given the potential benefits of these technologies, the operational and methodological aspects of video interviewing have started to receive research attention from survey methodologists. Although it is

Brady T. West is a Research Associate Professor with the Survey Research Center, Institute for Social Research, University of Michigan-Ann Arbor, 426 Thompson Street, Ann Arbor, MI 48106-1248, USA. Ai Rene Ong is a PhD Candidate with the Michigan Program in Survey and Data Science, University of Michigan-Ann Arbor, 426 Thompson Street, Ann Arbor, MI 48106-1248, USA. Frederick G. Conrad is a Research Professor with the Survey Research Center, Institute for Social Research, University of Michigan-Ann Arbor, 426 Thompson Street, Ann Arbor, MI 48106-1248, USA. Michael F. Schober is Professor of Psychology with the New School for Social Research. Kallan M. Larsen is a Research Area Specialist Intermediate and Andrew L. Hupp is a Survey Specialist Senior III with the Survey Research Center, Institute for Social Research, University of Michigan-Ann Arbor, 426 Thompson Street, Ann Arbor, MI 48106-1248, USA.

The authors acknowledge financial support for this research from NSF awards #SES-1825113 and #SES-1825194, and NIH awards P30 AG012846 and UL1TR002240. This study's design and analysis were preregistered with the Open Science Framework. The detailed preregistration can be found at https://osf.io/9jusx.

*Address correspondence to Brady T. West, Survey Research Center, Institute for Social Research, University of Michigan-Ann Arbor, 426 Thompson Street, Ann Arbor, MI 48106-1248, USA; E-mail: bwest@umich.edu.

doi: 10.1093/jssam/smab040

remote, video interviewing still involves respondent–interviewer interaction that introduces the possibility of interviewer effects. No research to date has evaluated this potential threat to the quality of the data collected in video interviews. This research note presents an evaluation of interviewer effects in a recent experimental study of alternative approaches to video interviewing including both LV interviewing and the use of prerecorded videos of the same interviewers asking questions embedded in a web survey ("prerecorded video" interviewing). We find little evidence of significant interviewer effects when using these two approaches, which is a promising result. We also find that when interviewer effects were present, they tended to be slightly larger in the LV approach as would be expected in light of its being an interactive approach. We conclude with a discussion of the implications of these findings for future research using video interviewing.

KEYWORDS: Interviewer effects; Video interviewing; Web surveys.

Statement of Significance

This research note presents an analysis of interviewer effects from an experimental study of alternative approaches to using video communication technologies to collect survey data. This work is especially important given the significant restrictions on survey research that have been introduced by the global pandemic, and the potential of LV interviewing to provide data quality benefits similar to face-to-face interviewing without putting interviewers at risk. The results provide good news about the use of this approach and at the same time introduce the importance of monitoring interviewer effects when using these types of video interviewing approaches.

1. INTRODUCTION

Live video (LV) communication tools (e.g., Zoom) have the potential to provide survey researchers with many of the benefits of in-person interviewing while also greatly reducing data collection costs because interviewers do not need to travel and make in-person visits to sampled households. The COVID-19 pandemic has exposed the vulnerability of in-person data collection to public health crises, forcing survey researchers to explore remote data collection approaches—such as LV interviewing—that seem likely to yield high-quality data without in-person interaction. Given the potential benefits of video communication technologies, the operational and methodological aspects of video interviewing have started to receive research attention from survey

methodologists (Endres and Hillygus 2019; Conrad, Schober, Hupp, West, Larsen, et al. 2020; Schober, Conrad, Hupp, Larsen, Ong, et al. 2020). However, we know of no survey organizations that routinely use LV in production surveys; more methodological research on LV interviewing is necessary before this becomes a routine survey research practice.

One recent methodological study of LV interviewing focused on experimental comparisons of data quality with in-person interviewing (Endres and Hillygus 2019). Comparing LV and face-to-face interviewing to self-administration, these authors found benefits of LV interviewing quite similar to those of in-person interviewing: less nondifferentiation, less item nonresponse, fewer do not know responses, and more participant satisfaction compared to a conventional web survey. Another recent methodological study performed an experimental comparison between a different type of video-based interviewing approach, namely prerecorded videos (PVs) of questions being asked by an interviewer in a web survey, and a traditional web survey (Haan, Ongena, Vannieuwenhuyze, and de Glopper 2017). These authors found few differences in disclosure of sensitive information and respondent engagement. They suggest that their "video-web" approach did not improve data quality because it lacked responsivity, using prerecorded questions. Although these studies have provided insights into the effects that LV interviewing or PVs of interviewers asking questions might have on data quality, they did not evaluate a potential threat to data quality associated with interviewer-administered modes of data collection: interviewer effects.

LV interviewing involves respondent–interviewer interaction, which introduces the possibility of interviewer effects on collected survey responses by allowing interviewers to question and probe responses differently from one another. PVs of different interviewers asking questions feature the exact same delivery of a question across all respondents assigned to a particular interviewer and do not involve respondent–interviewer interaction. This reduces the possibility of interviewer effects due to responsive nonverbal behaviors, verbal probing and clarification, and other tailored behaviors that are possible in LV interviewing. However, interviewer effects may still arise in surveys using PVs due to observable or inferable interviewer characteristics, such as their gender, age, or race/ethnicity (Krysan and Couper 2003; Conrad, Schober, Nielsen and Reichert 2020), or the magnification of differences in question delivery between interviewers due to the identical presentation each time a question is asked. Different interviewers conducting LV interviews or recording videos of themselves asking survey questions may therefore introduce the same types of effects on survey measures that have been reported in myriad prior studies of in-person interactions (West and Blom 2017). The resulting variability among interviewers in the distributions of the responses collected reduces the efficiency of survey estimates, lowering effective sample sizes and statistical power in a manner similar to cluster sampling (Elliott and West 2015).

