How to get better survey data more efficiently[*]

Mollie J. Cohen[†] and Zach Warner[‡]

February 2, 2019

Abstract

A key challenge facing many large, in-person public opinion surveys is ensuring that enumerators follow fieldwork protocols. Implementing "quality control" processes can improve data quality and help ensure the representativeness of the final sample. Yet while public opinion researchers have demonstrated the utility of quality control procedures such as audio capture and geo-tracking, there is little research assessing the relative merits of such tools. In this paper, we present new evidence on this question using data from the 2016/17 wave of the AmericasBarometer study. Results from a large classification task demonstrate that a small set of automated and human-coded variables, available across popular survey platforms, can recover the final sample of interviews that results when a full suite of quality control procedures is implemented. Taken as a whole, our results indicate that implementing and automating just a few of the many quality control procedures available can streamline survey researchers' quality control processes while substantially improving the quality of their data.

[*] For their advice and comments, we are grateful to Noam Lupu, Mitch Seligson, Chris Warshaw, Liz Zechmeister, and seminar participants at Cardiff. We thank the Latin American Public Opinion Project for generously sharing their data with us, and the LAPOP team (Rubi Arana, M. Fernanda Boidi, Nicole Hinton, Sebastian Larrea, J. Daniel Montalvo, Georgina Pizzolito, Mariana Rodriguez, and Carole J. Wilson) for its excellent work in data collection for this project. We also thank Alyssa Chvasta, Christine Huang, and Linzy Scott for their research assistance. A previous version of this paper was presented at the 2018 annual conference of the American Political Science Association. Both authors contributed equally; the ordering is alphabetical.

[†] Assistant Professor, University of Georgia. email: mj.cohen@uga.edu, web: http://www.molliecohen.com.

[‡] Postdoctoral Research Fellow, Cardiff University. email: WarnerZ@cardiff.ac.uk, web: http://www.zachwarner.net.

Political scientists are increasingly relying on large-scale public opinion surveys (Heath, Fisher, and Smith 2005). These studies provide important insights into how citizens relate to legislators, understand democratic norms, and participate in electoral politics, among many other areas of scholarly interest. A central challenge for the researchers who field such surveys is to ensure the quality of the data, particularly when conducting surveys in the developing world (Lupu and Michelitch 2018). Among the most persistent threats to data quality is enumerators deviating from fieldwork protocols. Enumerators may fail to properly screen respondents for eligibility, instead interviewing people who are outside the population of interest. They may also misread or interpret questions, potentially biasing respondents' answers. Other common problems include enumerators venturing outside the sampling area, recording answers incorrectly, failing to report unsuccessful interview attempts, and falsifying interviews (Montalvo, Seligson, and Zechmeister 2018).

Observations with deficiencies arising from inconsistent enumeration—which we call *low-quality* data—can limit researchers' ability to make inferences. These data may bias statistical estimates and understate the uncertainty associated with those estimates (Sarracino and Mikucka 2017; Gomila et al. 2017). Further, low interview quality and persistent violations of sampling protocol impede efforts to replicate the data collection process, threatening a foundational principle of rigorous public opinion research. To prevent these problems, scholars have developed a number of tools for assessing interview quality, particularly through Computer-Assisted Personal Interviewing ("CAPI"), which allows for monitoring quality in real time. Yet there is little evidence as to these methods' relative effectiveness outside of single-case studies in which they have been developed and implemented.[1] Scholars are left with little guidance for preventing low-quality data because the comparative merits of these tools are essentially unknown.

In this paper, we conduct the first (to our knowledge) systematic examination of methods to prevent and eliminate low-quality interviews in large-scale public opinion surveys. Our empirical strategy relies on three unique features. First, we draw on data

[1] Important exceptions include recent attention to screening out duplicate and near-duplicate interviews discussed below (Kuriakose and Robbins 2016; Blasius 2018).

collected in nine countries during the 2016/17 round of the AmericasBarometer surveys conducted by the Latin American Public Opinion Project (LAPOP) at Vanderbilt University. LAPOP generously provided us with not only the published data, but also all interviews screened out due to quality concerns. These data allow us to observe a binary indicator of interview quality—cancellation versus publication—in a large-scale, cross-national survey that is internationally recognized for its methodological rigor.[2] Second, LAPOP also provided us with 141 distinct quality control checks conducted on each interview in these data. These checks, described in detail below, include all tools for assessing interview quality of which we are aware,[3] allowing us to compare them on a common sample. Finally, we conduct a large classification task to identify the tools which are most informative for identifying low-quality data, using standard variable importance metrics used in machine learning applications (Hill and Jones 2014). By measuring each check's ability to predict low-quality interviews, we are able to identify the most powerful tools for ensuring surveys do not suffer from the problems associated with inconsistent or inadequate enumeration.

We find that a full suite of quality control tools is extremely effective in identifying low-quality interviews: on average, the root mean squared error (RMSE) of prediction for models estimated with all available tools is just three percentage points. In other words, using all currently-available methods for evaluating interview quality, we typically assign a 97% probability of being low-quality to interviews that LAPOP ultimately canceled, and a 3% probability to those that LAPOP did not.

However, we also find that a small subset of these tools produces very similar—and in some ways better—results. Specifically, we show that light manual auditing of random audio recordings, an interview timer, and a few metrics easily calculable from the data (such as Percentmatch and completion percentage; Kuriakose and Robbins 2016) are together sufficient to recover a sample nearly identical to that produced by a

[2] . The AmericasBarometer is the 2018 recipient of the Lijphart/Przeworski/Verba Data Set Award, given by the American Political Science Association's Comparative Politics section.

[3] . We found no additional quality control procedures in our review of publicly available technical reports published by the United States Census Bureau, the United Kingdom's Office for National Statistics, the American National Election Studies, the Afrobarometer, the Arab Barometer, and the Latinobarómetro.

full suite of checks. Taken as a whole, our estimates suggest that scholars can diagnose problems of interview quality with just a few quality control procedures, forgoing many of the methods used in the 2016/17 round of the AmericasBarometer.[4]

Our results provide an emphatic answer to recent calls for more rigorous research into identifying and preventing problems stemming from low-quality data (e.g., Lupu and Michelitch 2018). We provide a blueprint for public opinion researchers to choose a suite of methods to efficiently diagnose low interview quality. In so doing, we hope to empower researchers to improve the overall quality of their data, and thus the reliability of the inferences that can be drawn from survey research in political science.

