Methodological Note #002

Assessing and Improving Interview Quality in the 2016/17 AmericasBarometer

Mollie J. Cohen and Sebastian Larrea

LAPOP, Vanderbilt University

mollie.j.cohen@vanderbilt.edu
sebastian.c.larrea@vanderbilt.edu

February 6, 2018

Key Findings:

- LAPOP's human auditing protocol improves interview quality over the course of fieldwork.

- Data fabrication is present in a very small proportion of interviews—7% of those cancelled, and less than 2% of all interviews conducted.

- The vast majority of errors in AmericasBarometer surveys are issues with reading questions completely and correctly. Even these errors decrease over the course of fieldwork, using human auditing protocols.

USAID FROM THE AMERICAN PEOPLE   VANDERBILT   LAPOP   AmericasBarometer Barómetro de las Américas
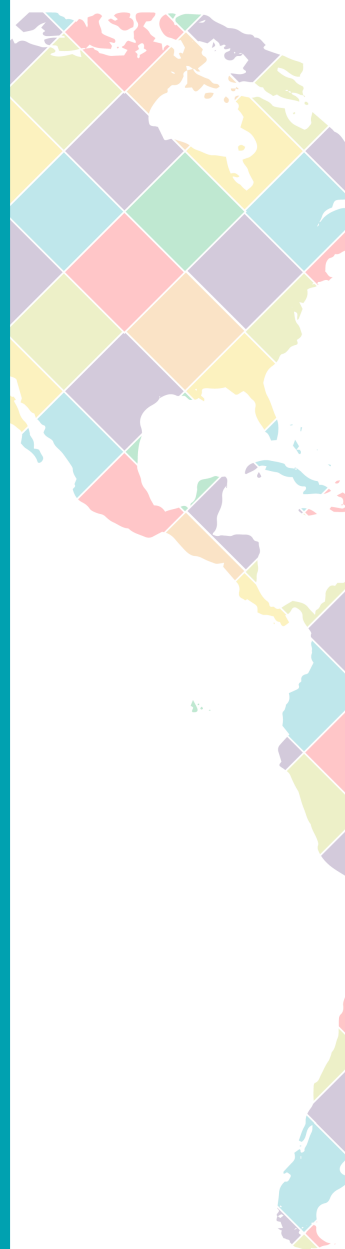
Recent methodological advances have enabled researchers to identify likely data falsification and fabrication in survey datasets (e.g., Kuriakose and Robbins 2016). However, survey researchers have historically been unable to identify and stop data fabrication from occurring in real-time, or to provide additional support and training to interviewers who need it while fieldwork is in progress. LAPOP's quality control system, particularly the implementation of near real-time auditing, represents a substantial advance toward this goal. This *Methodological Note* describes LAPOP's human auditing practices[1] to ensure data quality (a protocol that relies heavily on the quality control module—the "QuAC"[2]—we have programmed into our data collection software), and describes the errors we documented and addressed during fieldwork in ten countries included in the 2016/17 AmericasBarometer study. The vast majority of these errors are issues with reading questions completely and correctly; that is, most errors are indicative of deficiencies in interview quality, and not with data fabrication, in the 14% of interviews that are eventually cancelled. We also show that the implementation of near real-time auditing can improve interview quality as fieldwork progresses.

Near Real-Time Auditing Using the QuAC

For the 2016/17 round of the AmericasBarometer, and as part of the larger FALCON©protocol, LAPOP implemented silent image capture of interviewers, GPS locations, and audio recording of key questions.[3] These practices enable teams working on LAPOP projects to assess the quality of interviews in near real-time and provide feedback to interviewers throughout fieldwork.[4] During fieldwork, auditors in LAPOP Central and staff in our field teams compare images silently captured via front-facing cameras to interviewer photos to ensure that the enumerators in the field were those trained by LAPOP staff. In addition, we scrutinize interviews' GPS coordinates to ensure that each interview takes place in the assigned location (and that the location is, in fact, a residence rather than a public park or shopping mall). Finally, we listen to audio record-

ings to ensure that enumerators read items completely and correctly, without interpreting the question in their own words, skipping items, or influencing respondents' answers.[5]

Based on these audits, we assign each interview a quality control score. Higher scores indicate more serious errors, and we refuse to accept (that is, we require the cancellation of) interviews with scores above 20. In the ten countries that used the final version of the QuAC, we cancelled 14.3% of all interviews for a broad range of errors.[6]