Table 1 summarizes theoretical mechanisms that could introduce interviewer effects for each of these two approaches. Based on these mechanisms, we hypothesize that LV interviewing will more often introduce interviewer effects, especially for more sensitive or complex items. The lack of studies evaluating interviewer effects in LV or PV interviews precludes hypotheses based on empirical evidence.

This research note presents an evaluation of interviewer effects in a recent experimental study of LV and PV interviewing. While one should study interviewer effects from a Total Survey Error perspective (West and Blom 2017) and prior work has suggested that interviewer effects may arise from variance among interviewers in the types of respondents recruited (e.g., West and Olson 2010), we focus exclusively on the variability in survey measures among interviewers because the interviewers were not responsible for recruitment. We seek answers to the following two research questions:

(1) How much interviewer variance arises when using each approach, in particular for responses to sensitive questions and measures of satisficing?
(2) Does the interviewer variance differ significantly across these two approaches?

2. METHODS

2.1 Study Overview

We implemented a randomized experiment in an original data collection that took place between August 2019 and March 2020 (with most of the data collection occurring before the COVID-19 pandemic). This study sought to evaluate three different data collection approaches: LV, PV, and a conventional (text-only) web survey. We focus on the LV and PV approaches exclusively as the web approach did not involve interviewers.

A nonprobability sample of individuals ages 18 and above was recruited to participate in the study via one of two online research services: CloudResearch (cloudresearch.com) or the University of Michigan online Health Research program (umhealthresearch.org). Study participants were therefore volunteer online panelists, and all interested participants were randomly assigned to one of the three approaches upon agreeing to participate. All individuals were offered a $5 Amazon gift code for participating, and participants randomly assigned to the LV approach were offered an additional $15, conditional on completing their interview along with an online debriefing offered to both LV and PV participants. About one-third of the participants completed the survey using a smartphone across the two approaches (31.9 percent in LV and 40.3 percent in PV).

Table 1. Theoretical Mechanisms That Could Introduce Interviewer Effects on Survey Measurement in Live Video and Prerecorded Video Interviewing (Yes or No), Considering Interviewer Behaviors, Observable Interviewer Characteristics, and Question Types

	Theoretical mechanism	Live video interviewing	Prerecorded video interviewing	Examples of relevant literature[a]
Interviewer Behaviors	Responsive nonverbal behaviors (e.g., nodding, gazing, laughing)	Yes	No	Fuchs (2002), Holbrook, Green, and Krosnick (2003), and Welles, Sun, and Miller (in press)
	Probing	Yes	No	Mangione et al. (1992)
	Not reading questions exactly as worded	Yes	No	Houtkoop-Steenstra (1995) and Haan, Ongena, and Huiskes (2013)
	Attempting to improve respondent comprehension with conversational/ flexible styles	Yes	No	Conrad and Schober (2000), Schober and Conrad (1997), West, Conrad, Kreuter, and Mittereder (2018), and van der Zouwen, Dijkstra, and Smit (1991)
	Attempts to establish rapport	Yes	No	Goudy and Potter (1975) and Garbarski, Schaeffer, and Dykema (2016)
	Speed asking questions/alternative delivery styles	Yes	Yes	Olson and Peytchev (2007) and Olson and Smyth (2015)
	Voice characteristics	Yes	Yes	Charoenruk and Olson (2018)
	Fixed/identical question presentation by a given interviewer	No	Yes	Haan et al. (2017)
	Inability to adjust question presentation and subsequent dialogue responsively	No	Yes	Conrad and Schober (2000), Schober and Conrad (1997), and Haan et al. (2017)

Continued

Table 1. *Continued*

	Theoretical mechanism	Live video interviewing	Prerecorded video interviewing	Examples of relevant literature[a]
Observed or Inferred Interviewer Characteristics	Gender	Yes	Yes	Kane and Macaulay (1993) and Liu and Stainback (2013)
	Age	Yes	Yes	Sudman and Bradburn (1974), Collins and Butcher (1982), and Brüderl, Huyer-May, and Schmiedeberg (2013)
	Race/ethnicity	Yes	Yes	Schuman and Converse (1971), Davis and Silver (2003), and Krysan and Couper (2003)
	Physical appearance	Yes	Yes	Bateman and Mawby (2004) and Eisinga, Te Grotenhuis, Larsen, Pelzer, and van Strien (2011)
Question Types	Sensitive, complex, attitudinal, open-ended (more opportunities for live interviewers to react, assist, or intervene)	Yes	No	Groves and Magilavy (1986), Billiet and Loosveldt (1988), Schnell and Kreuter (2005), and O'Muircheartaigh and Campanelli (1998)
	Factual	No	No	Kish (1962)
	Show cards	Yes	No	Dijkstra and Ongena (2006) and Jäckle, Roberts, and Lynn (2010)

NOTE.— [a]These are not meant to be comprehensive lists of all studies on a particular source of interviewer effects; see West and Blom (2017) and Davis et al. (2010) for more comprehensive overviews.