1 Strategies for preventing low-quality interviews

Since the advent of polling, survey researchers' attempts to ensure data quality have focused on the problem of enumerator "cheating" (Crespi 1945), and particularly "curbstoning," the wholesale fabrication of interviews. Scholars have developed a variety of methods to detect fake interviews. Early strategies include asking enumerators to sign statements affirming that they correctly followed protocols (Bennett 1948) and sending fieldwork supervisors to conduct partial re-interviews with participants to verify their participation (Winker 2016). More recently, researchers have introduced checks for interview duplication and straightlining, wherein enumerators generate fraudulent interviews by filling out identical answers across a battery of questions (Blasius 2018).

Rapid expansion in the use of hand-held electronic devices for survey enumeration has created new opportunities to detect cheating. Researchers who conduct face-to-face surveys integrate CAPI methods to capture detailed metadata about each interview. These metadata are then processed to identify violations of fieldwork protocols (Seligson and Moreno Morales 2015). Among methods using such metadata, most common are those that rely on Geographic Positioning System (GPS) data: researchers unobtrusively

[4] . Because the costs of both collecting survey data and implementing quality control checks vary widely across contexts, we cannot definitively answer how much money our proposed strategy would save researchers. Our goal is instead to guide researchers on how to utilize their resources efficiently by implementing the most informative quality control methods.

capture GPS coordinates, documenting the precise location in which an interview was conducted (Montalvo, Seligson, and Zechmeister 2018). Such methods quickly identify interviews conducted at locations outside the assigned area of enumeration. A more involved method relies on silent audio recordings of enumerators' work, allowing researchers to audit interviews to ensure they reflect answers given by a real respondent (computer audio-recorded interviewing; Cohen and Larrea 2018; Hicks et al. 2010; Mitchell, Fahrney, and Strobl 2009).

These innovative quality control procedures have a decided advantage over their predecessors: they enable researchers to identify and resolve problems in semi-real time—just a few hours or days after the interview is conducted. By uncovering potentially serious errors so quickly, survey firms can replace substandard interviews at relatively low cost, since enumerators are likely to still be in the field. Methods for detecting fraud that rely on patterns observable only in a complete sample may uncover serious problems that are much costlier to fix *ex post*. Given the challenges of sending enumerators back into the field in a second wave to address earlier mistakes, researchers may decide not to correct these problems, resulting either in a smaller or lower-quality sample than originally designed.

However, these methods pose their own challenges. They incur non-trivial time and overhead costs, since interview metadata must be audited continuously while fieldwork is ongoing. Recent studies have introduced statistical and computational techniques to ease this burden, typically by imposing distributional assumptions on the metadata and automatically identifying interviews which are anomalous under those assumptions. For instance, researchers can analyze incoming interviews for too little missingness: long survey instruments are unlikely to be consistently 100% complete, so scripts can automatically compute completion percentages for interviews arriving from the field and quickly flag those that have suspiciously few missing answers. Another group of widely-used automated methods are algorithms such as Percentmatch, which detect near-duplicate interviews and flag them as likely to be fraudulent in semi-real time (Kuriakose and Robbins 2016). Abnormal participation rates and interview duration, among other patterns in the metadata, can similarly be mined for clues about data fabrication (Birnbaum et al. 2012; Blasius 2018; Blasius and Thiessen 2012, 2018; Bredl, Storfinger, and Menold 2011; Murphy et al. 2004).

In addition to detecting outright fraud, scholars have deployed audit-based and automated methods for detecting genuine but low-quality interviews. While curbstoning is an obvious and evocative problem, smaller and less deliberate deviations from fieldwork protocols may be a greater problem for total survey error (TSE; Biemer and Lyberg 2003). For example, silent audio captures can be used to correct fieldwork mistakes and retrain enumerators, improving overall data quality (Bhuiyan and Lackie 2016). When even minimal information is passed back to enumerators—e.g., only whether their interviews had been accepted or rejected on quality grounds—it can be sufficient to improve data quality across the duration of a project (Gomila et al. 2017). Enumerators may even hew more closely to fieldwork protocols based solely on the knowledge that an auditor may be listening (Mitchell, Fahrney, and Strobl 2009).

Taken as a whole, these studies provide survey researchers with a large suite of tools to weed out low-quality data arising from fraud and enumerator error. Manual duplication checks, respondent re-contacting, GPS and audio auditing, and automated metadata parsing can identify enumerator deviations from fieldwork protocols. Yet there remains a key challenge facing survey researchers who wish to prevent low-quality interviews from creeping into their data: lack of evidence as to the relative merits of these tools.

In an ideal world, every survey would include a lengthy battery of quality control procedures. In reality, however, implementing these tools requires time and money, and survey researchers are typically extremely short on both. Faced with these resource constraints, they may wish instead to employ a narrower, more streamlined range of quality control checks. Yet the evidence for which tools are most efficient is scant, with very little scholarly research testing, validating, and assessing the generalizability of these checks. Even studies that have introduced innovative quality control procedures have typically tested them in isolation, and analysis of their utility is typically informal (Bhuiyan and Lackie 2016; Gomila et al. 2017; Mitchell, Fahrney, and Strobl 2009). There are good reasons for this lacuna: quality control checks are often proprietary, and there are few survey projects that can facilitate such a broad study. Nevertheless, without any aggregation of knowledge about these tools, researchers are left with little guidance on how to mobilize their resources efficiently. Scholars need to know each quality control procedure's contribution to reducing TSE, and its ability to complement other tools as part of a broader quality control package, in order to ensure a high-quality sample.

2 Quality control in the 2016/17 AmericasBarometer

We address this problem directly by documenting a full suite of 141 quality control procedures used in the 2016/17 round of the AmericasBarometer, and evaluating them with a large classification task. These data, shared with us by the Latin American Public Opinion Project, are unique for two reasons: they comprise a nearly-identical instrument across a large cross-national sample, and they include every quality control procedure of which we are aware.

The 2016/17 AmericasBarometer study's quality control system consisted principally of four levels, each of which provided an opportunity to cancel an interview deemed low-quality. In the field, enumerators who were unable to complete a particular interview indicated that the interview had "terminated early." Survey teams in each country then used trained auditors to listen to audio recordings captured during each interview. Next, auditors employed by third party firms or in LAPOP's central office ran spot checks such as reviewing interview logs and verifying the field team's auditing. Finally, a staff member at LAPOP's central office ran weekly (and sometimes daily) checks of interview metadata. Additionally, LAPOP conducted extensive enumerator training ahead of fieldwork to both reduce enumerator cheating and increase the overall quality of data collected.[5] We note that although interviews were thus removed from the data at different points in this process, by different actors, and using different

[5] . During these two-day training sessions, enumerators were informed that portions of the interviews would be recorded (though not *which* portions would be), and that their GPS location would be monitored for sample compliance. It is possible that our findings would change if enumerators were not pre-warned about LAPOP's auditing procedures. However, we think this unlikely, not least because interviewers still attempted to cheat in ways they knew had a high probability of detection (such as conducting interviews with no respondent present). Further, enumerators were unaware of the vast majority of quality control procedures, such as those relying on metadata, and so could not have adjusted their behavior accordingly.

information about quality, our analysis includes all quality control procedures for all interviews: even if an interview is terminated in the field, we are still able to evaluate whether the other checks would have flagged it as being potentially low-quality.