How does near real-time quality control work in practice? In the first week of each study, the staff audit all interviews.[7,8] In a matter of days, it becomes clear which interviewers are outstanding (members of an "A Team"), and which interviewers need additional guidance to improve (members of a "B Team"). We then shift our strategy, reviewing the A Team's work periodically, but focusing most of our efforts on members of the B Team.[9] Within 48 hours of each interview, we send local fieldwork teams information identifying problematic interviews and suggestions for improving the work of specific interviewers. LAPOP is thus able to provide feedback to enumerators in near real-time.[10]

Typically, the feedback we provide is general. "Please remind Juan Pérez to read questions completely and correctly." "María López often speaks too quickly for respondents to understand; please remind her to speak at a more conversational pace." Often, we identify specific interviews that exemplify these issues. Providing this feedback works: over the course of fieldwork, we find fewer problematic interviews as enumerators remember to follow best practices—reading all questions and response options as written, reading at a conversational pace, and reading the study information sheet (SIS)[11] without error.

Positive feedback is also important. We believe that members of both the A and B teams deserve recognition for their good work, so we do our best to provide good news to local survey teams. These comments are also general, and might read, "Pablo Flores' interviews have improved markedly this week; he is incorporating feedback to read all items com-
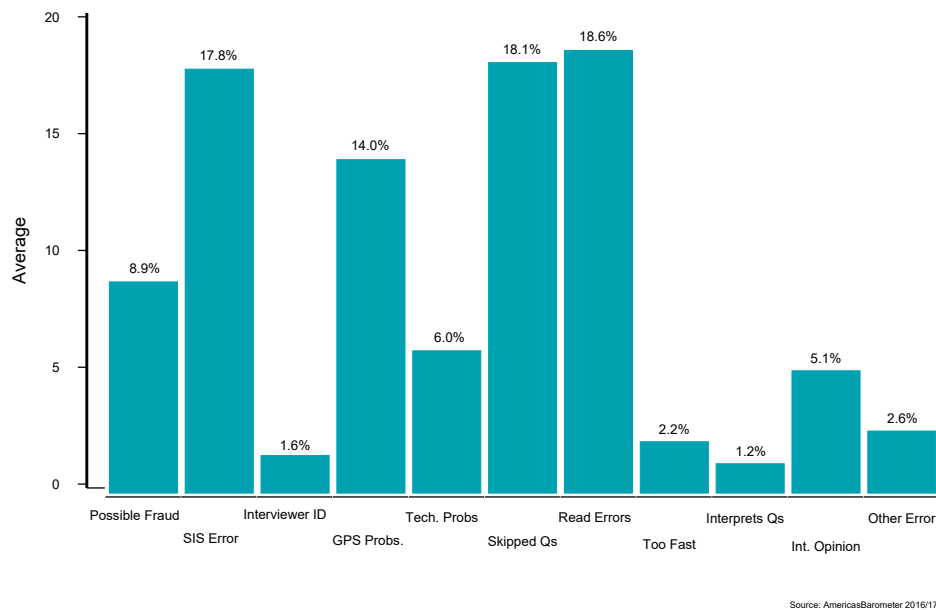
Figure 1: Recorded Errors among Cancelled Interviews

pletely and correctly. Please tell him that we've noticed the improvement!" or "Juana Díaz has done outstanding work, as shown in the following interview. Please tell her to keep up the good work!"

A plurality of issues we encounter with interviews are reading errors. Figure 1 shows the distribution of errors among interviews that were subsequently cancelled.[12] Interviews can have more than one error, so these values do not add up to 100%. Reading errors, issues with the GPS, and errors reading the SIS were far more common than errors consistent with data fabrication (which occurred in only 8.9% of the 14.3% of interviews that were eventually cancelled). This suggests that most of the error in our surveys is the result of interviewers forgetting or ignoring foundational lessons from their training. If reminding enumerators of these best practices improves their work, our feedback mechanism should result in declining quality control scores (that is, improved interview quality) and fewer cancellations over the course of fieldwork, as enumerators gradually stop making these mistakes. We analyze these metrics below.

QuAC Scores and Cancellations over Time

How does interview quality change over the course of fieldwork? We answer this question by analyzing two key metrics–QuAC scores and total cancellations–over each interviewer's tenure on the project.[13] Because new enumerators often begin to work on the project once fieldwork is already underway,[14] we assess QuAC scores and cancellations as a function of how much time a given *interviewer* has spent in the field.[15] While trends in quality control scores and cancellations vary somewhat across contexts and individuals, we find that, on average, both the percentage of interviews cancelled due to poor quality and the average QuAC score decline over the course of fieldwork.

