Check for updates

Special Issue Article

The Audio Check: A Method for Improving Data Quality and Detecting Data Fabrication

Social Psychological and Personality Science
2017, Vol. 8(4) 424-433
© The Author(s) 2017
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1948550617691101
journals.sagepub.com/home/spp
SAGE

Robin Gomila[1], Rebecca Littman[1], Graeme Blair[2], and Elizabeth Levy Paluck[1]

Abstract

Data quality and trust in the data collection process are critical concerns in survey research, particularly when surveyors are needed for reaching "diverse and inconvenient subject pools." In response to irregularities in a smartphone-based pilot survey data collection in Nigeria, we developed an audio check method that unobtrusively recorded surveyors reading aloud questions to participants. We present evidence that this method detected wholesale data fabrication in 14% of our surveys, prevented further fabrication, and improved data quality through provision of regular feedback to surveyors. Using simulation, we demonstrate that undetected fabrication would have introduced significant bias in our analyses. The audio check performs well compared to more traditional methods of detecting fabrication, and a comparative cost–benefit analysis reveals a savings of more than US$1,500 per surveyor by relying on the audio check. The audio check is a viable tool for psychologists who work with survey teams.

Keywords

survey research, data quality, data fabrication, smartphones, audio recorded survey

Historically, psychological research has explored human behavior using laboratory studies in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies. But an increasing awareness that knowledge acquired from WEIRD societies is not necessarily generalizable (Henrich, Heine, & Norenzayan, 2010) has led psychologists to target more "diverse and inconvenient subject pools" (p. 29). Given that over half of the world's population still has no access to the Internet (not to mention that among those with Internet, researchers frequently focus on those who access Amazon's Mechanical Turk; Sanou, 2016) and that literacy rates are below 50% in 14 countries (United Nations Educational, Scientific and Cultural Organization, 2015), targeting other subject pools may require adjusting our methodologies. For instance, survey research on severely underrepresented populations in psychology, such as poor and illiterate people, often must be conducted on the phone or in person by trained surveyors, rather than online or in the laboratory.

Social scientists from other disciplines have a longer tradition of collecting data on non-WEIRD samples and have developed useful logistical and methodological strategies. To address logistical challenges, investigators often provide electronic tablets or smartphones to teams of surveyors who conduct face-to-face interviews with respondents. Surveyors enter participants' responses on the device and the data are uploaded immediately to a secure online server that can be accessed from anywhere in the world. For example, Blattman and Annan (2016) used this technology to survey over 1,000 ex-combatants in remote mining sites and rural villages across Liberia, with surveyors uploading the data as soon as their devices reached an area with mobile network coverage.

Beyond logistics, collecting data using a survey team presents a class of methodological challenges with which psychologists are less familiar. In particular, introducing surveyors into the data collection chain raises the likelihood of data quality concerns due to human error and, in the worst case, fabrication of data. For example, minor errors in the way a question is asked or response options are presented can lead to significant differences in how participants respond to the question (Krosnick, 1999). While outright data fabrication is thought to be less common, a significant number of authors and institutions have reported occurrences of surveyor data fabrication. For instance, the U.S. Census Bureau revealed that 6% of their data were fabricated (Schreiner, Pennie, & Newbrough, 1988), and marketing researchers have reported the rates of fabrication ranging from 1% to more than 13% (Case, 1971; Kiecker & Nelson, 1996).

A recent meta-analysis on fraud in public opinion surveys estimates that about 20% of widely used international data sets

[1] Princeton University, Princeton, NJ, USA
[2] University of California, Los Angeles, Los Angeles, CA, USA

Corresponding Author:
Robin Gomila, Princeton University, Peretsman Scully Hall, Princeton, NJ 08544, USA.
Email: rgomila@princeton.edu

contain more than 5% of fraudulent data (Kuriakose & Robbins, 2016). In particular, the meta-analysis finds that in data from non-Organisation for Economic Co-operation and Development (OECD) countries, sites of largely underrepresented populations in psychology, 28% of observations are likely to be fraudulent. This is substantially more than the 4.6% of likely fraudulent observations in data from OECD countries, where psychological data are typically collected. Thus, two challenges for research with hard to reach and underrepresented populations that require survey teams for data collection are (1) to ensure that all surveyors administer each question properly throughout the duration of the study and (2) to minimize the chances of outright data fabrication.

In a recent experimental study on encouraging corruption reporting in Nigeria (Blair, Littman, & Paluck, 2016), we surveyed approximately 4,000 people across three waves of data collection in 106 southern Nigerian communities. The 45-min survey used at each wave measured psychological reactions to the experimental anticorruption campaign. Collaborating with a Nigerian research firm, we selected and hired a team of local surveyors who could develop a rapport with the members of our sample population and could administer the survey orally in the local English-derived Pidgin language. The oral nature of the interview was crucial because participants in our sample had little to no experience participating in a survey, nearly a quarter had not completed secondary education, and 10% were illiterate.

Although our surveyors already had experience with oral survey administration, our questionnaire was long and addressed complicated topics. We were aware that certain types of errors would be easy to make: misreading the questions, forgetting to read the answer options, or misrecording participants' responses. Furthermore, a large part of our survey was about corruption, a highly sensitive topic for this area. Surveyors had to approach people on the street and read aloud questions such as "how many people in this area have to give money to get out of trouble with police?" or "do you think that government workers are corrupt?" Finally, the 106 communities in our sample were spread across 21,994 square miles and four states in the Niger Delta region of Nigeria, which is largely rural, inaccessible country with bad roads. Sensitive topics and difficult survey conditions constitute the perfect storm for data fabrication (Birnbaum, 2012; Crespi, 1945).

