QualMix: Using Mixture Models to Assess Survey Quality

Simon Hoellerbauer[*]

May 8, 2022

Abstract

When we work with surveys in the social sciences, we are often unsure about the quality of the data collected by third-party actors, such as survey firms. Consequently, researchers typically either assume away problems of data quality or discard any data where doubts exist. This is costly in monetary terms and for analysis. Part of the issue is the inability to measure data quality effectively. To address the issue of quality measurement, I propose the QualMix model, a mixture modeling approach to derive estimates of survey data quality in situations in which two sets of responses exist for all or certain subsets of respondents. I apply this model to the context of survey backchecks. Through simulation based on real-world data, I demonstrate that the model successfully identifies incorrect observations and recovers latent enumerator and survey quality. I further demonstrate the model's utility by applying it to data from a large survey in Malawi, using it to identify significant variation in data quality across observations generated by different enumerators.

[*] PhD Candidate, University of North Carolina at Chapel Hill. Email: hoellers@unc.edu

1 Introduction

Researchers are usually neither the primary collectors of their data nor observers of the majority of the data collection process. As a consequence, they are often unsure about the quality of their data. This is particularly true when it comes to surveys — the assumption is generally that the information obtained about a respondent is actually from that respondent *and* is accurate. But how sure of that can we be? Even if we are not sure, what do we do about it? Data quality issues can induce measurement error, which in turn can bias analyses and lead researchers to draw incorrect conclusions. Often, the only solution when we are not sure about data quality is to drop observations or to to ignore the issue altogether, but both can be costly in economic and in analytic terms.

A large subset of the literature on survey data quality seeks to assess two core data quality concerns: data falsification (Murphy et al. 2016; De Haas and Winker 2014; Bredl et al. 2013; Forsman and Schreiner 1991; Schreiner et al. 1988; Crespi 1945) and data reliability (Tourangeau 2021; Alwin 2016; of Survey Quality 2016; Blasius and Thiessen 2012; Alwin 2011; Madans et al. 2011). A lot of this work, however, looks at only individual survey items or at aggregated levels (producing a single quality measure for the whole survey, for example). Furthermore, there is little agreement about how to assess survey data quality (Tourangeau et al. 2021). As a consequence, it is unclear how to incorporate uncertainty about data quality from existing methods into subsequent analyses. In this paper, I propose QualMix, a general approach to assess survey data **qual**ity using **mix**ture models in situations in which researchers have two sets of information, ostensibly from the same respondent. In addition, the proposed method allows us to estimate uncertainty about data quality at the observation level and at the survey level.

I first summarize the literature on survey quality, focusing on data reliability – data issues that will induce measurement error – and discussing present approaches to dealing with data quality concerns. I then describe the general QualMix model, which relies on the logic underpinning probabilistic record linkage (Enamorado et al. 2018; Fellegi and Sunter

1969). Next, I apply the QualMix model to a specific case: backchecks, also called re-interviews. I use a simulation study to show that the model can accurately identify matches and non-matches and can successfully estimate enumerator quality. Finally, I use the model in a real-world context, estimating survey and enumerator quality for a large survey carried out in Malawi.

The simulations and real-world-use example show that the model gives informative estimates of survey quality in the context of backchecks. The simulations additionally demonstrate that the method works well even when only a small proportion (5%) of responses is chosen for backchecking. Using QualMix to assess survey data quality is not meant to replace other approaches to estimating survey response quality.[1] Yet, it streamlines and makes less arbitrary a step that is already part of researchers' and survey firms' quality assessment workflow. In addition, it provides respondent-level (and potentially enumerator-level) summary assessments that can be incorporated into analysis.

2 Reliability and Survey Quality

There are many sources of data quality issues with surveys. Measurement error—mismatches between respondents' "true" responses and collected responses—can result from respondent satisficing, mode effects, implementer policies, poorly thought out questions, survey data fabrication, and low quality enumerators. Unfortunately, "[w]hile there has been considerable conceptual work regarding the measurement of [survey data] validity, translating the concepts into measurable standards has been challenging" (Madans et al. 2011, 2) In response to this, Alwin (2016) has proposed that "*the reliability of measurement should be used as a major criterion for assessing differences in measurement quality*" (3, italics in original). Reliability refers to "agreement between two efforts to asses the underlying value using *maximally similar*, or replicate, measures" (Alwin 2016, 7). Alwin points out that without data reliability, there cannot be data validity. In other words, without accurately

[1] For example, this method may not be optimal for assessing issues of data quality due to lack of concept or construct validity.

recording data, it is difficult to make judgments about whether data reflect concepts.