Our sample consists of every interview uploaded to LAPOP's primary software for CAPI interviews in 2016/17 (SurveyToGo, or STG), from countries where all quality control checks are available (Cohen and Larrea 2018).[6] The data consist of 13,253 interviews across Argentina, Bolivia, Chile, Guatemala, Haiti, Jamaica, Mexico, Peru, and Uruguay,[7] gathered between January 28 and June 2, 2017. In our sample, 933 observations (7%) were coded as 1 for canceled, with the remaining 12,320 coded 0 for published. This binary indicator is our outcome of interest.

We then matched these interviews to all metadata collected for the 2016/17 round. These include some 150,000 audio recordings and image captures. We also obtained the logs automatically generated by STG, which record the button presses and actions taken by enumerators during each interview, as well as silent actions such as GPS captures. For our covariates of interest, we used these data to code 141 distinct quality control variables, one for each procedure used to decide whether interviews are acceptable for publication in the AmericasBarometer. Full descriptions and coding rules for each variable are provided in the Appendix.

Broadly, these checks fall into three groups. First are 12 automatic flags in STG which screen interviews in real time. These include, for instance, whether a respondent terminated an interview early (as marked by the enumerator), and whether the username of the enumerator was different from that of the person who uploaded the interview to the server—which can indicate that an interview was started, stopped, and then restarted later in a different location or by a different enumerator, in violation of fieldwork protocols. Second are 65 variables generated by automatic parsing of metadata in R. These scripts evaluate interview quality in semi-real time, as they are run daily or weekly as data come in from the field. Such checks include *inter alia*: measures of cluster

[6] . The AmericasBarometer study was conducted in 29 countries; 20 of these countries do not include the complete suite of quality control procedures.

[7] . Our results are robust to using multiple imputation instead of listwise deletion. They are also robust to excluding auditor variables and using listwise deletion on the remaining sample, which is approximately 33,000 observations across 16 countries.

size and dispersion, to ascertain if the AmericasBarometer sampling procedure is being followed; Percentmatch; the quality of silent forward-facing enumerator photos, to ensure interviews are conducted only by trained enumerators; average question timings, to detect when enumerators skip items; whether devices in are airplane mode, to see when enumerators attempt to conceal their location; and GPS captures. The third group includes 64 checks coded manually by the auditors discussed above. These include information about such items as careful reading of the consent form, enumerators skipping or interpreting questions, the presence of enumerators who were not hired to work on the project, or evidence that the enumerator is otherwise not following fieldwork protocols. [8]

Most of these variables are binary: an observation is coded as 1 if the quality control procedure would have "caught" it as potentially low-quality, and 0 otherwise. For instance, one of the automated scripts generates a flag when GPS captures record a large geographic jump between consecutive attempted interviews. The metadata underlying this flag is measured in kilometers (calculated from GPS coordinates using spherical distance), and thus is continuous and nonnegative. An interview takes the value of 1 for this variable if the script detects a jump above a threshold, which LAPOP set at 10 kilometers for this study. Sixteen of the 141 variables cannot be discretized in this way or there was no clear threshold by which to separate flagged cases. These variables are kept untransformed and (semi-)continuous.[9]

3 Evaluating quality control with machine learning

Our primary goal is to rigorously evaluate which quality control procedures are most useful for identifying low-quality interviews. We want to know how well these methods predict whether an interview will be canceled due to quality concerns, as well as which variables are most useful for making those predictions. Variables that best separate high- and low-quality interviews are considered the most informative, and therefore the most

[8] . The auditing process also includes an open-form comment box for auditors to relay "other problems" encountered, which three research assistants coded into categorical variables.

[9] . These include checks which were developed late in the round, but were calculable from data collected in early-round countries, such as measures of geographic dispersion within sampling clusters.

valuable to survey researchers. Both of these quantities—predictive performance and variable importance—are key metrics produced by machine learning (ML), the science of learning patterns from data. We therefore study a supervised ML classification task in which a series of models separate interviews into canceled and published categories, using only the 141 covariates drawn from the AmericasBarometer quality control procedures. All analysis is conducted using the caret package in R (Kuhn 2008; Kuhn and Johnson 2013).

More specifically, we partition our final sample into training and validation sets, comprising 75% and 25% of the data, respectively. These sets preserve the marginal distributions of the outcome and all predictors. We then iterate through 36 models drawn from a variety of ML algorithms, including discriminant analysis, neural networks, decision trees, random forests, generalized linear models, and others, listed in the Appendix. Because our data are unbalanced, with the majority of observed outcomes being 0s, each model run begins by using the synthetic minority oversampling technique to achieve better balance (SMOTE; Chawla et al. 2002). We train each model using five-fold cross-validation, repeated five times, using common resampling indices across models. Each model's optimal hyperparameters are chosen by maximizing the area under the curve (AUC) across receiver operating characteristics (ROCs), a widely-used measure of classification accuracy, as computed during cross-validation. These optimal models are then fit to the training sample as a whole, variable importance summaries are computed, and the fitted models are used to predict outcomes on the validation sample which was held out from model training. We focus on interpreting results from the ten best-performing models due to space constraints, but the results are consistent across models.[10]

We are interested in two sets of quantities. The first are measures of predictive performance: how well quality control methods collectively screen out low-quality

[10] 36 models is likely excessive for this application, but we want to ensure that our results are not idiosyncratic to particular models (Fernández-Delgado, Cernadas, and Barro 2014). Further, these represent just a small subset of the approximately 200 models implemented in caret. Still, we dedicate most of our computing power to properly tuning hyperparameters instead of estimating additional models, because tuning has been found to exert a greater influence on overall performance than model choice (Bagnall and Cawley 2017).

interviews. Good predictions would indicate that standard quality control procedures are effective for recovering the AmericasBarometer sample, well-regarded in the academic and professional communities for its quality; bad predictions would indicate that they are ineffective, producing inconsistent information and leaving LAPOP heavily reliant on staff discretion in choosing to publish or cancel interviews. We expect predictive performance on par with other data-rich predictive models studied in political science, with AUCs of approximately 0.8-0.9 and root mean squared errors of prediction on the order of 0.2-0.3 (both on scales ranging from 0 to 1).[11]

To be clear, this modeling strategy does not assume that the AmericasBarometer sample is perfect, with publication and cancellation perfectly capturing high- and low-quality interviews, respectively. Instead, interview quality is better conceptualized as a latent variable generated by numerous factors. Yet as a "ground truth" for our classification task, this latent variable is essentially impossible to observe and measure, so we instead rely on publication versus cancellation in the AmericasBarometer as a reasonable approximation of overall quality.