How can researchers ensure high-quality data and prevent data fabrication with survey teams, particularly those who are targeting "diverse, inconvenient sample pools," as we did in our research? A number of well-studied data quality checks exist that can be implemented after data collection, although most approaches feature notable downsides in terms of time, precision, and money. Common data quality checks include searching for patterns of rare combinations in participants' responses (e.g., a report of smoking marijuana but never smoking cigarettes; Murphy, Baxter, Eyerman, Cunningham, & Kennet, 2004), looking for specific patterns in extreme answers or skipped questions in each surveyor's record (Bredl, Winker, & Kötschau, 2008; Schäfer, Schräpler, Müller, & Wagner, 2004), and checking for outlier quantities of nonresponses to survey questions. Another statistical approach based on Benford's Law, also known as the first-digit law, examines the distribution of the first digit of each set of numeric responses, which is expected to follow a specific, positively skewed distribution in nonfabricated data (Bredl et al., 2008; Durtschi, Hillison, & Pacini, 2004). A different type of approach involves extra data collection: resurveying a random selection of participants by different teams of surveyors to compare responses and verify the original data (Li, Brick, Tran, & Singer, 2009).

The primary downside to these common quality checks is their reliance on statistical analyses after most or all of the data set has been collected. As such, they often do not allow for real-time corrections in the midst of data collection, and researchers only have the capacity to detect low data quality or fabrication after weeks or months of data collection. This loss of time usually leads to financial loss: researchers need to pay for new data collection and for expenses incurred from project delays. Another downside is that these data quality checks do not offer clear standards for adjudicating which surveys constitute high- versus low-quality data and can thus only raise suspicion of data fabrication rather than constituting irrefutable evidence that a given surveyor is making up data.

Fortunately, recent technological innovations and applications for survey data collection have given birth to an additional set of quality check opportunities. Data collected through smartphone and tablet software usually come with crucial information such as time stamps and Geographic Information System (GIS) data, allowing researchers to confirm that surveyors worked from the correct location, during working hours, and spent the expected amount of time on each survey. Computer- and phone-based data collection softwares also offer tools to audio record in-person and phone surveys. The company RTI international, for instance, has developed survey quality monitoring solutions such as the computer audio-recorded interviewing system, used by the U.S. Census Bureau in the past years (Mitchell, Fahrney, & Strobl, 2009; Thissen, 2014). Practices and prescriptions about the number of survey questions to record using these types of systems vary; they usually depend on the specifics of the survey methodology and the researcher's needs. Some may decide to only record one question from the beginning, middle, and end of the survey (Thissen, 2014). Others have preferred to assign a probability of recording to each question, that is, between 0.25 and 1 (Hicks et al., 2010), or to select a few specific closed-ended and open-ended questions of interest, that is, between 3 and 5 (Mitchell et al., 2009).

In our own research, we developed a similar kind of "audio check" strategy for improving data quality and detecting fabrication. The smartphone survey software that we used (SurveyToGo, 2013) allowed us to record the voice of our surveyors while they were asking questions aloud to the participants.[1] We implemented this audio check strategy after suspecting data fabrication in the first of three survey waves. In this article, we present and evaluate the benefits of this technique. We first

demonstrate the evidence of the data quality problem that we discovered after completing the first pilot wave of our survey in Nigeria. Next, we describe the audio check technique and demonstrate how much fabrication was detected after introducing an audio check during subsequent waves of survey data collection. We provide the evidence of improvements in surveyor behavior and in data quality after introducing the audio check, using before-and-after comparisons and simulation techniques. Finally, we compare our audio check method to other well-known data quality methods and present a cost–benefit analysis of using an audio check compared with other methods. We conclude that our audio check is a viable and accurate data quality technique valuable for psychologists interested in data collection with survey teams.

Method

We conducted a large-scale field experiment to promote corruption reporting in 106 communities in the Niger Delta region of Nigeria, which included three waves of survey data collection: pilot, baseline, and end line waves (Blair et al., 2016). To collect the data, we partnered with a Nigerian survey company that hired a team of experienced local surveyors to conduct each survey wave. Before the start of each wave, the principal investigators and research assistants of this project conducted intensive in-person trainings with surveyors and selected only the top performers to continue as part of the data collection team.

Surveyors were trained to conduct an oral interview with randomly selected participants using a smartphone that contained one of the more well-known survey software packages available on the market (SurveyToGo, 2013). This software allowed surveyors to upload their data immediately over the mobile phone network to ensure that no or little data would be lost if, for instance, a surveyor misplaced or damaged their phone. Once a survey was uploaded, it no longer appeared on the phone, so this also assured data confidentiality and participant protection in the unlikely event that a phone was lost or stolen.

Surveyors launched the questionnaire on the smartphone by opening a SurveyToGo application on the phone, which displayed each question and answer options individually. Surveyors read the question and answers aloud, entered the participant's response to a question, and tapped a "next" button to advance the survey at a pace appropriate for each question and participant. At the end of the interview or the working day, depending on the mobile network availability, surveyors uploaded their data to an online server that only the principal investigators could access. In addition to the participant responses, the survey software recorded meta-data such as the time stamps of the survey start and finish.

During our pilot survey, examination of the daily upload of survey data led our team to suspect problems with the way the surveys were being conducted and perhaps also data fabrication. Specifically, we noticed significant surveyor differences in average survey durations. We expected that each surveyor would generate a range of survey durations, given that randomly selected participants would range from slow to fast in their responses. We also expected to observe differences among individual surveyors' speed. However, the differences among surveyors were large and, based on our experience with practice surveys during the training sessions, many survey durations did not fall within the expected range. Although we expressed our concerns to the managers of the survey firm multiple times during the pilot wave, the differences in average duration by surveyor persisted (see Figure 1).

In response to these irregularities in survey duration during Wave 1, we developed and implemented the audio check procedure to improve data quality for survey Waves 2 and 3. This preprogrammed option in the survey software allowed us to turn on the phone's microphone and record sound when the surveyor was asking certain questions, without notifying the surveyor that the recording was taking place. Importantly, all surveyors were informed that their voices would be recorded for some questions, but they were not told which questions would be recorded. We selected questions from the beginning, middle, and end of the survey, so that we could sample the surveyor's technique throughout the survey. We made sure to record a few questions with long or complicated instructions and response options and to check if surveyors were administering these questions as they had been trained.

As part of our audio check procedure, a research assistant listened to these audio files each night, after the surveyors uploaded their data. Each survey either passed or failed the audio check if some or all of the survey question and response items were not asked. For example, we encountered audio files that revealed the surveyor was chatting with other people as he or she entered fabricated responses into the survey software. All surveys that failed the audio check were dropped from the study, and the research firm was required to repeat the survey at their cost.