Measurement error can seriously impact analyses: Figure 1 shows bias in linear regression coefficients on predictors from a survey as proportion ($\in [0, 1]$) of the original effect size with varying levels of simulated measurement error.[2] Even when there is only a small amount of measurement error, the parameter estimates are clearly biased, sometimes unpredictably so. Although recent work in political science (Castorena et al. 2021) has shown that if measurement error produces data that are similar to the true data, bias may not be pronounced, this simulation *does* produce observations with measurement error that are overall very similar to existing ones in the survey. In addition, the simulation that produced this figure did not have measurement error vary with respondent characteristics, which can introduce even more bias.

At their core, the most common ways to assess measurement error caused by data quality concerns (i.e. a lack of reliable data) rest on the idea of repeated measures — measure the same item in (mostly) the same way, repeatedly, and check for deviations. However, the most common versions of repeated measurements focus on within respondent reliability—internal consistency approaches (Bohrnstedt 2010) —require multiple questions to assess the reliability of one survey item—multitrait-multimethod approaches (Alwin 2007)—require longitudinal data with more than three waves—quasi-simplex models (Alwin 2007)—or can be overly simplistic, such as the gross difference rate (Tourangeau et al. 2021). There are other recently proposed methods for detecting low-quality data, such as applying supervised machine-learning to survey para- and metadata (Cohen and Warner 2021) and checking for duplicates (Kuriakose and Robbins 2016). However, while both useful, the former relies on existing truth data to train a model, and the latter is more suitable for finding fabricated data, rather than assessing quality outright.

While it *is* important to assess the reliability for single survey questions, it can become too time intensive to scale this approach to a survey as whole, as it would require extended

[2] See Appendix E.2 for a description of how this data was simulated.