2.2 Variables of Interest

The survey featured 36 questions from major US social surveys and other methodological studies (e.g., the General Social Survey; see the Supplementary data online). Ten questions required an open numerical response (e.g., hours watching television per day), nine presented categorical responses (e.g., Likert-type agree/disagree responses to statements like "It is important to maintain a healthy diet"), and 17 items (statements) comprised three batteries, offering the same options for each statement in a battery (e.g., agree/disagree). The batteries were implemented as a series of individual items (i.e., not as grids in PV) for comparability across the approaches, and responses to these items were dichotomized as neutral (e.g., neither agree nor disagree) versus non-neutral (e.g., somewhat disagree) about particular topics to allow for modeling the probability of a substantive (i.e., non-neutral) response.

The questions were organized by topic and presented in order from the least sensitive topic to the most sensitive topic. These sensitivity ratings were determined by a prior norming study in which respondents rated how uncomfortable they thought people would be answering each question and selecting each response option. This norming study was based on approaches used in prior assessments of question sensitivity (e.g., Fail, Schober, and Conrad 2021; Feuer, Fail, and Schober 2019). Questions about credit card balance, attending religious services, volunteer work, helping the homeless, participating in local elections, sex frequency, and frequency of watching pornography were all considered more sensitive (based on the norming study).

We also analyzed three measures of data quality: *disclosure*, measured by the average sensitivity of the response options selected by a participant (e.g., because 24 percent of the respondents in the sensitivity rating study rated reporting having only one sex partner in the past 12 months to be "Very Uncomfortable" or "Somewhat Uncomfortable," the sensitivity of that response is 0.24); *rounding*, measured by the number of numerical answers divisible by 10 (e.g., ten versus eleven movies seen in the last month); and *near straightlining*, measured by the participant selecting the same response option for all or all but one of the items in *any* of the three batteries. Initial analyses have shown that LV interviewing produced significantly less disclosure, a significantly higher proportion of rounded answers, and significantly less straightlining than the PV approach (Conrad et al. 2020); whether these measures of data quality tend to vary across interviewers in each of these two approaches remains an open question.

LV interviewing (279 respondents) was implemented as synchronous two-way video using the BlueJeans video platform (https://www.bluejeans.com/), through which the interviewer administered a questionnaire programmed using Blaise and displayed on the interviewer's screen below the BlueJeans window.

Participants randomly assigned to LV interviewing scheduled appointments with one of eight interviewers from the University of Michigan Survey Research Center. The same eight interviewers were video recorded asking the same 36 survey questions, and these recordings were embedded in a Blaise 5 web survey comprising the PV approach. For each PV respondent, one of the eight interviewers asked all 36 questions. The response options did not appear until the recording played in its entirety; the video recordings autoplayed except on mobile devices.

2.3 Assignment of Participants to Interviewers

At the beginning of data collection, we required participants from CloudResearch to schedule interviews with only one randomly assigned interviewer, attempting to implement interpenetrated sample assignment for the purpose of estimating interviewer effects. When this approach resulted in many missed and unscheduled appointments, we allowed Michigan Health Research respondents to schedule appointments with *any* interviewer at *any* of the available interview slots. In instances where multiple interviewers were available at the same time, the scheduling software we used, Calendly, assigned interviewers based on who had completed the fewest interviews thus far. Because this process did not result in true random assignment of participants to interviewers, we reviewed the distributions of demographic measures (sex, age, race, and education) of the participants assigned to each interviewer on a weekly basis. When notable imbalances across the interviewers were identified, we adjusted which interviewers' appointment slots participants from certain demographic groups could view for scheduling.

Table 2 shows the final distributions of selected demographic characteristics for each of the eight interviewers. This table shows that outside of the gender distributions for LV interviewing, the eight interviewers ultimately interviewed participants with similar demographic features. The same was true for the PV interviews.

2.4 Analytic Methods

For each of the thirty-six survey variables and the data quality measures, we fit the following mixed-effects model, where i indexes respondents, j indexes interviewers, LV and PV are indicators of assignment to the two approaches, and $g()$ is a link function appropriate for a given type of dependent variable (e.g., the logit link for a binary dependent variable):

$$g(y_{ij}) = \beta_0 + \beta_1 LV_{ij} + u_{1j} LV_{ij} + u_{2j} PV_{ij} + \varepsilon_{ij}, \quad (1)$$

Table 2. Comparisons of Eight Interviewers in Terms of Demographic Distributions of the Interviewed Participants under the LV and PV Approaches

Live video (LV) interviewing						Prerecorded video (PV) interviewing					
Iwer	Under 65	White	>HS	Male	Total	Iwer	Under 65	White	>HS	Male	Total
1	0.70	0.85	0.85	0.56	27	1	0.60	0.74	0.69	0.37	35
2	0.74	0.77	0.74	0.38	39	2	0.72	0.83	0.75	0.47	36
3	0.90	0.78	0.83	0.40	40	3	0.69	0.83	0.64	0.33	42
4	0.67	0.90	0.82	0.33	39	4	0.72	0.74	0.70	0.33	43
5	0.79	0.91	0.91	0.24	33	5	0.65	0.69	0.55	0.33	49
6	0.78	0.69	0.84	0.41	32	6	0.67	0.64	0.80	0.33	45
7	0.66	0.76	0.90	0.31	29	7	0.68	0.70	0.72	0.40	50
8	0.68	0.78	0.80	0.30	40	8	0.73	0.84	0.69	0.42	45
TOTAL	0.74	0.80	0.83	0.36	279	**TOTAL**	0.68	0.75	0.69	0.37	345

NOTE.—Iwer, interviewer ID; >HS, greater than high school education (the numbers indicate proportions of respondents with certain characteristics and total counts of surveys completed for each interviewer).

where $\begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right)$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ for continuous items.