The audio check procedure also allowed us to improve data quality in real time by catching smaller survey administration mistakes. Audio checks often revealed other problems, such as the surveyor speaking too fast or leaving out a response option such as "other." In such cases, the survey passed the audio check but we provided the surveyor with feedback on their performance. Feedback was e-mailed to the research firm manager each night and conveyed to the surveyors by phone from their manager the following day. In this way, the audio check allowed us to provide real-time feedback and to immediately identify and drop fabricated data.

Results

We received audio files for 1,880 surveys following the implementation of our audio check. Of these surveys, a total of 1,620 passed the audio check (86.17%) while 260 failed (13.83%; see Table 1). For the great majority of the surveys that failed the audio check, the audio file revealed that the surveyor was completely silent or was chatting with family or friends about unrelated topics while entering survey responses in the software.[2]

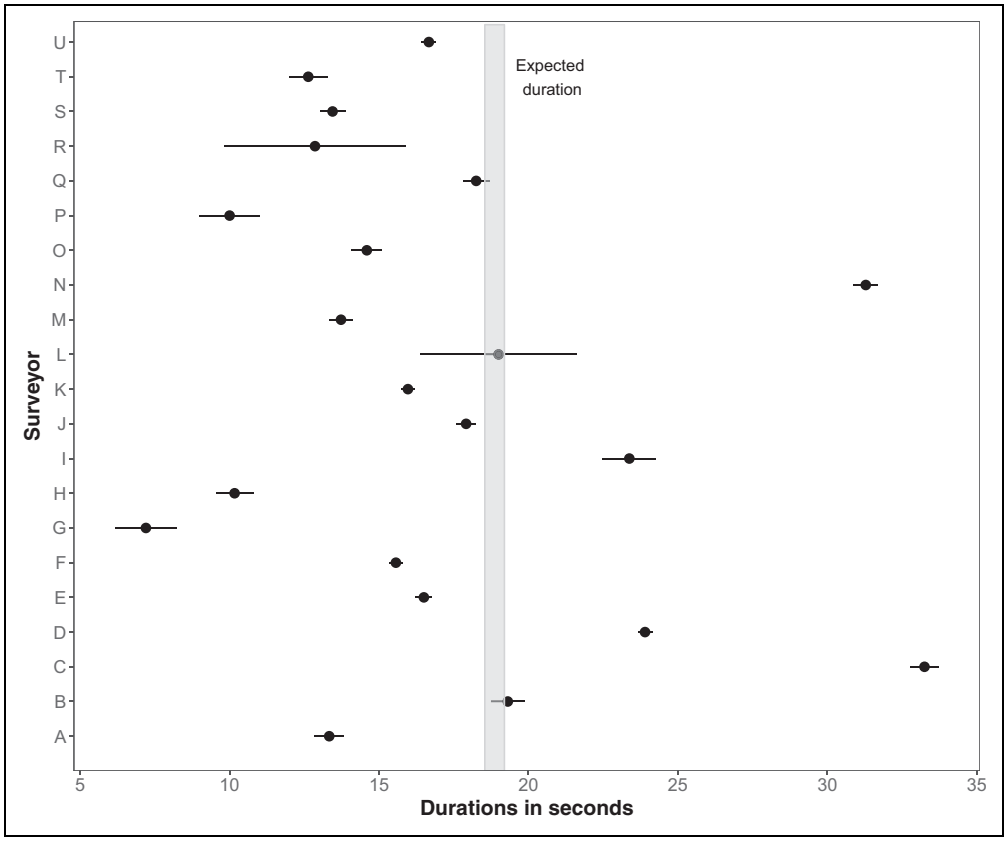

Figure 1. Average question duration by surveyor in survey Wave 1. Average duration of a single question posed by each surveyor in Wave 1, with 95% confidence intervals (CIs). For each survey, we calculated the number of questions asked (which varied due to the presence of skip patterns) and divided the total survey duration by this number to calculate the average time spent per question (*x*-axis). The shaded area represents the ideal time spent per question with 95% CIs. It was calculated from the average duration following the implementation of the audio check and after fabricated surveys were dropped (see Figure 2).

Table 1. Overview of Audio Check Activity.

Variable	Pilot	Baseline	End Line	Total
Survey wave dates	August 30, 2013, to September 23, 2013	October 21 2013 to January 08 2014	February 16 2014 to May 11 2014	
Surveyors	21	26	14	29
Surveys	510	1,904	1,695	4,109
Audio files	No audio check	1,102	778	1,880
Audio files passed check	No audio check	902	718	1,620
Audio files failed check	No audio check	200	60	260
Proportion failed	No audio check	.18	.077	.14

Audio files did not upload to the server for the remaining 1,719 surveys from Waves 2 and 3, likely due to the weak mobile network coverage in rural parts of Nigeria. In our analyses of the audio check's impact on our data, we include those surveys without audio files with the surveys that have "passed" because surveyors could not control whether the audio files uploaded to the server. This ensures a conservative test for a positive impact of the audio check, since it is possible and in fact likely that at least a small percentage of the surveys without audio files include fabricated data. Over the two waves, each surveyor had at least 8% of their surveys audio-checked, with a median of 57%.

The number of surveys that failed the audio check decreased over time, showing that we were not only able to detect but also to prevent fabrication. As shown in Figure 2, data fabrication persisted at much lower rates after the implementation of the audio check and was concentrated at the beginning of each survey wave. This suggests that our daily feedback on the surveyors' performance decreased and eventually eliminated data fabrication over time.