Figure 2 illustrates our results with regard to average QuAC scores. In the left panel, interviewer field days are on the x-axis. A value of 1 indicates an interviewer's first field day, which might take place at the beginning of national fieldwork or after several days' or weeks' worth of interviews have been completed. Higher values indicate additional days of interview experience. As expected, average QuAC scores trend downwards as fieldwork progresses, although there are also notable spikes in observed QuAC scores toward the end of fieldwork.

The right panel of Figure 2 shows the QuAC score predicted by a linear regression analysis including interviewer field day, a squared term to capture nonlinearities by interviewer, and country fixed effects. The regression analysis similarly shows a downward trend in predicted QuAC scores over the course of fieldwork, indicating improved interview quality. We note that the vast majority of interviews are conducted prior to the 100-day mark. Our estimates of interview quality increase slightly following the 100-day mark.[16] We believe this reflects the fact that teams sometimes rush to conduct interviews at the end of fieldwork in order to meet deadlines. This can yield a spike in low-quality interviews–and cancellations–in the final days of fieldwork. Because of the relatively small number of interviews that take place at these late dates, we see substantially more uncertainty around the average QuAC score as the
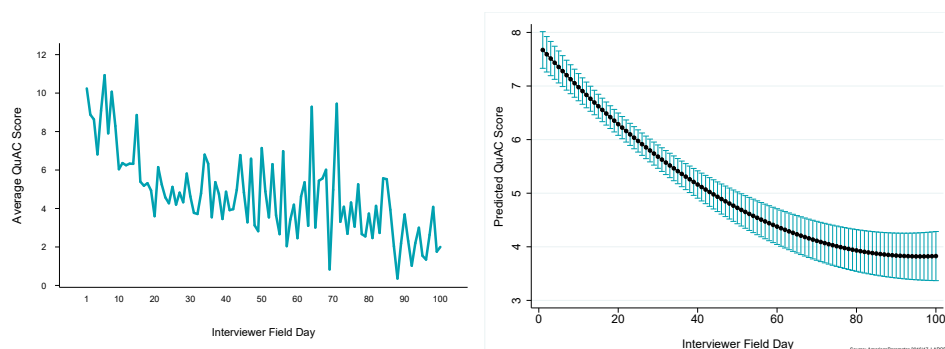
Figure 2: Average and Predicted QuAC Scores, by Interviewer-Fieldwork Day

number of days an interviewer has spent in the field increases beyond about 30 days.

We see a similar trend when we consider the percentage of interviews that are cancelled as interviewers work on the project. Figure 3 shows descriptive averages of the percentage of interviews cancelled by fieldwork day (left panel) and an identical regression analysis to that presented in Figure 2, using the percentage of interviews cancelled as the dependent variable (right panel). Both figures show a decrease in cancelled interviews as fieldwork progresses, with very low average levels of cancellations and sudden spikes in the final days of fieldwork. As before, we believe this is a reflection of fieldwork teams' rush to meet project deadlines, resulting in low-quality interviews that are then cancelled. Like the descriptive analysis, the regression analysis shows a downward trend in cancellations as an enumerator spends more time in the field, with more uncertainty over the level of cancellations towards the end of fieldwork.

It is possible that interview quality would improve over the course of fieldwork regardless of our quality control protocols, as firms fire low-quality interviewers and the remaining enumerators become increasingly familiar with the questionnaire and therefore better able to read questions fluently. We cannot rule this out on the basis of these data; however, the uptick in interviewer errors we observe towards the end of field-
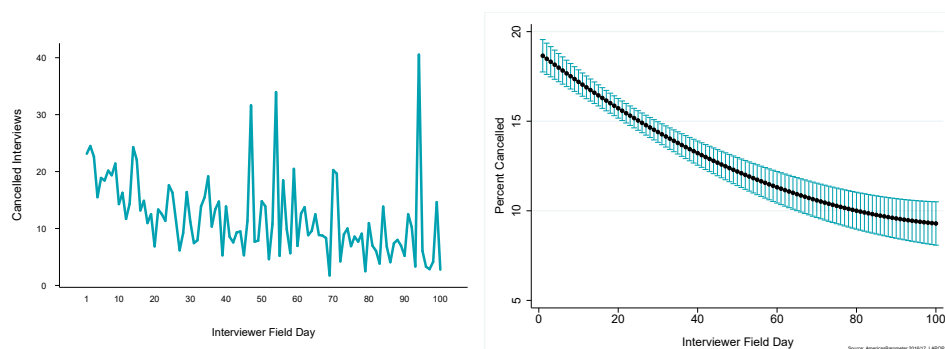
Figure 3: Average and Predicted Cancellations, by Interviewer-Fieldwork Day

work suggests interviewers may in fact adhere *less strictly* to standard protocols as they become more familiar with the instrument.