Given the small number of interviewers available for this study, we used penalized quasi-likelihood (as implemented in the GLIMMIX procedure of SAS Version 9.4) to fit logistic regression models of the form in (1) to each of the binary measures (excluding residual terms), and restricted maximum likelihood to fit linear regression models of the same form to the numeric measures. These estimation procedures reduce the bias in estimates of variance components that can arise with small samples of higher-level clusters (McNeish and Stapleton 2016). We specified the model in (1) for each of our measures because each interviewer ultimately produced data for both approaches. This model, which varies from more traditional multilevel models for studying interviewer effects, allows for unique interviewer variance depending on the approach used (LV or PV), and enables us to answer our second research question. Furthermore, the model allows the random interviewer effects associated with each approach to covary, enabling an assessment of the correlation of the two random effects for each interviewer across the two approaches.

For our first research question, we initially tested the significance of the covariance of the two random effects by fitting the model in (1), referred to as Model 1, along with a reduced model with the covariance of the random effects constrained to be zero (Model 2), and performing a likelihood ratio test of the null hypothesis that the covariance was equal to zero. We note that estimating the covariance of the random effects based on Model 1 was only feasible for outcome measures (responses) that presented evidence of nonzero variance in the random interviewer effects for both approaches. We then tested the significance of the interviewer variance components for each approach by using an appropriate mixture-based likelihood ratio test of the null hypothesis that a given variance component is equal to zero (e.g., West and Olson 2010). For our second research question, we tested whether the interviewer variance components were equal by fitting a reduced model (Model 3) with the interviewer variance constrained to be equal for each mode, and then performing a likelihood ratio test of the null hypothesis that the two variance components were equal (West and Elliott 2014).

Given the relatively low statistical power of this study to detect interviewer variance and differences in interviewer variance between the approaches (with only eight interviewers), we supplemented our analyses with descriptive estimates of intra-interviewer correlations (IICs) for each item measured in each approach (based on ratios of the estimated interviewer variance component to the total variance for each item; see West et al. [2018] for details). Because multilevel models constrain estimates of IICs to be greater than or equal to zero, we also estimated the IICs using the ANOVA-based method outlined by Kish (1962) and frequently employed in prior studies of interviewer effects

Table 3. Estimated IICs of the Outcome Measures, by Approach

Variable (sensitivity rating)	Type of variable	IICs: REML/PQL (Kish)	
		Live video	Prerecorded video
No. of hours of television watched/day (0.07)	Numeric	0 (−0.013)	0.025 (0.012)
No. of movies watched in theater, PY (0.06)	Numeric	0.019 (0.005)	0 (−0.009)
No. of movies watched, PM (0.04)	Numeric	0 (−0.017)	0 (0.002)
Times eaten in restaurants, PM (0.06)	Numeric	0 (−0.012)	0 (0)
Times eaten spicy food, PM (0.05)	Numeric	0 (−0.007)	0 (−0.005)
Times shopped in grocery store, PM (0.03)	Numeric	0.006 (0.003)	0 (−0.013)
Fluid oz. of drinking water yesterday (0.04)	Numeric	0.009 (0.003)	0.005 (0.016)
Food battery (Neutral versus Substantive)			
Avoiding fast food (0.09)	Binary	0.022 (−0.013)	0 (0.016)
Maintaining healthy diet (0.07)	Binary	0.046 (0.005)	0 (0.002)
Monitoring cholesterol (0.10)	Binary	0 (−0.022)	0.005 (0.005)
Emphasizing taste (0.05)	Binary	0.050 (0.029)	0 (−0.002)
Attention to packaging (0.02)	Binary	0 (−0.016)	0 (−0.008)
Limiting red meat diet (0.09)	Binary	0 (−0.008)	0.033 (0.022)
Balancing food groups (0.04)	Binary	0.024 (0.010)	0 (−0.001)
Money battery (Neutral versus Substantive)			
Imp. to have nice things (0.08)	Binary	0 (−0.013)	0 (−0.013)
Can buy anything (0.06)	Binary	0.019 (−0.015)	0.016 (0.004)
Afford to buy more things (0.07)	Binary	0 (−0.001)	0 (−0.014)
Bothers when cannot afford (0.13)	Binary	0 (−0.003)	0 (−0.006)
Money can buy happiness (0.21)	Binary	0.063 (0.010)	0.002 (0.001)
Things I own give pleasure (0.09)	Binary	0.007 (0.005)	0.021 (0.017)
Pay off CC balance (never/ever) (0.18)	Binary	0 (−0.027)	0.012 (0.006)
Frequency attending religious services (seldom/never/other) (0.11)	Binary	0.007 (0.008)	0.011 (0.009)
Freq. offer bus seat to stranger (never/ever) (0.09)	Binary	0 (0.017)	0.012 (0.009)
Freq. volunteer work (never/ever) (0.12)	Binary	0.084** (0.060)	0 (−0.009)
Freq. help homeless (never/ever) (0.14)	Binary	0.026* (0.038)	0.017 (0.013)