Nor does this modeling strategy prime favorable results. Scholars may be concerned that because LAPOP makes cancellation decisions using the quality control checks we study, there may a deterministic link between these variables and the outcome of interest—necessarily yielding high predictive performance. However, a number of factors break this simple dependence. For one, because LAPOP allows enumerators to respond to conditions in the field, the team leaves considerable room for auditor discretion in deciding whether to publish an interview. Further, at no stage in the quality control workflow does any individual have access to the full suite of quality control checks; as discussed above, there are multiple points in the review process at which an interview may be rejected. These decisions to cancel or publish an interview are often made with less than ten variables to hand. Finally, many of the quality control checks we study were implemented *after* fieldwork was completed, that is, after all decisions to cancel or publish interviews were made. For example, the measure of geographic

[11] . These ranges are drawn from results obtained in recent applications of machine learning to political science data. See, for example, Bonica (2018), Cranmer and Desmarais (2017), Montgomery and Olivella (2018), Muchlinski et al. (2016), Neunhoeffer and Sternberg (2019), and Warner (2018).

dispersion within sampling clusters was only developed and implemented after the entire round was complete. In short, the structure of (and continual improvements to) LAPOP's workflow breaks any simple correspondence between quality control procedures and cancellation decisions.

The second set of quantities are measures of variable importance, which indicate how much information each quality control check provides for predicting interview quality. Each model computes variable importance differently,[12] but scales all 141 variables according to how useful they are for prediction, such that the most informative procedures are scored as 100 and completely uninformative procedures are scored as 0. Given the dearth of evidence comparing quality control procedures' efficacy, we do not have strong expectations about which methods will be most important for classifying interviews. Anecdotal and qualitative evidence suggests, however, that cost- and effort-intensive auditing are most useful for weeding out low-quality interviews (e.g., Mitchell, Fahrney, and Strobl 2009).

4 Does quality control work?

Table 1 reports performance measures for the ten models that predict the best out-of-sample, using the held-out validation set. Most notable is that these models perform very well: the AUCs appear far above the range typically observed in predictive models of political science data. Nor is this performance limited to just a few "best" models. 29 of our 36 models achieve an AUC within or above the expected range of 0.8-0.9. Scholars may be concerned that, as an improper scoring rule, the AUC is an imperfect measure of performance—especially when using techniques such as SMOTE to address class imbalance (Merkle and Steyvers 2013). Yet the Brier (1950) scores (RMSEs of prediction) indicate that this is not a concern here. The best model fits produce predicted probabilities of cancellation that are on the order of three percentage points, suggesting that on average, the best-performing models predict canceled (published)

[12] . For instance, random forests use permutation importance while the elastic net uses the absolute magnitude of coefficient estimates after rescaling predictors (after rescaling predictors; for details, see Kuhn 2008).

Table 1: Predictive performance for the ten best models

Model	AUC	Brier score (RMSE of prediction)	Precision (PPV)	Recall (Sensitivity)	Specificity	NPV
rf	0.98	0.03	0.79	0.78	0.98	0.98
parRF	0.97	0.03	0.82	0.72	0.99	0.98
pcaNNet	0.97	0.03	0.71	0.85	0.97	0.99
C5.0	0.97	0.04	0.80	0.82	0.98	0.99
multinom	0.96	0.04	0.63	0.79	0.97	0.98
RRFglobal	0.96	0.03	0.70	0.73	0.98	0.98
pda	0.96	0.03	0.71	0.75	0.98	0.98
glmnet	0.96	0.04	0.67	0.79	0.97	0.98
bayesglm	0.95	0.04	0.65	0.78	0.97	0.98
avNNet	0.95	0.04	0.69	0.79	0.97	0.98

All metrics are constrained to $[0, 1]$, with higher values indicating better performance for all except Brier scores. Alternative names for some performance measures are in parentheses.

interviews have a 97% (3%) chance of being low-quality—a small margin for a true validation sample of over 3,000 observations across nine countries. Thus, in general, we find that when flags, scripts, and auditors suggest a problem with an interview, it often has one; when they do not, it rarely does. Survey researchers have developed many methods to diagnose low interview quality, and very clearly, they work.

The accuracy of these predictions is underscored in Figure 1, which plots predicted probabilities of cancellation from the best model, rf (a random forest), against true outcomes for interviews held-out in the validation set. Predicted probabilities cluster near zero for the overwhelming majority of published interviews. While the predictions for canceled interviews are slightly more dispersed, the vast majority are well above the 50% cutoff typically used to generate a "hard-type" (binary) prediction.

To investigate this pattern more closely, we code all interviews which rf predicts will be canceled with 50% probability or higher as a predicted cancellation, taking the value 1, and anything with a lower predicted probability as a publication, taking the value 0. We compare these binary predictions against their true outcomes with a "confusion matrix," given in Table 2. For survey researchers interested in zeroing in on the interviews most likely to be of low-quality, these results suggest that quality control procedures can sort the wheat from the chaff very effectively. Just 3% of interviews are misclassified in this exercise. And while rf performs better at letting through high-quality interviews than screening out potentially low-quality interviews, this is entirely expected given the relative frequency of cancellation versus publication in the underlying data, as classification tasks generally struggle to predict outcomes that occur infrequently (He and Garcia 2009). The recall rates in Table 1 indicate that despite these limitations, most models detect just shy of 80% of low-quality interviews in the validation sample.