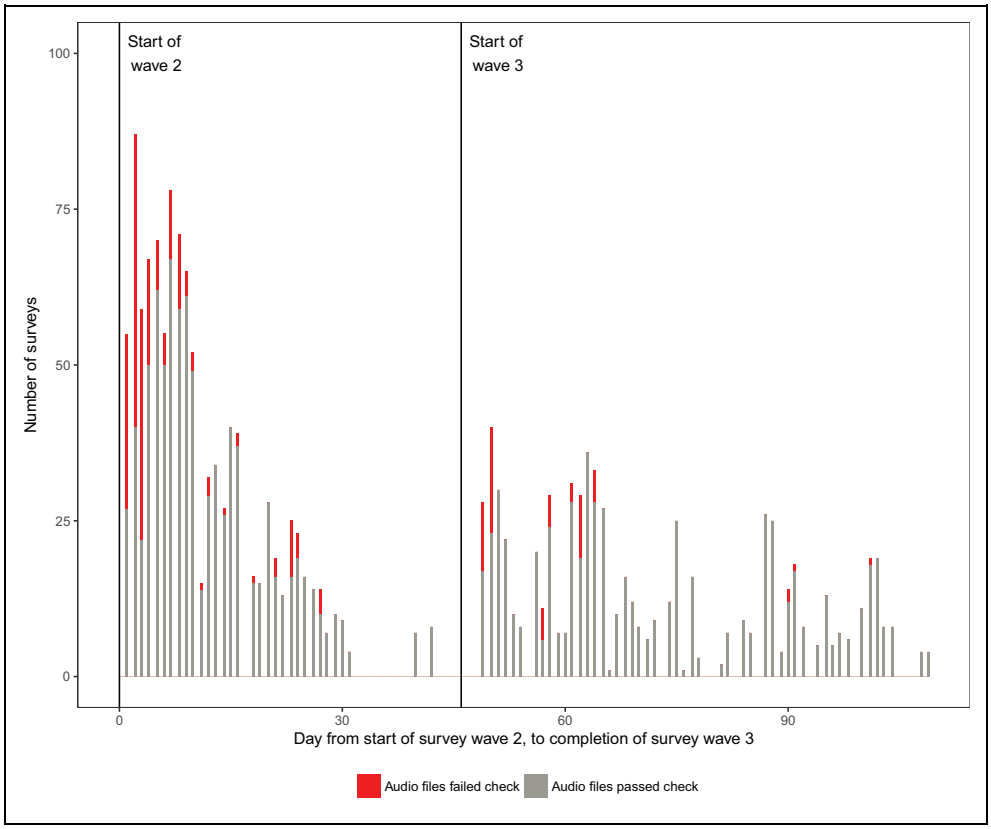

Figure 2. Data fabrication decreases over time in survey Waves 2 and 3.

Turning to data quality, the audio check procedure had an immediate effect on the duration of the survey, which was the signal of low data quality during the first survey wave. Following the announcement of the audio check, we observed sharp improvements in the average question duration, which increased and hovered around our expected time of 18 s per question. Each local regression (loess) depicted in Figure 3a was computed to draw the nonparametric curves using 100% of the data points displayed in the figure (i.e., smoothing parameter $\alpha = 1$) and a linear polynomial (i.e., parameter $\lambda = 1$), thereby providing the highest possible level of smoothness.

In total, the distribution of survey times after the audio check demonstrates a higher average time than prior to the audio check and follows a normal instead of a skewed distribution, suggesting that surveyors were responding to a normal distribution of participant response styles (this is what we would expect, given that participants were randomly selected; see Figure 3b). In other words, some participants were slow and others were quick to respond to each question, but the distribution of times suggests that surveyors were responding to their needs rather than imposing their own pace on participants.

We observed a positive effect of the audio check on surveyors who were "consistently reliable" and also those who were "less consistent." Figure 4 plots the time course performance of surveyors who failed the audio check less than 3 times ever (representing one third of the team) and those who failed 3 times or more. Surveyors who failed three or more audio checks did so largely in the beginning of the second wave of the survey and became more careful and on average took more time with the survey by Wave 3; top surveyors' average times became highly stable, suggesting expertise.

We also conducted simulation analyses, which suggest that introducing the audio check led to significantly different survey results, that is, that the fabricated surveys would have significantly biased our analyses. Using the 62 variables that appeared in both Waves 2 and 3 of the survey, we ran simulations that randomly selected with replacement observations for each variable from two separate pools of surveys: those that passed and did not pass the audio check. We then tested whether the mean of each variable changed significantly with the injection of fabricated observations versus observations that passed the audio check. Specifically, we used 1,000 draws from surveys that passed and failed the audio check to construct one thousand simulated data sets constituted by either 2.5%, 5%, 10%, 20%, 30%, 45%, 60%, 80%, or 100% of resampled observations.[3] We then calculated the difference in means for each variable between the data sets using only surveys that passed the audio check and the data sets with fabricated data, for each of the nine considered proportions. Figure 5 presents the average of all 62 variables' mean differences and the confidence intervals (CIs) of the average difference, at each of our nine levels of fabricated data saturation. The pattern demonstrates that leaving fabricated data in our data set would have led to significantly different mean-level outcomes for our