Continued

Table 3. *Continued*

Variable (sensitivity rating)	Type of variable	IICs: REML/PQL (Kish)	
		Live video	Prerecorded video
Sports battery (Neutral versus Substantive)			
Too much sports (0.06)	Binary	0 (−0.013)	0.030 (0.020)
Brings races together (0.09)	Binary	0 (−0.003)	0 (−0.014)
Intern. sports create tension (0.11)	Binary	0 (−0.013)	0.029 (0.022)
Gov't should spend more (0.21)	Binary	0 (−0.017)	0.037 (0.020)
Frequency of voting in local elections (never versus ever) (0.11)	Binary	0.003 (0.001)	0 (−0.003)
Frequency of sex in past 12 months (never versus ever) (0.45)	Binary	0.013 (0.010)	0 (−0.013)
No. of sex partners in past 12 months (0.53)	Numeric	0 (−0.007)	0.015 (0.005)
No. of sex partners in past 12 months (none versus any) (0.53)	Binary	0.005 (0.006)	0 (−0.002)
No. of female sex partners (0.54)	Numeric	0.005 (−0.008)	0 (−0.008)
No. of male sex partners (0.55)	Numeric	0.003 (0.026)	0.003 (−0.004)
Gender of sex partner (both gender/same gender/none versus heterosexual) (0.45)	Binary	0 (−0.005)	0.031 (0.020)
Frequency of visiting sexually explicit website (never versus ever) (0.59)	Binary	0 (0.001)	0 (−0.003)
Data quality variables			
Rounding (sum out of seven)	Numeric	0.001 (0.002)	0 (−0.007)
Average response sensitivity	Numeric	0.012 (0.002)	0.008 (0.018)
Near straightline (ever versus never)	Binary	0 (0)	0.024 (0.013)
Near straightline (Food battery)	Binary	0 (−0.003)	0.152 (0.010)[a]
Near straightline (Sports battery)	Binary	0.084 (0.021)	0.026 (0.009)
Near straightline (Money battery)	Binary	0 (−0.019)	0 (−0.013)
Mean IIC		0.013 (0.0002)	0.012 (0.003)
Number of zero/negative IICs		22/22	22/19

NOTE.— *$p < .1$, **$p < .01$.

REML, restricted maximum likelihood; PQL, penalized quasi-likelihood; PM, past month; PY, past year. IICs estimated using the ANOVA-based method (Kish 1962), which allows for negative estimates of the IICs, are in parentheses. Estimates of 0 indicate variance components that were very close to zero (i.e., less than 0.001) when using the multilevel modeling methods.

[a] There were only seven cases of near straightlining on the food battery, and two of them were interviewed by one interviewer in the PV approach; this estimated IIC under multilevel modeling was therefore not considered reliable.

(e.g., Groves and Magilavy 1986). We also generated plots of the two predicted values (EBLUPs) of the random effects for each interviewer (corresponding to the two approaches) to visually examine the correlations of the two random effects for each interviewer when we found evidence of interviewer variance.

3. RESULTS

3.1 Research Question 1: How Much Interviewer Variance Arises When Using Each Approach?

Table 3 presents estimates of the IICs for each measure by approach and indications of whether the interviewer variance for a given approach was found to be significantly greater than zero. In general, the majority of the estimated IICs were very small for both approaches, with twenty-two out of forty-three measures computed as zero for each approach when using multilevel modeling, and 22 (LV) and 19 (PV) computed as negative when using Kish's ANOVA-based method. Table 4 shows that the means and ranges of the IICs are largely in line with similar descriptive summaries reported in prior studies of interviewer effects on multiple survey items. We see no evidence of the mean IIC for the LV approach being substantially higher than the mean IIC for PV or the means reported in prior studies, and the ranges of the estimated IICs for both approaches are actually somewhat smaller than observed in the prior literature.

We found no evidence of significant interviewer variance in the PV approach. Five of these estimated IICs were at or above 0.02 when using the multilevel modeling and ANOVA-based methods to compute them. Such IICs would generally be considered "large" and would likely impact the precision of survey estimates in a negative fashion (Groves and Magilavy 1986; O'Muircheartaigh and Campanelli 1998). We also found five relatively large IICs in the LV approach, using the same criterion; two were significantly greater than zero at the 0.10 or 0.01 level, despite our reduced power.

The survey items that seemed to introduce larger interviewer effects in the LV interviews included a binary indicator of ever performing volunteer work (IIC = 0.084, $p < .01$) and a binary indicator of ever helping the homeless (IIC = 0.026, $p < .10$). We also found a relatively large IIC in LV (IIC = 0.084) for the measure of straightlining based on the battery of items on sports. We did not find any evidence of a correlation between the rated sensitivity of the item and the magnitude of the estimated IIC in either approach.

3.2 Research Question 2: Does the Interviewer Variance Differ across the Two Approaches?

Table 5 summarizes the model fitting and testing results for selected outcome measures with notable differences in the estimated IICs across the approaches

Table 4. Overall Means and Ranges of Estimated IICs of the Outcome Measures by Approach with Similar Results from Other Prior Studies of Interviewer Effects on Multiple Survey Items for Context

	LV: MLM/Kish[a]	PV: MLM/Kish[a]	Kish (1962)	Feather (1973)	O'Muircheartaigh and Marckwardt (1980)	Tucker (1983)	Groves and Magilavy (1986)[b]	O'Muircheartaigh and Campanelli (1998)[c]	West et al. (2018): Standardized interviewing	West et al. (2018): Conversational interviewing
Mean	0.013/0.0002	0.012/0.003	0.02	0.006	0.07	0.004	0.009	0.015	0.018	0.030
Range	(0.000–0.084)/ (−0.027–0.060)	(0.000–0.152/ (−0.014–0.022)	(−0.03–0.09)	(−0.01–0.03)	(0.00–0.19)	(−0.003–0.008)	(−0.042–0.171)	(−0.296–0.216)	(0.001–0.364)	(0.001–0.401)
Topics	Various	Various	Job attitudes	Health	Fertility	Political/ social attitudes	Various	Various	Various	Various

NOTE.— [a]MLM, multilevel modeling methods outlined in section 2.4; Kish = Kish's ANOVA-based method outlined in section 2.4, allowing for negative estimates of ICCs.