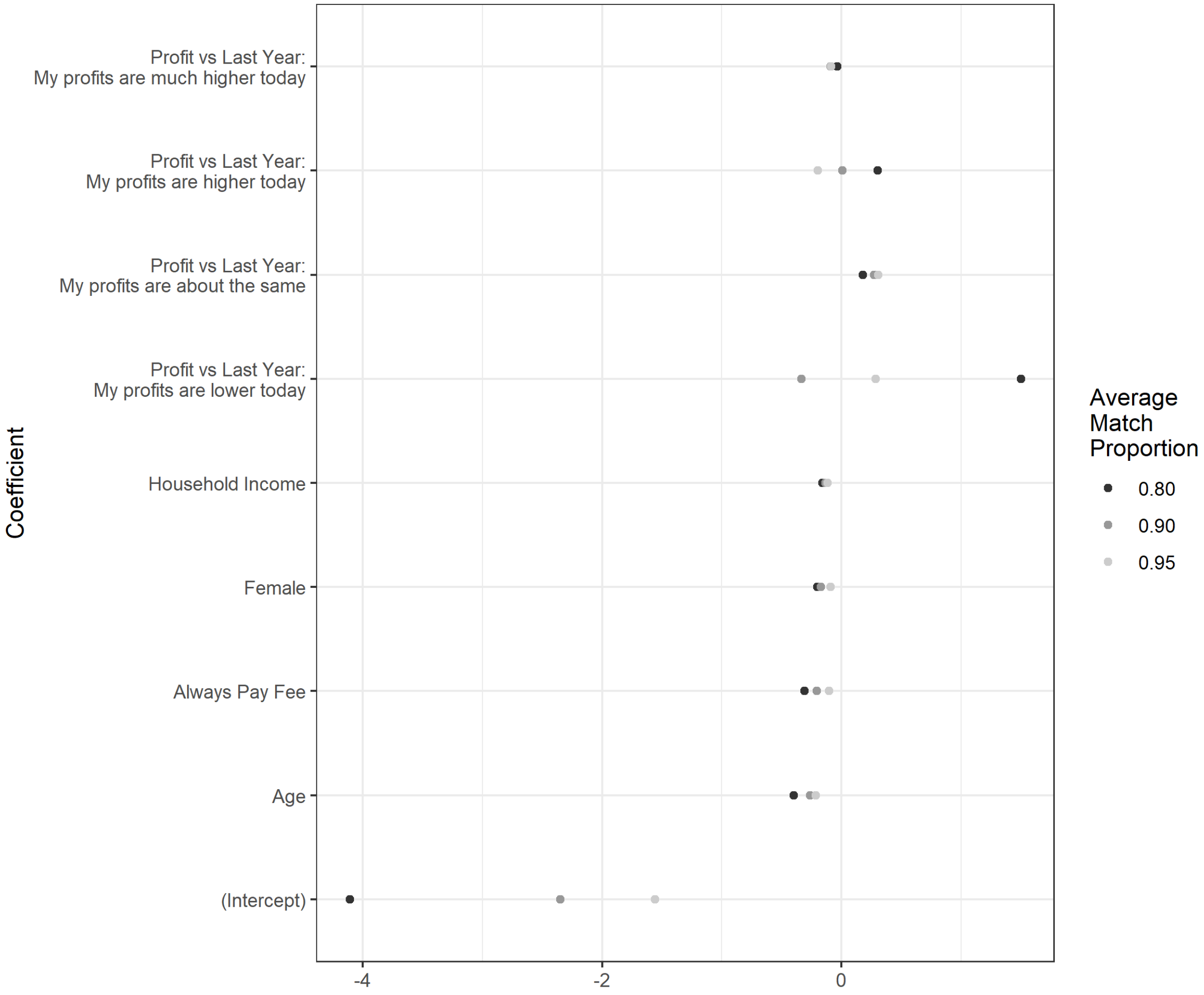


Figure 1: Linear Regression Coefficient Bias as Proportion of Original Effect Size. Regressions used the correct model specification on data where measurement error had been introduced during the simulations used to assess the QualMix model. See Appendix E.2 for more information on this process.

analysis for many questions. This disincentivizes researchers and survey implementers from using them (Madans et al. 2011, 2). Therefore, we need approaches for estimating general data quality in surveys that are easier to implement on a larger scale. The QualMix model I propose here builds on the idea of repeated measurements but applies to sets of questions, as opposed to individual questions. There are many scenarios in which implementers will have repeated measurements by design. One such scenario is the use of backchecks.

2.1 Backchecks

Backchecks—also called re-interviews, recontacts, callbacks or field audits—form a core part of the data quality assessment strategy at most major survey firms that do interviewer-administered surveys (Tourangeau et al. 2021; Murphy et al. 2016). For example Innovations for Poverty Action (IPA) includes re-interviews (backchecks) in their "Minimum Must Dos" for "every research project at IPA" (IPA 2018). The World Bank states that "[b]ack checks are an important tool to detect fraud" and "help researchers assess accuracy and quality of the data collected" (DIME n.d.). The U.S. Census uses re-interviews extensively as part of its quality assessment procedures (Schreiner et al. 1988; Forsman and Schreiner 1991; Krejsa et al. 1999). Forsman and Schreiner (1991) explain that re-interviews can be used to "evaluate field work" and "estimate error components in a survey model" (280-281). Conceptually, therefore, re-interviews can be used to assess survey quality *and/or* find falsified data. This is reflected by the protocols developed by IPA and the World Bank. Nevertheless, the survey literature primarily discusses re-interviews in the context of finding falsified data, starting with Crespi (1945).

Random re-interviews as a way to identify data falsification by enumerators is inefficient (Schreiner et al. 1988; Krejsa et al. 1999; Bredl et al. 2013). Random sampling might lead to too many "good" enumerators being chosen for backchecking. As such, survey analysts and statisticians have proposed a series of methods for detecting interviewer falsification without using backchecks, relying instead on paradata and characteristics of the response data, such as applying Benford's Law to numeric data entries. Researchers have suggested using these features in logistic regression (Li et al. 2011), unsupervised clustering algorithms (De Haas and Winker 2014, 2016; Rosmansyah et al. 2019), and random forests (Birnbaum et al. 2013). Each of these methods shows promise for identifying observations that may be fraudulent.