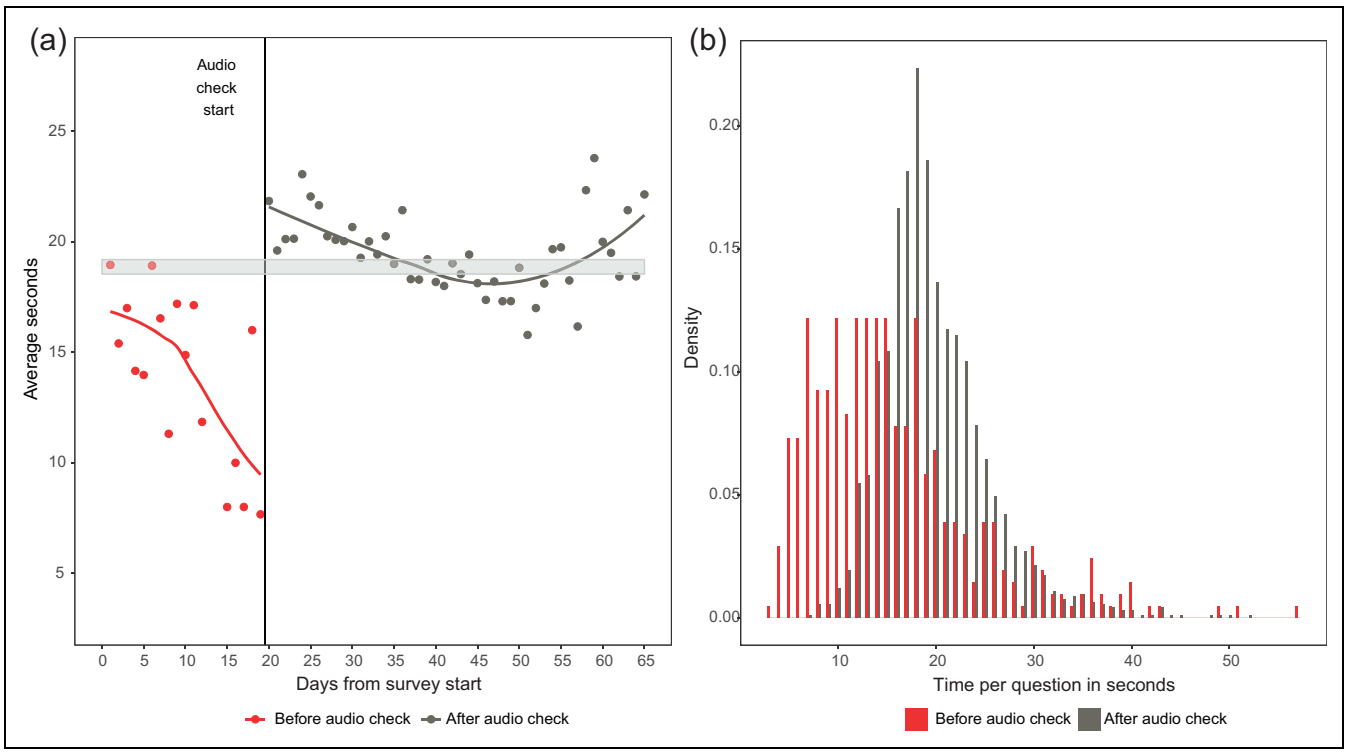

Figure 3. Mean and distribution of time per survey question before versus after the audio check. Each point in (3a) represents the survey team's mean duration in seconds per question, for each working day represented on the *x*-axis. A locally weighted smoothing line (loess) depicts the trend in durations over time before (red) and after (gray) the implementation of the audio check. The gray shaded area represents the researchers' expectation of time spent per question with its 95% confidence interval ($\mu = 18.86$, CI = [18.54, 19.19]).

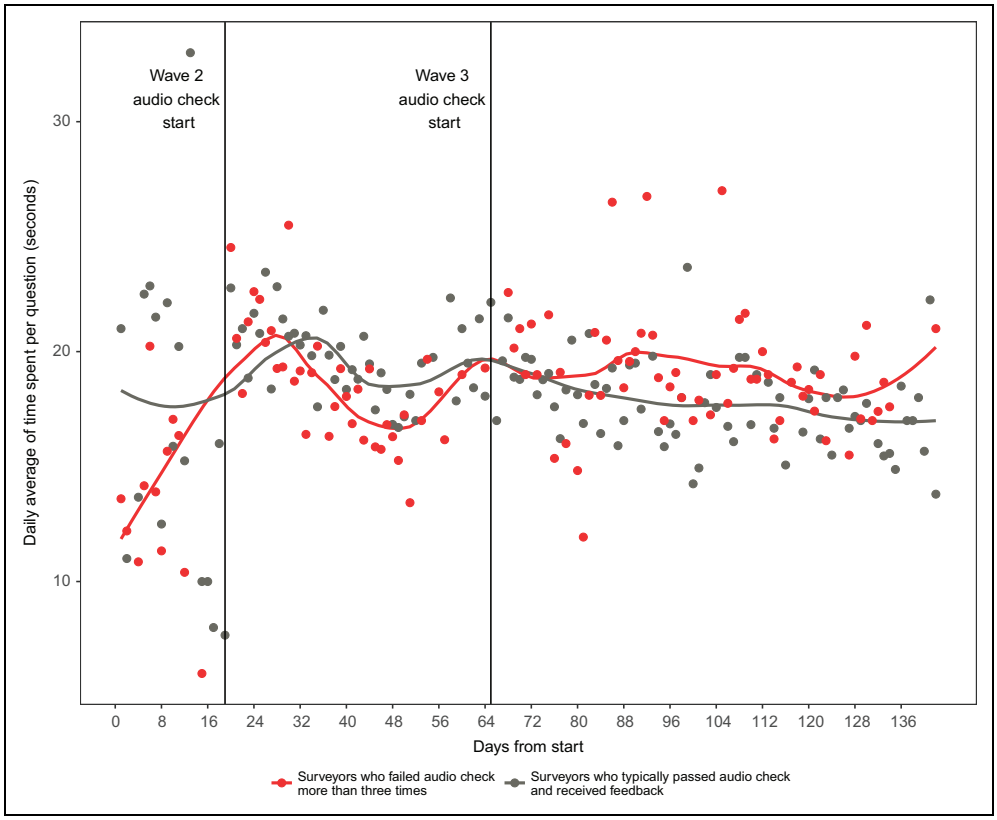

Figure 4. Performance of surveyors improves over time in response to audio check feedback. Each point represents the mean duration per question among surveyors who failed the audio check more than 3 times (red) and among those who failed 3 times or less (gray), in seconds per question, for each working day represented on the *x*-axis. A locally weighted smoothing line (loess) depicts the trends over time for each group.