[b] Based on aggregation of the data in Table 1 from Groves and Magilavy (1986).

[c] Based on aggregation of the data from the literature review in Table 1 (O'Muircheartaigh and Campanelli 1998). Based on their own study, these authors report that 80 percent of the IICs computed were less than 0.02; see figure 1 in O'Muircheartaigh and Campanelli (1998).

Table 5. Summaries of Model Fitting Results (−2 Log-Likelihood Values and Results of Likelihood Ratio Tests of Equality Of Interviewer Variance Components across the Approaches) for Items Presenting Evidence of Differences in Interviewer Variance across the Two Approaches

	Maintaining Healthy Diet	Emphasizing Taste	Red Meat Diet	Money Buys Happiness	Volunteer Work	Help Homeless	Too Much Sports
Model 2: −2 log-likelihood	639.58	697.65	639.58	442.46	772.29	841.31	772.87
Model 3: −2 log-likelihood	640.83	699.05	640.83	442.87	777.82	840.34	773.85
Difference: Model 3 - Model 2	1.25	1.40	1.25	0.41	5.53*	−0.97[a]	0.98

	International sports	Gov't spend more	Gender of sex partner	Near straightlining: sports battery
Model 2: −2 log-likelihood	665.82	551.30	805.84	459.71
Model 3: −2 log-likelihood	666.70	552.38	806.87	458.25
Difference: Model 3—Model 2	0.88	1.08	1.03	−1.46[a]

Notes:—*$p < .05$. Model 2: unique variance components for the LV and PV approaches. Model 3: common interviewer variance for both approaches. Per Kim, Choi, and Emery (2013), PQL methods, while important for reducing bias in estimates of variance components due to small sample sizes, are not suitable for comparing model fit. All likelihood ratio tests therefore employed adaptive Gaussian quadrature for purposes of the likelihood ratio testing (Kim et al. 2013).

[a] The −2 log-likelihood was in fact lower for the model with the interviewer variance component constrained to be equal (Model 3), suggesting a (slightly) better fit.

(from table 3). We were largely unable to reliably estimate correlations of the random interviewer effects in the two approaches, given that the majority of the items presented evidence of no interviewer variance in one of the two approaches. As a result, we focus primarily on the tests of equality in the interviewer variance between the two approaches.

We only found evidence of the interviewer variance in LV being significantly larger than in the PV approach for one of the items (volunteer work). We visualize these differences in interviewer variance for three of the items in figure 1, which presents predictions of the interviewer effects for the eight interviewers for each of these items in each approach. The larger variance in the black dots representing predictions of the random effects in LV for the eight interviewers is evident, as is the general lack of correlation between the predicted random effects.

Closer inspection of the predicted random interviewer effects in figure 1 indicates that the second interviewer had relatively large positive effects on the responses to each item in LV. This interviewer completed a total of thirty-nine LV interviews (table 1). Both LV and PV respondents were asked to complete an online debriefing after the survey, which was designed for quality control purposes and to provide more insight into possible differences in response distributions between the approaches. An analysis of selected debriefing items indicated that fewer respondents interviewed by the second interviewer reported that they were "Very Satisfied" with the interview compared to the other interviewers. Respondents interviewed by this individual also reported the lowest comfort with the interviewer (only 54 percent said they were "Very Comfortable" with the interviewer). Two of this interviewer's respondents also

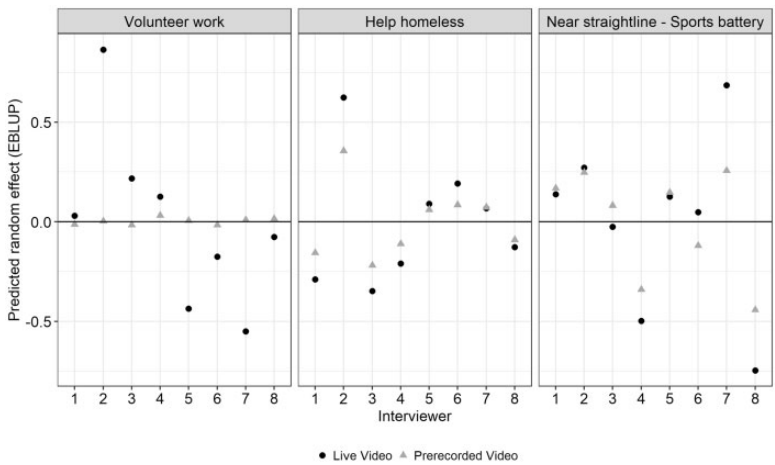

Figure 1. Predicted Values of Random Interviewer Effects for Each Interviewer in Each Approach for Selected Survey Questions.

reported that they felt that they could not answer honestly. The behavior of the interviewer that produced these respondent reactions might have affected the answers of the respondents (e.g., more socially desirable responses), which may have introduced higher interviewer variance on selected items relative to the PV approach.

4. DISCUSSION

We found that interviewer effects in LV interviewing tended to be rare and did not arise at substantially higher rates when compared to the PV approach. Out of 43 variables analyzed, including six measures of data quality, we only found evidence of IICs greater than 0.02 for five items in the LV interviews (two of which had significant interviewer variance components) and five items in the PV approach (none of which had significant variance components). When comparing the two approaches, we did find significantly higher interviewer variance for the item on frequency of volunteering in LV interviews, suggesting that this approach may introduce opportunities for larger interviewer effects on selected items. While this lack of significant findings may have been due to relatively low statistical power, given that we only studied eight interviewers, descriptive analyses of the estimated IICs for the two approaches using multiple estimation methods suggested that the large majority of the IICs were small (<0.02), meaning that a higher-powered study may have led to the same results.