5 Which procedures are most useful?

While these results should be encouraging for researchers, very few surveys have the financial freedom and technical capacity to implement this full suite of quality control procedures. For many studies, a more pressing question is which tools are most useful

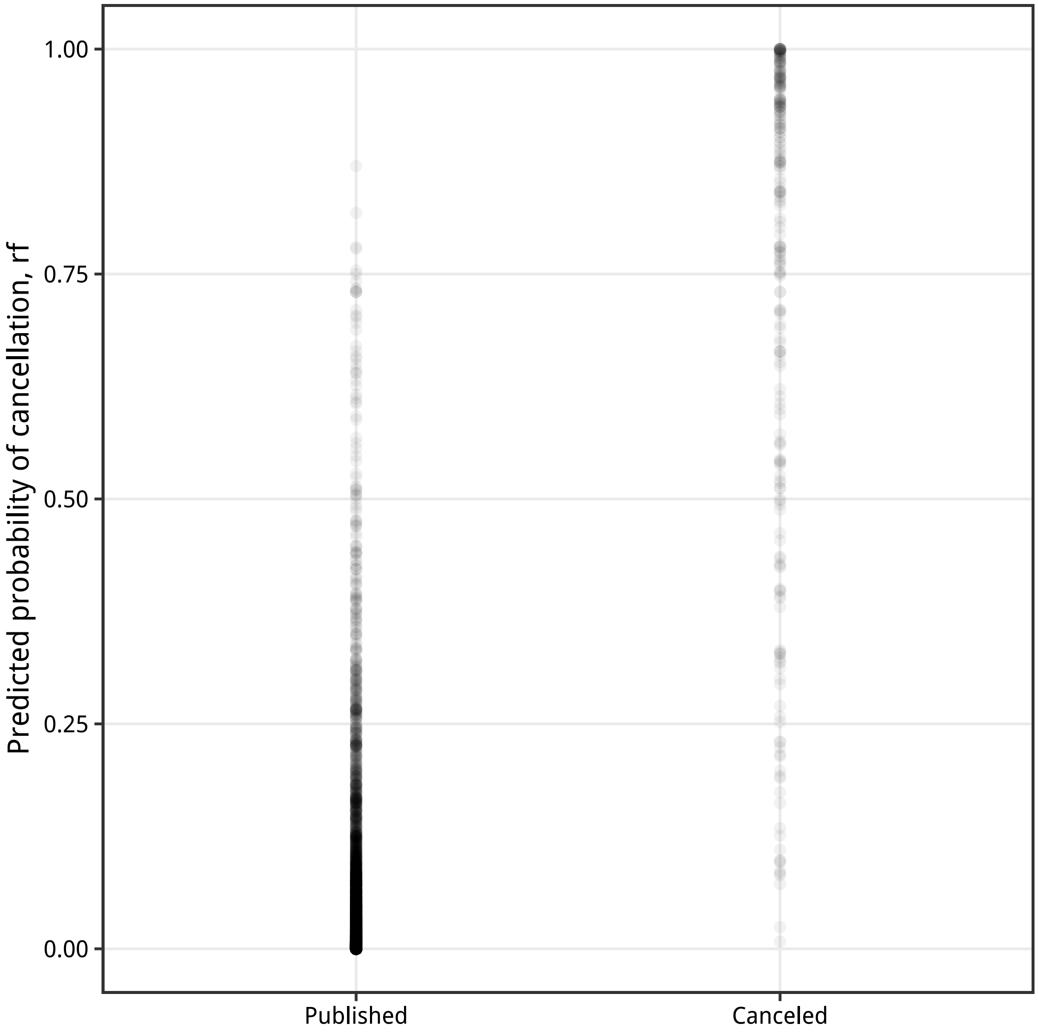

Figure 1: Predictive performance of rf. Each dot represents an interview in the validation set. The y-axis indicates the predicted probability of cancellation, with the x-axis giving the real-world outcome. Darker dots indicate more observations.

Table 2: Confusion matrix, rf

	True published	True canceled
Predicted published	3033	52
Predicted canceled	47	181

for recovering a high-quality sample. To answer this question, Figure 2 plots variable importance for each of the 141 procedures. Dots represent median values, and segments the interquartile range, across the ten best-performing models.

Our results indicate that while quality control measures work well as a whole, many procedures are essentially superfluous. Beyond the approximately 30 top performers, additional methods for detecting low-quality interviews add very little new information. This finding may be intuitive, since many of the procedures listed in the Appendix are close correlates of each other. For instance, given automated variables that calculate each interview's completion percentage and duration, it is unclear how much further information another variable to compute the average time spent on each question will add. However, at the same time, many of the "poor performers" would appear to add new information not otherwise captured by the more informative variables. For instance, a script which identifies large jumps in geolocation between attempted interviews does not provide much information to these models. It would seem that divergent GPS captures tell us more about satellite and mobile coverage than they do about interview quality.

Table 3 lists the 30 most informative quality control procedures. Among the interesting patterns that emerge, two important findings stick out. First, the most informative variables are a mix of those generated by flags, scripts, and auditors. That is, we find that no one approach to quality control is itself sufficient to ensure a high-quality sample, in line with anecdotal evidence. Effective quality control requires multiple passes at the data, taken in real and semi-real time. The consistency and speed of automated metadata parsing must be paired with the flexibility of auditor discretion (and ingenuity in identifying new problems as they arise) in order to get high-quality data. Our results suggest that if budget constraints truly bind, real-time STG flags may