Falsified data is a core concern, as multiple studies have shown that it can bias results, especially when used in multivariate analysis (Schnell 1991; Schräpler and Wagner 2005; Ahmed et al. 2014; Finn and Ranchhod 2017; Sarracino and Mikucka 2017). However, another expressed goal of using re-interviews is a more general quality assessment. How can researchers and survey implementers use the statistical models proposed to identify faking enumerators to generate statements about the quality of a survey as a whole? Little work has been done on how to analyze backchecks effectively from this perspective. Partly, this is because quality control at major survey firms are often proprietary and not open to researcher or public scrutiny (Cohen and Warner 2021, 124). Forsman and Schreiner (1991) discuss "reconciliation"—that is, finding out which information is correct if there are disagreements—in their in-depth look at re-interviews, but do not offer advice on how to use the re-interview information itself to measure quality. IPA has developed the very helpful Stata (StataCorp 2019) package *bcstats* (White 2016) to help with analyzing re-interviews, but it only helps identify mismatches in a deterministic, not probabilistic, way. It also does not offer a simple way of summarizing these mismatches or generating uncertainty about whether two sets of information match.

3 QualMix: Survey Quality and Mixture Models

This section describes the general approach and the probabilistic model behind QualMix. The QualMix model can be used to assess overall quality *and* to detect falsified data when applied to backchecks and can also generate uncertainty estimates about the quality of individual observations. The method is inspired broadly by the probabilistic record linkage

	Survey Questions			
	Last Name (*String*)	Monthly Income (*Ordered*)	Occupation (*Categorical*)	Age (*Continuous*)
Response Set $\boldsymbol{R_a}$				
$\boldsymbol{r}_{\boldsymbol{a}1}$	Melzer	[\$250, \$500) (2)	Market Vendor	65
$\boldsymbol{r}_{\boldsymbol{a}2}$	Karlsen	<\$250 (1)	Market Vendor	21
Response Set $\boldsymbol{R_b}$				
$\boldsymbol{r}_{\boldsymbol{b}1}$	Beier	<\$250 (1)	Tax Collector	57
$\boldsymbol{r}_{\boldsymbol{b}2}$	Karls	<\$250 (1)	Business Owner	31
Agreement Vectors				
γ_1	Complete Disagreement	Complete Disagreement	Complete Disagreement	Similar
γ_2	Complete Agreement	Complete Agreement	Similar	Complete Disagreement
Agreement Summary Vectors	**Agreement Levels**			
	Complete Disagreement	Similar	Complete Agreement	Sum (K)
$\boldsymbol{\nu}_1$	3	1	0	4
$\boldsymbol{\nu}_2$	1	1	2	4

Table 1: Example of General Approach

model proposed by Fellegi and Sunter (1969; see also Enamorado et al. (2018)).[3]

3.1 General Approach

Suppose that for n survey respondents we have two sets of responses to the same K questions, $\boldsymbol{R_a}$ and $\boldsymbol{R_b}$, both with dimensions $n \times K$, where $\boldsymbol{r}_{\boldsymbol{1}i}$ and $\boldsymbol{r}_{\boldsymbol{2}i}$ represent the two response vectors for respondent $i, \forall i = 1, \ldots, n$. We can compare the values for the k-th question by looking at $\boldsymbol{r}_{\boldsymbol{a}k,i}$ and $\boldsymbol{r}_{\boldsymbol{b}k,i}$. If we define information about the agreement or disagreement between $\boldsymbol{r}_{\boldsymbol{a}k,i}$ and $\boldsymbol{r}_{\boldsymbol{b}k,i}$ as γ_{ik}, we can create a length-K agreement vector $\boldsymbol{\gamma}_i$. We can discretize the information about agreement or disagreement for each question into L ordered categories, which I term agreement-levels. For example, if $L = 3$, we could set $1 =$ complete disagreement, $2 =$ similar, $3 =$ complete agreement. Because each element of $\boldsymbol{\gamma}_i$ has the same number of possible levels, we can count up the number of times each level appears in $\boldsymbol{\gamma}_i$. This results in a length L *agreement summary* vector $\boldsymbol{\nu}_i$, the entries of which will add up to K.

[3] However, in contrast to probabilistic record linkage, the aim here is to identify potential non-matches where identifiers for two sets of responses already exist.

Turning the comparison information into L agreement-levels requires pre-specified decision rules, which may be different for different variable types.[4] Table 1 presents a concrete hypothetical example of the general approach, with examples of four different variable types 1) strings, 2) ordered categorical, 3) unordered categorical, and 4) continuous numeric. I first set $L = 3$, for "Complete Agreement," "Similar," and "Complete Disagreement."