This is generally good news for LV interviewing. Although there is no obvious reason why video mediation would increase standardization compared to in-person interviewing, the fact that interviewer effects were generally no greater in LV than PV could mean that the interaction was as standardized in LV as in PV. This could be the case if video mediation reduced the amount of probing (Mangione, Fowler, and Louis 1992), but there is not a clear theoretical reason why this would have been the case. Without having video-recorded LV interviews, we cannot test this possibility. Because LV interviewing mimics in-person interviewing in several important ways, potential interviewer effects introduced by LV in different contexts need monitoring as in in-person interviewing. The same is true for the PV approach, which may produce significant data quality benefits relative to a text-only web survey at far lower cost than LV (Conrad et al. 2020). Future experiments evaluating the LV approach on a larger scale could use our results for power analysis (West 2020), compare interviewer effects in LV interviewing and in-person interviewing, and carefully examine any differences between LV and in-person interviewing in terms of the factors introducing interviewer effects. The possibility that interviewer effects introduced by LV or PV may also vary by the device used (e.g., computer versus mobile device) also warrants future examination.

Analyses of *explanations* for any interviewer effects observed will be crucial for future work in this area (West and Blom 2017). These explanatory factors may include observable features of the interviewers (which we could not access in this study), behaviors during the interviews (e.g., in debriefings, some interviewers expressed frustration with technical difficulties and coaching people through the use of a mobile phone for the interview), and nonverbal expressions (e.g., some interviewers said that trying to keep a "poker face" when hearing responses to sensitive questions was difficult). We did not have the resources to record and analyze the recordings of the LV interviews, but future studies using more interviewers could code a subsample of recorded LV interviews and study potential correlates of any interviewer effects observed, such as respondent comfort answering questions.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam

REFERENCES

Bateman, I. J., and J. Mawby (2004), "First Impressions Count: Interviewer Appearance and Information Effects in Stated Preference Studies," *Ecological Economics*, 49, 47–55.

Billiet, J., and G. Loosveldt (1988), "Improvement of the Quality of Responses to Factual Survey Questions by Interviewer Training," *Public Opinion Quarterly*, 52, 190–211.

Brüderl, J., B. Huyer-May, and C. Schmiedeberg (2013), "Interviewer Behavior and the Quality of Social Network Data," in *Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, pp. 147–160, Frankfurt: Peter Lang Academic Research.

Charoenruk, N., and K. Olson (2018), "Do Listeners Perceive Interviewers' Attributes from Their Voices and Do Perceptions Differ by Question Type?," *Field Methods*, 30, 312–328.

Collins, M., and B. Butcher (1982), "Interviewer and Clustering Effects in an Attitude Survey," *Journal of the Market Research Society*, 25, 39–58.

Conrad, F. G., and M. F. Schober (2000), "Clarifying Question Meaning in a Household Telephone Survey," *Public Opinion Quarterly*, 64, 1–28.

Conrad, F. G., M. F. Schober, A. L. Hupp, B. T. West, K. Larsen, A. R. Ong, and T. Wang (2020), "Interviewers, Video, and Survey Data Collection," paper presented at the 2020 AAPOR Virtual Conference, June 12, 2020.

Conrad, F. G., M. F. Schober, D. Nielsen, and H. Reichert (2020), "Social Identities of Virtual Interviewers and Their Impact on Survey Responses," in *Interviewer Effects from a Total Survey Error Perspective*, eds. K. Olson, J.D. Smyth, J. Dykema, A.L. Holbrook, F. Kreuter, and B.T. West, pp. 149–164. Boca Raton, FL: CRC Press.

Davis, R. E., M. P. Couper, N. K. Janz, C. H. Caldwell, and K. Resnicow (2010), "Interviewer Effects in Public Health Surveys," *Health Education Research*, 25, 14–26.

Davis, D. W., and B. D. Silver (2003), "Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge," *American Journal of Political Science*, 47, 33–45.

Dijkstra, W., and Y. Ongena (2006), "Question-Answer Sequences in Survey-Interviews," *Quality and Quantity*, 40, 983–1011.

Eisinga, R., M. Te Grotenhuis, J. K. Larsen, B. Pelzer, and T. van Strien (2011), "BMI of Interviewer Effects," *International Journal of Public Opinion Research*, 23, 530–543.

Elliott, M. R., and B. T. West (2015), "Clustering by Interviewer": A Source of Variance That is Unaccounted for in Single-Stage Health Surveys," *American Journal of Epidemiology*, 182, 118–126.

Endres, K., and D. S. Hillygus (2019), "A Future for Video Interviewing? An Experimental Assessment of Video Mode Compared to Face-to-Face and Online Self-Complete Interviewing," Initiative on Survey Methodology, Duke University, May 17, 2019.

Fail, S., M. F. Schober, and F. G. Conrad (2021), "The Time It Takes to Reveal Embarrassing Information in a Mobile Phone Survey," *International Journal of Social Research Methodology*, 24, 249–364.

Feather, J. (1973), "*A Study of Interviewer Variance*," Department of Social and Preventive Medicine, Canada: University of Saskatchewan.

Feuer, S., S. Fail, and M. F. Schober (2019), "Empirically Assessing Survey Question and Response Sensitivity," paper presented at the 74th Annual Conference of the American Association for Public Opinion Research, Toronto, Canada.

Fuchs, M. (2002), "The Impact of Technology on Interaction in Computer-Assisted Interviews," Chapter 20 in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, ed. D. W. Maynard, et al., pp. 471–491, New York: John Wiley and Sons.