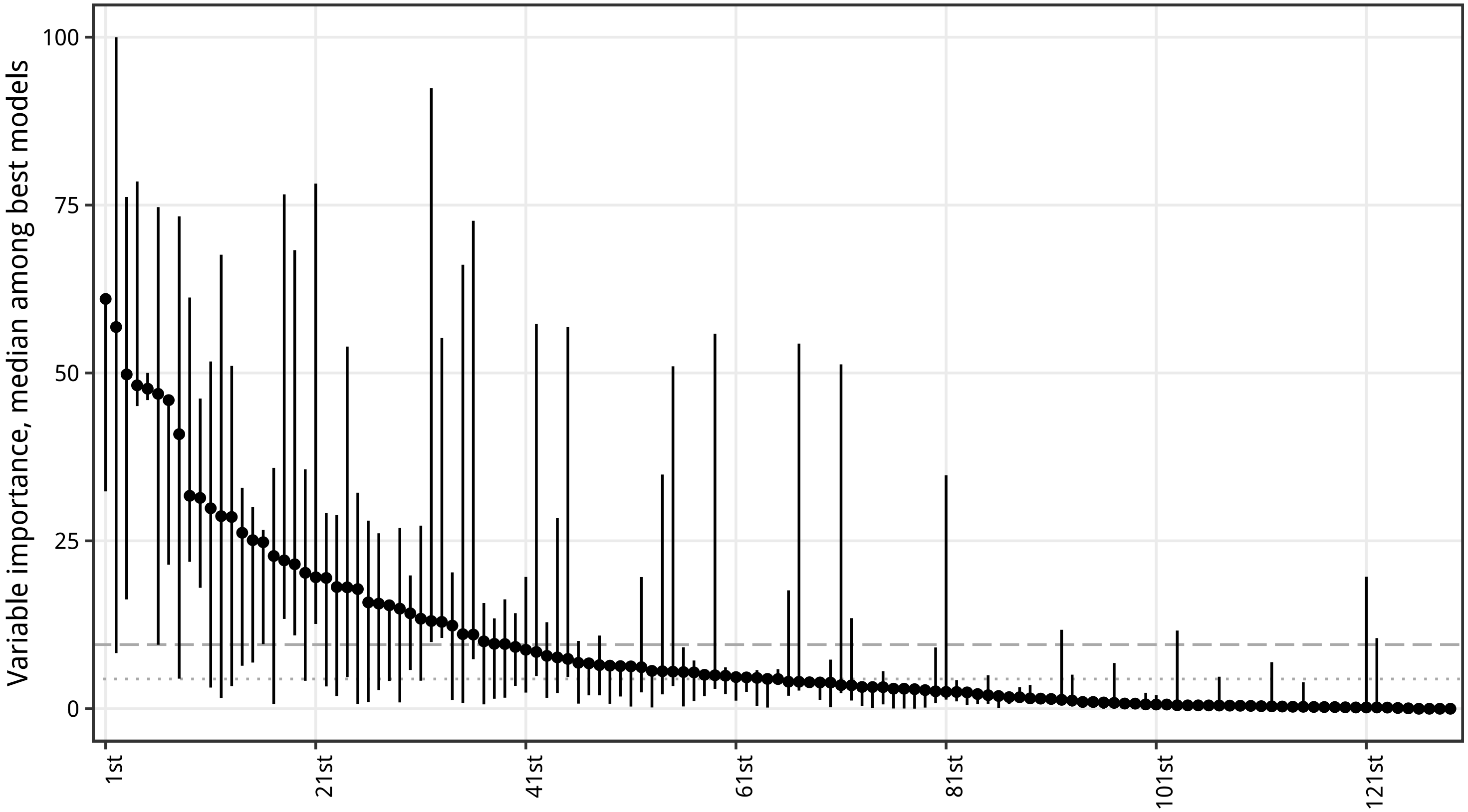

Figure 2: Variable importance for all quality control procedures across the ten best-performing models. Each dot represents a procedure's median variable importance (y-axis), ranked along the x-axis. Larger values indicate methods that provide more information for distinguishing high- and low-quality interviews. Segments represent the interquartile range for each variable across the ten models. The dashed (dotted) line indicates mean (median) variable importance. The long tail of the low-importance predictors indicates that many quality control checks provide very little added value.

Table 3: The most informative quality control procedures

1. Completion percentage (S)
2. Sampling cluster too big (S)
3. Interview duration, net (S)
4. Consent not read (A)
5. Interview terminated early (F)
6. Enumerator success rate (S)
7. One question skipped (A)
8. Enumerator “no one home” rate (S)
9. Two questions skipped (A)
10. Percentmatch (S)
11. Interview duration (S)
12. No real GPS captures (S)
13. Enumerator success, rural gap (S)
14. Enumerator refusal rate (S)
15. Interviewee abandoned (A)
16. One question interpreted (A)
17. Percentmatch, top decile (S)
18. No respondent heard (A)
19. Many questions skipped (A)
20. Sampling cluster dispersed (S)
21. Wrong location type (A)
22. Consent form incomplete (A)
23. Many questions misread (A)
24. Too short or too long (A)
25. GPS settings altered (S)
26. Other enumerator error (A)
27. “No one home” rate, rural gap (S)
28. One question misread (A)
29. Stopped and restarted (F)
30. Enumerator completion, rural gap (S)

Variables ordered according to their median variable importance as computed from the ten best-performing models. Letters in parentheses indicate whether the source of the information is an auditor check (A), STG flag (F), or R script (S). See the Appendix for details of each variable.

be foregone at a lower cost to quality than automated scripts and human auditors.[13] To test this intuition further, we re-ran each of our models three times using the same training and validation sets, but only the 12 STG flags, the 65 automated R scripts, and the 64 auditor checks as predictors. While the models relying only on information gleaned from scripts or auditors achieved AUCs in the range of 0.90, those drawing solely on STG flags were all below 0.75.

Second, the most informative variables are observed at different levels of analysis. Items like “interview terminated early” or “consent not read” are observed for each interview. On the other hand, “enumerator success rate” (the percent of attempts that result in interviews, by enumerator) is observed at the enumerator level. “Sampling cluster too big” is observed at each sampling cluster. And “Percentmatch” is observed

[13] The exception to the generally lower informational value of STG flags is, unsurprisingly, “early termination.” As discussed above, enumerators select this flag when a respondent refuses to continue with an interview; it therefore results in nearly-automatic cancellation for the AmericasBarometer.

across the entire sample. This finding provides clear evidence that interview quality is a function of multiple data-generating processes—not just enumerator fraud or error. Only by using an array of quality control procedures can researchers account for these complex causal pathways producing low-quality data.[14]

6 A more efficient quality control system

Because researchers do not typically have access to all 141 quality control procedures assessed here, and since many of these procedures appear to provide little information to our models, an important question is the extent to which sample quality would degrade if a more limited suite of quality control checks were used instead. To answer this question, we replicate our classification task using the same training and validation samples (and cross-validation resamples), but limiting the predictors to the top 30 most informative quality control procedures from Table 3. Our goal is to compare these models' predictive power against those estimated above. Large differences in performance would indicate that investment in a large number of quality control procedures is of primary importance, while smaller differences would suggest that a limited set of tools is more or less sufficient to recover a high-quality sample. To facilitate comparison, we present findings for the ten best-performing models from the main analysis. Table 4 provides the same performance metrics for these models as were presented in Table 1 for the main analysis.

The results indicate that limiting the predictors to a small set of quality control procedures may lead to somewhat worse overall performance, evident in the lower AUCs and higher Brier scores. However, this decline is relatively small, and predictive power still remains very good. Further, Table 4 reveals that some models are actually *better* at predicting cancellations with fewer predictors, evident in higher recall rates than those generated by models using the full suite of quality control procedures. Compared to the models studied above, these models correctly identify more true

[14] . This result highlights the limitations of developing quality control methods entirely within off-the-shelf software, since these programs typically only allow for checks which are observed at the interview level but not at the attempt, sampling unit, enumerator, or country level.