3.2 QualMix Model

If data quality issues exist, we can think of two *clusters* of agreement summary vectors: one with more agreements — like $\boldsymbol{\nu}_2$ in Table 1 — and one with more disagreements — like $\boldsymbol{\nu}_1$ in Table 1. We can think of these two clusters as representing high quality and low quality data, respectively. Not all agreement summary vectors for sets of high quality responses will consist of *only* complete agreements (due to random chance and sporadic data entry mistakes), nor will agreement summary vectors for sets of low-quality responses consist of *only* complete disagreements.[5] This complicates what we do with the agreement summary vectors. We could use them deterministically, by establishing another decision rule. For example, if $K = 4$ and $L = 3$, we could say that agreement summary vectors with at least three complete agreements represent matches between $\boldsymbol{r_{a}}_i$ and $\boldsymbol{r_{b}}_i$. However, such a decision rule is highly arbitrary and becomes harder to make as the number of questions and agreement categories grow. With deterministic methods, it is hard to determine what to do with fringe cases — in our example, how do we categorize an agreement summary vector with two complete agreements and two similar values? Even once we have made the decisions, it is hard to conceptualize our uncertainty about their validity. The only solution we are left with is to drop observations we are unsure about. This potentially results in wasted observations, a reduction of power, and selection problems.

[4] See App. A for an in-depth description of the decision rules used to create this table and in the applications in this paper.

[5] A low-quality agreement summary vector does not necessarily represent fabricated data. In analytic terms there is no distinction between a falsified response and one full of errors — both induce measurement error.

The solution is to take a probabilistic approach. We can use the agreement summary vectors as the data for a two-component finite mixture model (McLaughlan and Peel 2000), resulting in the following model

$$\boldsymbol{\nu}_i | Q_i = q \overset{\text{i.i.d}}{\sim} \text{Multinomial}(\boldsymbol{\pi}_q)$$
$$Q_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\lambda)$$

where $q = 1$ when the two response vectors generally match (are of high quality) and $q = 0$ when they do not (are of low quality), λ characterizes the overall probability that the agreement summary vectors from $\boldsymbol{R_a}$ and $\boldsymbol{R_a}$ are high quality or not, and $\boldsymbol{\pi}_m$ is an L length vector of the agreement-level probabilities for distribution q. The probabilistic structure—and the distribution of the individual elements of the agreement summary vector—make it possible that pairs of matched observations can fail to coincide exactly on some of variables of interest, yet still count as high-quality.[6]

The observed-data likelihood for this model is

$$\mathcal{L}(\boldsymbol{\Pi}, \lambda | \{\boldsymbol{\nu}_i\}_{i=1}^N) \propto \prod_{i=1}^{N} \left(\sum_{q=0}^{1} \lambda^q (1-\lambda)^{1-q} \prod_{l=1}^{L} \pi_{ql}^{\nu_{il}} \right)$$

It is possible to estimate the model parameters using the Expectation-Maximization (EM) algorithm or, as I do in the empirical applications below, in a Bayesian framework. If all $\boldsymbol{\nu}_i$ seem to come from the same distribution, then the estimated λ will be close to 1.

We can also estimate the observation-specific probability that observation i represents a high-quality observation using the posterior probability of coming from the high-quality component. Intuitively, this is just the amount that the observation i contributes to the

[6] An inherent risk with any unsupervised learning approach is that the model may overfit and find patterns in the data that may not exist in reality. Thus, it is important to inspect the parameter estimates for the discovered distributions. See Appendix C for recommendations on diagnosing issues.

likelihood when $Q_i = 1$ divided by observation i's total contribution to the likelihood:

$$\xi_i = \Pr(Q_i|\boldsymbol{\nu}_i) = \frac{\lambda \prod_{l=1}^{L} \pi_{1l}^{\nu_{il}}}{\sum_{q=0}^{1} \lambda^q (1-\lambda)^{1-q} \prod_{l=1}^{L} \pi_{ql}^{\nu_{il}}}$$

I discuss in Section 3.3 how these posterior probabilities can be used as a measure of the quality of observation i.