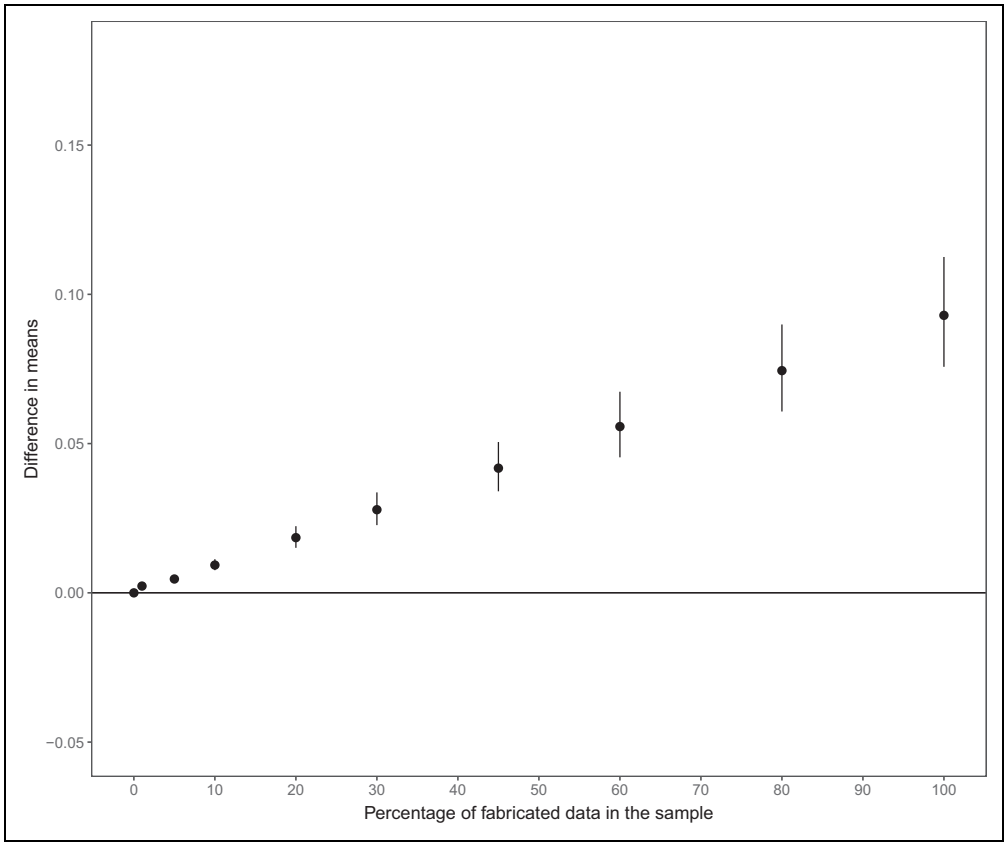

Figure 5. Fabricated data lead to different mean-level outcomes across all variables in both Waves 2 and 3. Average mean differences (and 95% confidence intervals) of 62 variables in which the difference is calculated between the mean of a data set simulated with fabricated observations versus the mean of a data set with all observations that passed the audio check, containing increasing proportions of the resampled surveys.

variables, starting with only 2.5% of fabrication in our data set (we identified 14% of our sample as fabricated and likely prevented even more fabrication by using the audio check; see Figure S2 in the Online Supplemental Materials for individual analysis of all numeric variables from our data set).

Two reliability analyses show that the sample of surveys that failed our audio check contains data patterns that are consistent with more traditional tests for fabrication. First, Figure 6 shows that questions from surveys failing our audio check had, on average, lower variance ($\mu = -0.09$; CI = [−0.13, −0.04]) compared to questions from surveys that passed the audio check (excluding surveys that were uploaded with no audio file). This test is based on the established observation that individuals who fabricate surveys underestimate the number of times participants provide extreme answers (e.g., on a scale from 1 to 7, responding 1, 2, 6, or 7; Bredl et al., 2008; Schäfer et al., 2004).

Second, we used linear regression to show that surveys that passed our audio check contained significantly fewer skipped questions than the surveys that failed the audio check (see Table 2). Specifically, we expect 0.40 fewer skipped questions in the surveys that passed the audio check (effect size = −.39.60; 95% CI = [−0.7423, −0.07]), controlling for survey wave. This finding is consistent with skipped question analysis techniques, based on the fact that surveyors who fabricate data use the shortest possible path toward the end of the survey (Bredl et al., 2008; Hood & Bushery, 1997).

One concern among prospective adopters of this audio check might be the time and financial investment involved in listening to each audio file and providing feedback in real time. We calculated the extra cost incurred by adding an audio check to our survey procedure and compared this to the cost of one common traditional data quality control technique. The cost of including the audio check in our survey software package was US$99 per month to store the uploaded audio files on the server. Additionally, we paid a research assistant to listen to the surveys and to identify those that showed signs of fabrication or data quality issues, requiring surveyor feedback. The research assistant was paid approximately US$17 per hour and needed training and practice such that in the first 3 weeks of the study, he worked 6 hr per day, and in the remaining 8 weeks of the study, he worked 2 hr per day. These numbers are particular to our study and likely represent a higher bound on time involvement, given that the task involved learning to comprehend spoken Nigerian Pidgin English and composing feedback across cultural and language lines. All told, over 8 months of data collection, we spent US$3,660.75 on the audio check procedure and identified 260 fabricated surveys. The cost of finding each fabricated survey was US$14.08, and the cost of replacing one fabricated survey with a new survey was