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

Goudy, W. J., and H. R. Potter (1975), "Interview Rapport: Demise of a Concept," *Public Opinion Quarterly*, 39, 529–543.

Groves, R. M., and L. J. Magilavy (1986), "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys," *Public Opinion Quarterly*, 50, 251–266.

Haan, M., Y. Ongena, and M. Huiskes (2013), "Interviewers' Questions: Rewording Not Always a Bad Thing," in *Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, pp. 173–193, Frankfurt: Peter Lang Academic Research.

Haan, M., Y. P. Ongena, J. T. A. Vannieuwenhuyze, and K. de Glopper (2017), "Response Behavior in a Video-Web Survey: A Mode Comparison Study," *Journal of Survey Statistics and Methodology*, 5, 48–69.

Holbrook, A. L., M. C. Green, and J. A. Krosnick (2003), "Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Bias," *Public Opinion Quarterly*, 67, 79–125.

Houtkoop-Steenstra, H. (1995), "Meeting Both Ends: Between Standardization and Recipient Design in Telephone Survey Interviews," in *Situated Order: Studies in the Social Organization of Talk and Embodied Activities*, eds. P. ten Have and G. Psathas, pp. 91–107, Washington, DC: University Press of America.

Jäckle, A., C. Roberts, and P. Lynn (2010), "Assessing the Effect of Data Collection Mode on Measurement," *International Statistical Review*, 78, 3–20.

Kane, E. W., and L. J. Macaulay (1993), "Interviewer Gender and Gender Attitudes," *Public Opinion Quarterly*, 57, 1–28.

Kim, Y., Y. Choi, and S. Emery (2013), "Logistic Regression with Multiple Random Effects: A Simulation Study of Estimation Methods and Statistical Packages," *The American Statistician*, 67, 171–182.

Kish, L. (1962), "Studies of Interviewer Variance for Attitudinal Variables," *Journal of the American Statistical Association*, 57, 92–115.

Krysan, M., and M. P. Couper (2003), "Race in the Live and the Virtual Interview: Racial Deference, Social Desirability, and Activation Effects in Attitude Surveys," *Social Psychology Quarterly*, 66, 364–383.

Liu, M., and K. Stainback (2013), "Interviewer Gender Effects on Survey Responses to Marriage-Related Questions," *Public Opinion Quarterly*, 77, 606–618.

Mangione, T. W., F. J. Fowler, and T. A. Louis (1992), "Question Characteristics and Interviewer Effects," *Journal of Official Statistics*, 8, 293–307.

McNeish, D. M., and L. M. Stapleton (2016), "The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration," *Educational Psychology Review*, 28, 295–314.

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

Olson, K., and J. D. Smyth (2015), "The Effect of CATI Questions, Respondents, and Interviewers on Response Time," *Journal of Survey Statistics and Methodology*, 3, 361–396.

O'Muircheartaigh, C., and P. Campanelli (1998), "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161, 63–77.

O'Muircheartaigh, C. A., and A. M. Marckwardt (1980), "An Assessment of the Reliability of WFS Data," World Fertility Survey Conference, Methodology Session No. 6.

Schnell, R., and F. Kreuter (2005), "Separating Interviewer and Sampling-Point Effects," *Journal of Official Statistics*, 21, 389–410.

Schober, M. F., and F. G. Conrad (1997), "Does Conversational Interviewing Reduce Survey Measurement Error?," *Public Opinion Quarterly*, 61, 576–602.

Schober, M. F., F. G. Conrad, A. L. Hupp, K. M. Larsen, A. Ong, and B. T. West (2020), "Design Considerations for Live Video Survey Interviews," *Survey Practice*, 13, 1, available at https://www.surveypractice.org/article/17839-design-considerations-for-live-video-survey-interviews.

Schuman, H., and J. M. Converse (1971), "The Effects of Black and White Interviewers on Black Responses in 1968," *Public Opinion Quarterly*, 35, 44–68.

Sudman, S., and N. M. Bradburn (1974), *Response Effects in Surveys: A Review and Synthesis*, Chicago: Aldine Publishing Co.

Tucker, C. (1983), "Interviewer Effects in Telephone Surveys," *Public Opinion Quarterly*, 47, 84–95.

van der Zouwen, J., W. Dijkstra, and J. H. Smit (1991), "Studying Respondent-Interviewer Interaction: The Relationship between Interviewing Style, Interviewer Behavior, and Response Behavior," in *Measurement Errors in Surveys*, eds. P. P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz, and S. Sudman, pp. 419–438, New York: John Wiley & Sons, Inc.

Welles, B. F., H. Sun, and P. V. Miller (in press), "Nonverbal Behavior in Face-to-Face Survey Interviews: An Analysis of Interviewer Behavior and Adequate Responding," *Field Methods*.

West, B. T. (2020), "Designing Studies for Comparing Interviewer Variance in Two Groups of Survey Interviewers," Chapter 23 in *Interviewer Effects from a Total Survey Error Perspective*, eds. K. Olson, J.D. Smyth, J. Dykema, A.L. Holbrook, F. Kreuter, and B.T. West, pp. 323–334, Boca Raton, FL: Chapman Hall/CRC Press.

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

West, B. T., F. G. Conrad, F. Kreuter, and F. Mittereder (2018), "Can Conversational Interviewing Improve Survey Response Quality without Increasing Interviewer Effects?," *Journal of the Royal Statistical Society (Series A)*, 181, 181–203.

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

West, B. T., and K. Olson (2010), "How Much of Interviewer Variance Is Really Nonresponse Error Variance?," *Public Opinion Quarterly*, 74, 1004–1026.