Data Fabrication over Time

While most cancelled interviews are cancelled because of sloppy reading or accidental error, this is not always the case. Occasionally, we identify members of an "X Team"–those who refuse to follow protocols after being reminded to do so, or who intentionally and continually cheat. We provide evidence of such shirking to local survey teams and ask that these interviewers be removed from the project. However, such cases are rare: only about 1% of all interviews from these ten countries had errors consistent with fabrication (7% of the 14.3% of interviews that were eventually cancelled).[17]

Further, problematic interviewers are usually identified in the early days of fieldwork. Figure 4 below shows the average percentage of all interviews with errors consistent with fraud by interviewer fieldwork day (left panel) and a regression analysis identical to those described above predicting the percentage of interviews with errors consistent with fraud (right panel). Rates of potentially fraudulent interviews are always low (less than 2% of all interviews). However, the proportion of all interviews
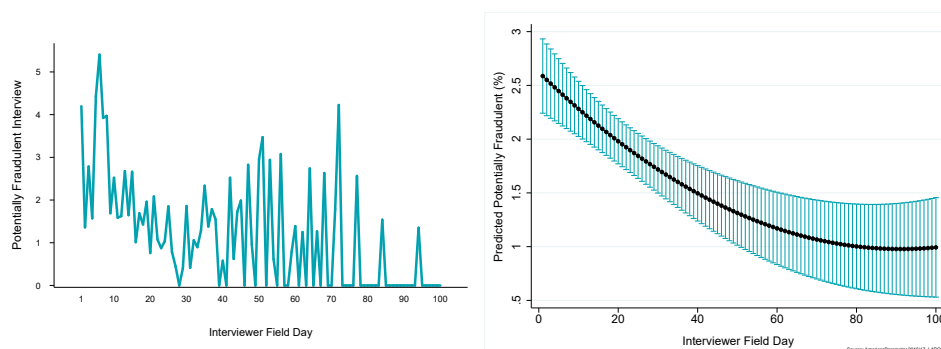
Figure 4: Average and Predicted Percentage of Interviews with Errors Consistent with Fraud, by Interviewer-Fieldwork Day

with errors consistent with fabrication declines substantially as fieldwork progresses, from about 1.7% of all interviews on the first day of fieldwork to 0.6% of all interviews by day 75.

Not only are fabricated interviews rare, with some variation by country and interviewer, but their prevalence decreases over time. That is, our strategy for rooting out very low-quality interviews and interviewers results in less data fabrication over time. One central message of the above figures, however, is that *even when fabrication is at its highest*, potentially fraudulent interviews account for a very small portion of all interviews conducted as part of the AmericasBarometer project.

Conclusion

Making scrupulous and systematic use of audio recording, image capture, and GPS location, LAPOP is able to assess and improve interview quality *and* provide timely, detailed feedback to interviewers and fieldwork teams in near real-time. We will continue to refine these tools and processes in an effort to ensure that LAPOP data are of the highest possible quality.

Our efforts to identify and codify the different types of errors that occur during the interview process also provide insight into the prevalence of

serious errors like those consistent with fabrication. We are pleased to report that such errors account for a very small portion of all errors in our interviews. The vast majority of errors, such as misreading questions, are consistent with sloppy or forgetful interviewing, *not* with data fabrication. These findings are consistent with recent analyses (for example, Robbins 2016) suggesting that public opinion data from developing contexts is generally trustworthy. Moreover, these are heartening results for those who work with international public opinion data.

Notes

1. LAPOP's auditing protocols include the human auditing described here as well as the implementation of automated scripts in our data collection software (in the 2016/17 round, SurveytoGo) and R (see also note 5 below). In this *Methodological Note*, "auditing" refers to the human component detailed here.

2. QuAC stands for "Quality Assurance Chapter."

3. All images used a front-facing camera to ensure that respondent anonymity was not compromised (that is, the camera only recorded images of the interviewers). Study participants were informed prior to consenting to be interviewed that some of their answers would be recorded for quality control purposes.

4. Prior to onboarding, all interviewers on the AmericasBarometer project undergo a two-day training that covers (among other topics) the project, the questionnaire, best practices, and also includes role-playing and practice interviews.