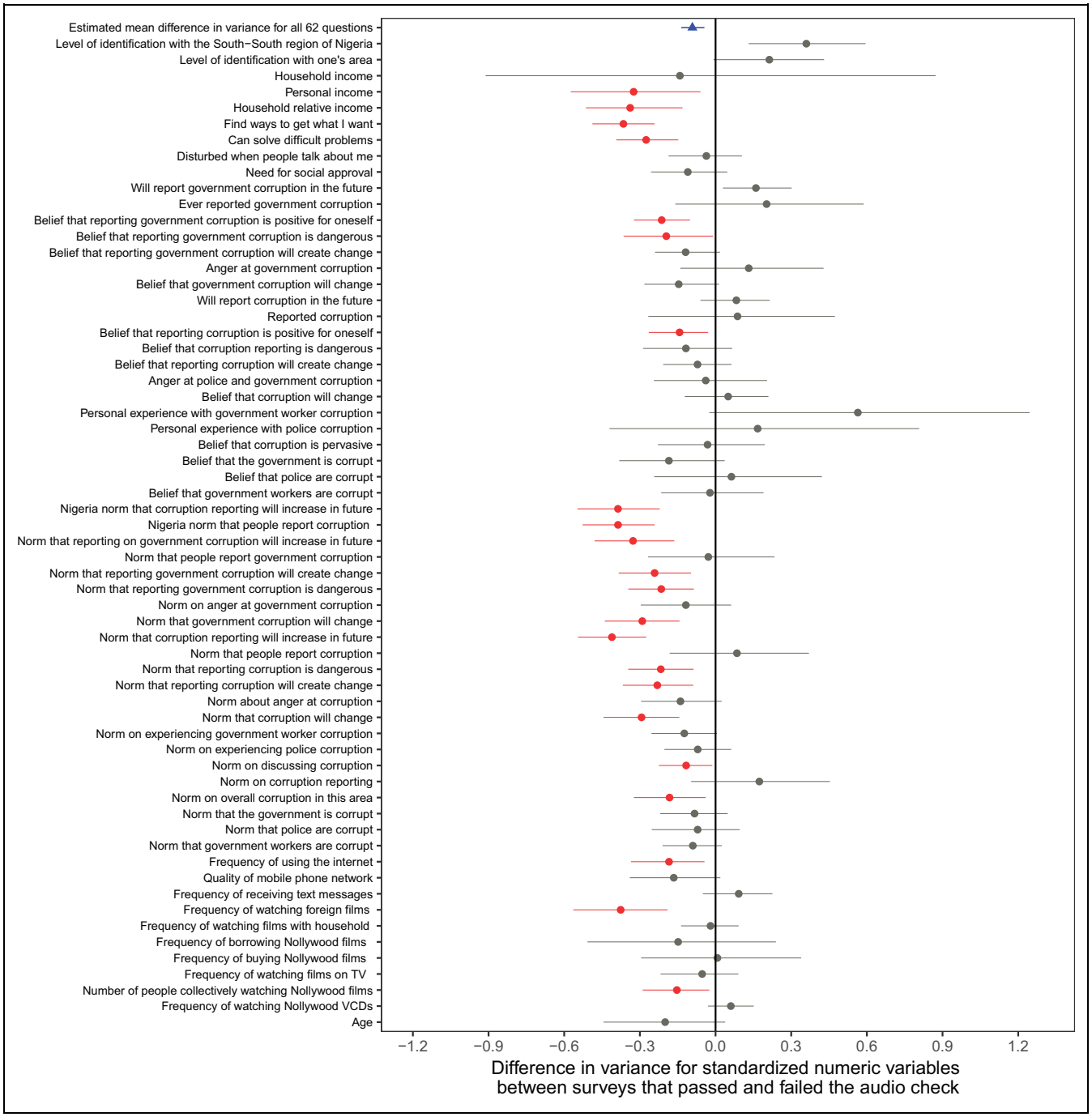

Figure 6. Difference in variance between surveys that passed versus failed the audio check.[4] The difference in variance for each variable[5] between surveys that passed and surveys that failed the audio check was calculated by subtracting the variance of those that passed from those that failed. The 95% confidence interval of each difference was calculated by bootstrapping 1,000 differences in variance and taking the 5th and 95th quantiles of each resulting distribution of differences in variance. The 21 differences displayed in red are significantly lower than 0. The blue triangle at the top of the figure represents the estimated mean difference in variance for all 62 variables.

US$17.47, so in total, the cost of identifying and replacing a fabricated survey was US$31.55.

We now compare this cost to a common alternative method, which involves analyzing data after the majority of data has been collected and identifying the surveyors who are likely to be serially fabricating data. The analysis might involve an examination of skip patterns or differences in variance as demonstrated above, to identify with some uncertainty a group of surveys that may have been fabricated or surveyors who may have made fabrication a practice. To check this allegation of fabrication, we would have to send a new surveyor to find and resurvey a participant. If this resurvey (at US$17.47, although finding the same person again would likely cost more) suggested that the quality of the original survey was indeed poor

Table 2. Surveys That Failed the Audio Check Had Significantly More Skipped Questions.

Variable	Number of Questions Skipped
Surveys passed audio check	−0.386* (0.175) $p = .028$
Survey wave	−0.536*** (0.122) $p = .00002$
Intercept	16.858*** (0.163) $p = .000$
Observations	1,880
Adjusted R^2	.013
Residual standard error	2.585

Note. This regression only includes surveys that explicitly passed or failed the audio check. The audio checked surveys were dummy coded (0 = *failed the audio check*, 1 = *passed the audio check*). The number of questions skipped was individually calculated for each survey and regressed on the audio check dummy variable, controlling for the survey wave.
*$p < .05$. **$p < .01$. ***$p < .001$.

or fabricated, there would be no recourse to correct the original surveyor in real time. The safest correction would be to resurvey all of that surveyor's participants. In our case, the median number of surveys conducted by each surveyor was 96, so this would mean resurveying 96 participants from already-visited communities, at US$1,694.59 (US$17.47 + 96 × US$17.47). Thus, our estimate of the financial gains of the audio check is US$1,694.59 − US$31.55, or US$1,663.04 saved per identification of a surveyor who fabricated.

Notably, one of the major advantages of the audio check method, compared to other postdata collection techniques, is that the audio check method allows the researcher to pinpoint problematic surveys. Thus, researchers avoid the more conservative data strategy of excluding the entirety of a surveyor's data because the surveyor conducted one or two problematic surveys. Additionally, the savings from the audio check method does not account for the ways in which the audio check allowed us to improve data collection among surveyors who were not fabricating data but who needed feedback such as reminders to read aloud response options. The savings does highlight the audio check's real-time precision—the fact that we could intervene the day after a survey was fabricated and tell the surveyor to redo the survey (or to fire the surveyor who was responsible if they did not improve).

Discussion

By implementing an audio check, in which we used smartphone software to listen to preselected questions in our orally delivered survey, we were able to detect and prevent data fabrication during two waves of a large survey in Nigeria. The audio check method can be used ethically; surveyors consented to be recorded for quality control but didn't know which questions would be recorded. Certainly, anticipation of such close monitoring alone helped to prevent fabrication but surprisingly we still identified 14% of our data set as fabricated. We were able to improve the quality of our data over time by providing personalized feedback to each of our surveyors on their work. Had we not used an audio check, our data suggest that the mean responses to our survey questions would be significantly biased. Using more traditional methods of data fabrication detection and quality control would have provided more circumstantial evidence of fabrication and would have incurred a significantly larger financial cost.

Implementing the audio check is straightforward, but the development of complementary technologies could reduce both time and money spent on the audio check. Indeed, a large part of our sample of audio files associated with fabricated surveys presented audio patterns that could fairly easily be detected by software. For instance, surveyors would remain silent the whole time while entering fake responses or chat with relatives or friends about other topics. Developing ways to automatically check the audio recordings would reduce the time investment of this method even further, although we hasten to point out that we learned a lot about our survey's implementation by listening to many different surveys. Another future direction is to introduce quality assurance as the topic of the research and record interactions between interview respondents and surveyors. In the present research, which focused on surveyors who were aware that they would be recorded for quality assurance purposes, the research did not fall within the realm of human subjects research (as determined by the Princeton Institutional Review Board). Recording respondents in future research would bring about additional ethical considerations, such as getting approval from the respondents to record and use their interview data for research purposes.