Table 4: Predictive performance for the ten best models, 30 most informative variables only

Model	AUC	Brier score (RMSE of prediction)	Precision (PPV)	Recall (Sensitivity)	Specificity	NPV
C5.0	0.94	0.06	0.59	0.79	0.96	0.98
parRF	0.94	0.05	0.62	0.75	0.97	0.98
rf	0.94	0.05	0.59	0.77	0.96	0.98
RRFglobal	0.94	0.05	0.55	0.77	0.95	0.98
pcaNNet	0.93	0.06	0.52	0.80	0.94	0.98
glmnet	0.92	0.08	0.41	0.72	0.92	0.98
multinom	0.91	0.08	0.40	0.75	0.91	0.98
bayeslgm	0.91	0.08	0.41	0.74	0.92	0.98
pda	0.91	0.06	0.48	0.70	0.94	0.98
avNNet	0.91	0.09	0.46	0.76	0.93	0.98

All metrics are constrained to $[0, 1]$, with higher values indicating better performance for all except Brier scores. Alternative names for some performance measures are in parentheses.

cancellations and let through fewer interviews that ended up canceled, despite a more limited set of covariates to use for prediction. The only cost to this improved performance is in false positives, with some 100 more interviews flagged for cancellation that ended up being published. Researchers may prefer the conservatism of this trade-off, accepting more false positives for better identification of interviews that might be problematic, despite the slight decline in overall performance.

A closer look at the interviews misclassified by these models provides further evidence that these procedures are sufficient to recover a high-quality sample. We randomly sampled 24 of the 179 misclassified cases and conducted a re-audit of each, where the experienced auditor was blind to both the real and predicted decision to cancel or publish the interview. Among this re-audited sample, we found no evidence of any systematic patterns that would suggest these quality control procedures are failing to identify a particular type of low-quality interview.

Further, among exactly half of these re-audited interviews, we found that the error was not with the models' predictions but rather with LAPOP's quality control workflow: in 12 of these 24 interviews, the models' prediction was upheld and the initial decision overturned.[15] In the other 12 interviews, the prediction was wrong and the initial decision was upheld. Although this sample is relatively small, the qualitative evidence suggests that these misclassified cases reflect statistical noise inherent to probabilistic models—noise which appears equal to that of the complete quality control process used by LAPOP. Limiting ourselves to just these 30 covariates does not appear to add in any error, systematic or otherwise, over a full quality control suite.

Taken together, these results suggest that researchers can produce very high-quality samples while relying on just a small sample of the available tools, significantly reducing their quality control effort compared to a full suite of procedures. Although few researchers have the resources or capacity to implement LAPOP's full quality control workflow, our results indicate that these limitations need not prevent them from

[15] . In the 2016/17 round, when LAPOP identified a particularly problematic enumerator, all of that interview's work was canceled—even interviews that appeared to be of high quality—because auditors indicated that a common cheating strategy was for enumerators to hide low-quality interviews in a batch of otherwise high-quality work. Our analysis suggests that this abundance of caution was likely unnecessary.

producing high-quality data.

7 Better data, more efficiently

One of the most important determinants of total survey error for large in-person household surveys is interview quality. Technological advances over the last decade have led to the rapid proliferation of tools for identifying and eliminating low-quality data to reduce TSE. Yet researchers seeking to implement these tools have little guidance on which procedures are cost- and time-effective, a pressing problem given that surveys are typically fielded under severe resource constraints. This paper provides the first steps toward solving this problem so as to help researchers produce more and better survey data.

We find that current tools are extremely effective in distinguishing high-quality interviews from low-quality interviews, as proxied by publication or cancellation in the 2016/17 round of the AmericasBarometer. However, they are also largely redundant: after dropping 111 of the 141 procedures we study, our models still predict interview quality as well as (and in some cases, better than) models fit using all of these variables. For survey researchers, the takeaway is clear: by implementing a limited, complementary set of quality control procedures, they can ensure a high-quality sample while freeing up resources to obtain more or richer data.

Our results indicate that researchers should implement quality control systems with two minimal characteristics. First, they should test for patterns that are observed at multiple levels of analysis, including indicators of fraudulent or low-quality data that can be detected in individual interviews, by enumerator, by sampling unit, and across the entire sample. Interview quality is determined by a number of distinct causal pathways; a workflow focused on just one level of analysis will necessarily miss some of these factors, leading to a lower-quality sample. Second, effective quality control systems should take multiple passes at the data, with automated flags that work in real time, regularly-scheduled scripts that analyze batches of interviews in semi-real time, and light auditing continuously throughout fieldwork. All of these steps are necessary to fully assess interview quality.

Beyond these broad findings, we suggest that researchers ensure they implement the quality control procedures identified in Table 3, which our results indicate are particularly useful for detecting interview quality. Among real-time flags automatically programmed into survey research software, we suggest only that researchers implement "early termination" or its equivalent. We recommend investing in the development of automated scripts to look for basic problems with required attachments (such as enumerator photos and GPS data); sample quota adherence (sampling cluster size and dispersion); enumerators failing to log unsuccessful attempts and incomplete interviews (e.g., interview success rates, refusal rates, and completion percentages); and complete or partial duplication (Percentmatch). Finally, among the checks auditors should carry out, we suggest listening for the consent of the interviewee, as well as random spot-checks to identify questions the enumerator may have skipped or interpreted.

Our results speak to the comparative effectiveness of quality control tools across the largest sample and the broadest range of countries of which we are aware. Yet they come with two important caveats. First, not all surveys will share LAPOP's definition of "low-quality." The AmericasBarometer assigns greater importance to some checks than others might; for instance, while informed consent is of primary importance to this study, this criterion may be less critical to other researchers. While we view LAPOP's weighting of these priorities as generally applicable, we nonetheless encourage researchers to keep their own research priorities in mind while implementing these general recommendations. Second, we emphasize that these data do not allow us to measure interview quality directly. Our empirical strategy relies on the coarse proxy that is interview cancellation versus publication. Quality is a much more nuanced concept than this binary measure can capture, incorporating enumerator characteristics, respondent features, and contextual factors. Our goal is not to generate a fine-grained measure of interview quality; rather, we seek to help scholars efficiently identify interviews of sufficiently low quality that they merit rejection.

Implementing the procedures we identify is a feasible, minimalistic approach to increase the baseline quality of large in-person household surveys. Yet there remain a number of ways by which researchers can reduce TSE in these surveys. Most obviously, they can develop better quality control procedures. Many of the tools studied here were implemented *ad hoc* to combat specific behaviors observed in the field, but can be fine-tuned to better identify potential problems. We also encourage researchers to continue to invest in the development of entirely new tools. For instance, the process of auditing interviews for "no respondent" and other problems could be partially automated by creating scripts to analyze audio captures and identify how many voices can be heard. Although our results indicate that many procedures add little value for determining interview quality, they do not suggest that researchers should stop innovating.