5. In addition to this review, we periodically analyze "Key Performance Indicators," which provide detailed information about fieldwork start and end times each day, among other metrics. We scrutinize mobile device logs to ensure that interviewers have not altered phone settings to impede, for example, the collection of GPS coordinates. We also track the route the interviewers traveled over the course of each fieldwork day to assess whether they correctly followed rules for selecting dwellings and individual respondents. These protocols have also improved our ability to track and verify attempted contacts and refusals. Each of these processes is detailed elsewhere in the *Methodological Note* series, so we do not focus on them here.
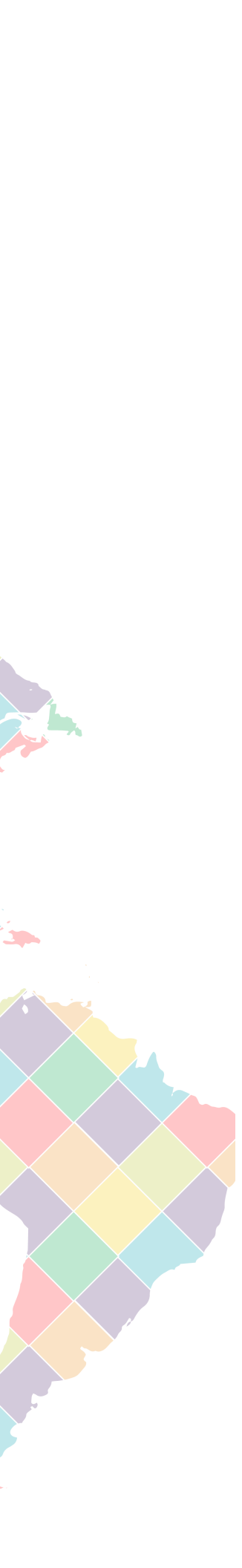
6. We updated the QuAC over the course of the 2016/17 round. The second version, assessed here, was used in: Argentina, Bolivia, Brazil, Chile, Guatemala, Haiti, Jamaica, Mexico, Panama, Peru, and Uruguay.

7. Field teams audit 100% of interviews for the entirety of the project. All interviews thus pass through at least one round of human auditing.

8. A single auditor is usually assigned to a group of interviewers. This way, the individual auditor becomes familiar with an interviewer's voice and speech patterns and, we believe, better able to identify errors consistent with fraud.

9. We periodically review interviews conducted by members of the "A Team" and, in the very rare instance that we find errors consistent with fabrication, conduct a thorough review of all their past work.

10. While we are unable to confirm whether local team leaders relay our feedback to the enumerators themselves, we do find that the quality of interviews conducted by each interviewer tends to improve over time, consistent with having received the feedback.

11. The SIS provides respondents with information about their rights as a study participant, including their right to refuse to answer questions and to stop the interview at any time.

12. Such errors are also common among interviews that were not cancelled. 40.6% of accepted interviews had at least one reading error, 16.4% had errors related to the study information sheet, 8.1% of accepted interviews had technical issues, and in 7.7% of accepted interviews, enumerators had skipped sections of questions.

13. Due to inconsistencies in the underlying information about interview timing, Argentina is excluded from this cross-time analysis.

14. New interviewers may join the project over the course of fieldwork due to turnover. The timing of interviews may also be staggered across subnational regions, leading to staggered start dates for enumerators.

15. When we do not normalize the data in this way, we see more spikes in both metrics, as new interviewers begin on the project, make errors, are corrected, and improve over time.

16. Fully 90% of all interviews in this subset of countries were completed by day 102, and

99% of all interviews were conducted prior to the 150-day mark.

17. Errors consistent with data fabrication include (but are not limited to): interviews that change location or are completed in a secondary location; interviews that have multiple respondents or interviewers; interviews for which GPS capture has been manually disabled; and interviews that take place in suspicious environments (e.g., a moving vehicle, or a bar). In a small number of instances, we captured enumerators discussing their plans to shirk as part of the audio recording. Some of these errors may have legitimate causes. For example, it is possible that a respondent could invite an enumerator into a vehicle to complete an interview in a private location, or that an interviewer could realize she forgot to enter her unique ID number to complete an interview and do so in a secondary location following the interview. We err on the side of caution by cancelling such interviews. Thus, our measure of fraudulent interviews is likely an over-estimate.

References

Kuriakose, Noble, and Michael Robbins. 2016. "Don't Get Duped: Fraud through Duplication in Public Opinion Surveys." *Statistical Journal of the IAOS* 32.3 (2016): 283–291.

Robbins, Michael. 2016. "Yes, You Can Trust International Surveys. Mostly." *Monkey Cage.* `http://wapo.st/21FBQb2?tid=ss_tw-bottom&utm_term=.0ad06aa88bc1`.