Overall, in the current context of increasing and justified trends toward collecting data with hard to reach and underrepresented samples, we believe the audio check is a reliable and useful method for psychologists who seek to preserve a high level of data quality while making necessary adjustments to their data collection. We show that the audio check is a technique that makes these data collections more efficient, more reliable, and cheaper. We believe psychologists would do well to add the audio check to their methodological tool kit.

Acknowledgments

Thanks to Brandon Stewart, Elisha Cohen, Han Zhang, Clark Bernier, Nhung Bui, Jeremy Cohen, Janet Xu, and members of Elizabeth Levy Paluck's lab for helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Anonymous donor contribution to E. L. Paluck.

Notes

1. Many survey software companies also offer voice capture and audio recording features.

2. Similarly, it was straightforward to recognize most of the surveys that passed the audio check because we could distinctly hear the surveyor asking the questions to the participant. A few instances were more difficult to assess, for instance, when could hear a surveyor asking the questions to the participant in a clear manner at the beginning of the survey but not in the middle or at the end. For these cases, we adopted a conservative approach by considering the survey unreliable, removing it from our final data set, and providing feedback to the survey team. Note that we could hear surrounding street noise in cases where the surveyor remained silent, which confirmed the reliability of the audio recording.
3. The observations from our final data set that were replaced by resampled observations were randomly sampled without replacement one time for each of the nine proportions under consideration.
4. We also provide a figure displaying the difference in means between surveys that passed versus failed the audio check in the Online Supplemental Materials (Figure S1).
5. The specific questions corresponding to each of these variables are detailed in the Online Supplemental Materials (Table S1).

Supplemental Material

The supplemental material is available in the online version of the article.

References

Birnbaum, B. (2012). *Algorithmic approaches to detecting interviewer fabrication in surveys* (Unpublished doctoral dissertation). University in Seattle, Washington.

Blair, G., Littman, R., & Paluck, E. L. (2016). *Motivating the adoption of new community-minded behaviors: An empirical test in Nigeria*. Manuscript in preparation.

Blattman, C., & Annan, J. (2016). Can employment reduce lawlessness and rebellion? A field experiment with high-risk men in a fragile state. *American Political Science Review*, *110*, 1–17. doi:10.1017/s0003055415000520

Bredl, S., Winker, P., & Kötschau, K. (2008). A statistical approach to detect cheating interviewers. *Survey Methodology*, *38*, 1–10.

Case, P. B. (1971). How to catch interviewer errors. *Journal of Advertising Research*, *11*, 39–43.

Crespi, L. P. (1945). The cheater problem in polling. *Public Opinion Quarterly*, *9*, 431. doi:10.1086/265760

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Bendford's law to detect fraud in accounting data. *Journal of Forensic Accounting*, *5*, 17–34.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–135. doi:10.1017/s0140525x0999152x

Hicks, W. D., Edwards, B., Tourangeau, K., Mcbride, B., Harris-Kojetin, L. D., & Moss, A. J. (2010). Using CARI tools to understand measurement error. *Public Opinion Quarterly*, *74*, 985–1003. doi:10.1093/poq/nfq063

Hood, C. C., & Bushery, J. M. (1997). Getting more bang from the reinterview buck: Identifying "at risk" interviewers. *Proceedings of the American Statistical Association*, *27*, 820–824.

Kiecker, P., & Nelson, J. E. (1996). Do interviewers follow telephone survey instructions? *Journal of the Market Research Society*, *38*, 161–176.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567. doi:10.1146/annurev.psych.50.1.537

Kuriakose, N., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, *32*, 283–291.

Li, J., Brick, J. M., Tran, B., & Singer, P. (2009). Using statistical models for sample design of reinterview program. *Proceedings of the American Statistical Association (Survey Research Methods Section)*, 4681–4695.

Mitchell, S. B., Fahrney, K. M., & Strobl, M. M. (2009). *Monitoring field interviewer and respondent interactions using computer-assisted recorded interviewing: A case study*. AAPOR Conference, Hollywood, FL.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D., & Kennet, J. (2004). *A System for detecting interviewer falsification*, (Vol. 8). American Association for Public Opinion Research 59th Annual Conference, Phoenix, AZ.

Sanou, B. (2016). ICT Facts and Figures 2016. Retrieved August 25, 2016, from http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf

Schäfer, C., Schräpler, J. P., Müller, K. R., & Wagner, G. G. (2004). *Automatic identification of faked and fraudulent interviews in surveys by two different methods*, (No. 441). DIW Discussion Papers.

Schreiner, I., Pennie, K., & Newbrough, J. (1988). Interviewer falsification in Census Bureau Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 491–496.

SurveyToGo [Computer software]. (2013). Retrieved from http://www.dooblo.net

Thissen, M. R. (2014). Computer audio-recorded interviewing as a tool for survey research. *Social Science Computer Review*, *32*, 90–104. doi:10.1177/0894439313500128

United Nations Educational, Scientific and Cultural Organization. (2015). "Adult and youth literacy." Retrieved from http://www.uis.unesco.org/literacy/Documents/fs32-2015-literacy.pdf

Author Biographies

Robin Gomila is a PhD student in psychology and social policy at Princeton University.

Rebecca Littman is a PhD candidate in psychology and social policy at Princeton University. Her research focuses on collective violence, group identification, and promoting behavior change.

Graeme Blair is an assistant professor of political science at the University of California, Los Angeles.

Elizabeth Levy Paluck is a professor in the Department of Psychology and in the Woodrow Wilson School of Public and International Affairs at Princeton University. Her research is concerned with the reduction of prejudice and conflict, including ethnic and political conflict, youth conflict in schools, and violence against women. She is the deputy director of the Kahneman–Treisman Center for Behavioral Science and Public Policy at Princeton.

Handling Editor: Nickola Overall