There is also ample room for researchers to empower the users of their data to make their own quality assessments. Current practice is to indirectly communicate quality through the binary decision to publish or cancel an interview. This approach makes low-quality interviews effectively unobservable to most scholars, leaving them with no options but to trust that the data are reliable (or to not use them). Researchers may instead develop a measure of quality as a latent variable informed by the outcomes from a series of quality control procedures. They would then be able to publish all data collected, including their measure of quality, and allow users of the data to decide which data are of sufficient quality to use in their specific analyses. Even better, researchers may provide the quality control variables, so that scholars can interrogate the measure of latent quality itself, and potentially adjust it to better suit their own uses. Such freedom can help scholars use these surveys more effectively.

Finally, in this vein, researchers can do much more to make their quality control workflow more transparent. The American Association for Public Opinion Research's Transparency Initiative has called on major survey research institutions to routinely disclose methodological information. This initiative has led to advances in areas such as sampling frames and response rates; however, quality control procedures remain mostly private. By publicizing their quality control workflow, researchers can contribute to better scholarly understanding of survey research methods, and ultimately, more credible social science.

References

Bagnall, Anthony, and Gavin C. Cawley. 2017. On the Use of Default Parameter Settings in the Empirical Evaluation of Classification Algorithms. ArXiv: 1703.06777v1, March.

Bennett, Archibald S. 1948. "Toward a Solution of the 'Cheater Problem' among Part-Time Research Investigators." *Journal of Marketing* 12 (4): 470–474.

Bhuiyan, Muhammad F., and Paula Lackie. 2016. "Mitigating Survey Fraud and Human Error: Lessons Learned from a Low Budget Village Census in Bangladesh." *IASSIST Quarterly* 40 (3): 20–26.

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to Survey Quality.* Hoboken, NJ: Wiley.

Birnbaum, Benjamin, Brian DeRenzi, Abraham D. Flaxman, and Neal Lesh. 2012. Automated Quality Control for Mobile Data Collection. Proceedings of the 2nd ACM Symposium on Computing for Development, Atlanta, GA, March 11-12.

Blasius, Jörg. 2018. "Fabrication of Interview Data." *Quality Assurance in Education* 26 (2): 213–226.

Blasius, Jörg, and Victor Thiessen. 2012. *Assessing the Quality of Survey Data.* London: Sage.

———. 2018. "Perceived Corruption, Trust, and Interviewer Behavior in 26 European Countries." *Sociological Methods & Research,* no. forthcoming.

Bonica, Adam. 2018. "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning." *American Journal of Political Science,* no. forthcoming.

Bredl, Sebastian, Nina Storfinger, and Natalja Menold. 2011. A Literature Review of Methods to Detect Fabricated Survey Data. Discussion paper no. 56, Zentrum für internationale Entwicklungs- und Umweltforschung, ZEU, Giessen.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16:321–357.

Cohen, Mollie J., and Sebastian Larrea. 2018. "Assessing and Improving Interview Quality in the 2016/17 AmericasBarometer." *AmericasBarometer.*

Cranmer, Skyler J., and Bruce A. Desmarais. 2017. "What Can We Learn from Predictive Modeling?" *Political Analysis* 25 (2): 145–166.

Crespi, Leo P. 1945. "The Cheater Problem in Polling." *Public Opinion Quarterly* 9 (4): 431–445.

Fernández-Delgado, Manuel, Evan Cernadas, and Senén Barro. 2014. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research* 15 (1): 3133–3181.

Gomila, Robin, Rebecca Littman, Graeme Blair, and Elizabeth Levy Paluck. 2017. "The Audio Check: A Method for Improving Data Quality and Detecting Data Fabrication." *Social Psychological and Personality Science* 8 (4): 424–433.

He, Haibo, and Edwardo A. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263–1284.

Heath, Anthony, Stephen Fisher, and Shawna Smith. 2005. "The Globalization of Public Opinion Research." *Annual Review of Political Science* 8:297–333.

Hicks, Wendy D., Brad Edwards, Karen Tourangeau, Brett McBride, Lauren D. Harris-Kojetin, and Abigail J. Moss. 2010. "Using CARI Tools to Understand Measurement Error." *Public Opinion Quarterly* 74 (5): 985–1003.

Hill, Daniel W., Jr., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108 (3): 661–687.

Kuhn, Max. 2008. "Building Predictive Models in R using the caret Package." *Journal of Statistical Software* 28 (5): 1–26.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling.* Berlin: Springer.

Kuriakose, Noble, and Michael Robbins. 2016. "Don't Get Duped: Fraud through Duplication in Public Opinion Surveys." *Statistical Journal of the IAOS* 32 (3): 283–291.

Lupu, Noam, and Kristin Michelitch. 2018. "Advances in Survey Methods for the Developing World." *Annual Review of Political Science* 21:195–214.

Merkle, Edgar C., and Mark Steyvers. 2013. "Choosing a Strictly Proper Scoring Rule." *Decision Analysis* 10 (4): 292–304.

Mitchell, Susan, Kristine Fahrney, and Matthew Strobl. 2009. Monitoring Field Interviewer and Respondent Interactions Using Computer-Assisted Recorded Interviewing: A Case Study. Paper presented at the annual conference of the American Association for Public Opinion Research (AAPOR).

Montalvo, J. Daniel, Mitchell A. Seligson, and Elizabeth J. Zechmeister. 2018. "Improving Adherence to Area Probability Sample Designs: Using LAPOP's Remote Interview Geo-locating of Households in real-Time (RIGHT) System." *AmericasBarometer.*

Montgomery, Jacob M., and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–744.

Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocker. 2016. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24 (1): 83–103.

Murphy, Joe, Rodney Baxter, Joe Eyerman, David Cunningham, and Joel Kennet. 2004. A System for Detecting Interviewer Falsification. Paper presented at the annual conference of the American Association for Public Opinion Research (AAPOR).

Neunhoeffer, Marcel, and Sebastian Sternberg. 2019. "How Cross-Validation Can Go Wrong and What to Do About It." *Political Analysis* 27 (1): 101–106.

Sarracino, Francesco, and Małgorzata Mikucka. 2017. "Bias and Efficiency Loss in Regression Estimates Due to Duplicated Observations: A Monte Carlo Simulation." *Survey Research Methods* 11 (1): 17–44.

Seligson, Mitchell, and Daniel E. Moreno Morales. 2015. "Improving the Quality of Survey Data Using CAPI Systems in Developing Countries." In *The Oxford Handbook of Polling and Polling Methods,* edited by Lonna Rae Atkeson and R. Michael Alvarez. Oxford: Oxford University Press.

Warner, Zach. 2018. "Divide to Rule: Deconcentration as Coalition Manipulation." PhD diss., University of Wisconsin–Madison.

Winker, Peter. 2016. "Assuring the Quality of Survey Data: Incentives, Detection and Documentation of Deviant Behavior." *Statistical Journal of the IAOS* 32 (3): 295–303.