This model is flexible: we can also incorporate respondent-level characteristics or survey metadata into the model.[7] For example, in the case of re-interviews, we can incorporate information on interviewers, if the survey mode is interviewer-implemented. In this case, the extended model becomes:

$$\boldsymbol{\nu}_i|Q_i = q \overset{\text{i.i.d}}{\sim} \text{Multinomial}(\boldsymbol{\pi}_q)$$

$$Q_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\lambda_e)$$

$$\lambda_e = \text{logit}^{-1}(\beta_0 + \beta_e)$$

$\text{logit}^{-1}(\beta_0)$ in this context represents the overall probability of a match, and the intercepts by enumerator (β_e) represent the deviations from this probability. λ_e represents the probability that $\boldsymbol{r}_{a_i}$ and $\boldsymbol{r}_{b_i}$ — $i \in I_e$ — match, i.e. that observations associated with enumerator e are of high quality. In short, we now have E different λ's. The benefit of this approach is that it allows for match probability to vary by enumerator.

3.3 Quantities of Interest: Assessing Survey Quality

QualMix can be used to assess different aspects of survey data quality. The general approach is a test-retest measure. As such, it is best situated to assess questions of reliability — how often do repeated measurements return the same response? The posterior probability of a match ξ_i encapsulates how likely it is that $\boldsymbol{r}_{\boldsymbol{a}i}$ and $\boldsymbol{r}_{\boldsymbol{b}i}$ are actually the same — they vary

[7] See Appendix B for an expanded discussion of this and other extensions to the general model.

from 0 to 1. We can designate the *mean* of $\boldsymbol{\xi}$ as an indicator of overall survey data quality

$$Q_S = \frac{\sum_i^N \xi_i}{N}$$

This quantity will also vary between 0 and 1; a 1 indicates that all agreement summary vectors represent high-quality data points, and a 0 would indicate that all agreement summary vectors represent low-quality entries.

$\hat{Q_S}$ and $\hat{\boldsymbol{\xi}}$ are our estimates of survey data quality and our confidence in the quality of an individual observation. We use the average of the posterior probability of a match to estimate survey quality instead of $\hat{\lambda}$ because the former quantity incorporates the actual data as well.[8] How we interpret these estimated quantities substantively depends on the types of questions we use for the model. If we use questions whose responses should not change between the two data sets, then we are assessing the possibility of the wrong person having been re-contacted, data falsification, or shoddy interviewer work. If we use questions that may conceivably be different, such as attitudinal questions, we are assessing the stability of respondents' opinions or preferences.[9]

3.3.1 Quantities of Interest Specific to Enumerator-Implemented Surveys

When we adjust the QualMix model to incorporate information on enumerators, we can refine existing quantities of interest and define new ones:

Posterior Probability of Match:

$$\begin{aligned}
\lambda_{e_i} &= \text{Logit}^{-1}(\beta_0 + \beta_{e_i}) \\
\xi_i &= \frac{\lambda_{e_i} \prod_{l=1}^L \pi_{1l}^{\nu_{il}}}{\sum_{q=0}^1 \lambda_{e_i}^q (1-\lambda_{e_i})^q \prod_{l=1}^L \pi_{ql}^{\nu_{il}}}
\end{aligned}$$

[8] This method allows us to express our uncertainty that two sets of responses match one another; it cannot tell us which response vector is more correct.

[9] See Appendix B.3 for an expanded discussion of how different questions types can change interpretation of the quality estimates derived from the model.

Enumerator Data Quality:

$$Q_e = \frac{\sum_{i_e}^{N_e} \xi_{i_e}}{N_e} \tag{1}$$

The interpretation of a posterior probability of a match remains the same. Similar to above, we use the average of the posterior probabilities of an enumerator's observations for enumerator quality instead of λ_{e_i} (an enumerator's overall probability of having an observation be in the high quality distribution) because it uses the actual observed agreement summary vector for respondent i. Some of the previously mentioned efforts to identify "cheating" enumerators are not useful for assessing whether individual observations are falsified because they rely on enumerator level characteristics. This approach allows us to assess the probability that each observation chosen for the re-interview process has been falsified or was originally recorded incorrectly.

As above, we can estimate these quantities using a variety of statistical approaches.[10] Because the posterior probability of a match varies between 0 and 1, this measure is also bounded by 0 and 1. This measure let us assess the quality of the data associated with each enumerator and compare this between enumerators involved with the same survey. Importantly, while enumerator data quality may be correlated with enumerator quality, these estimates are not directly a measure of enumerator quality. They just allow us to assess the quality of the data associated with an enumerator and the potential chance for measurement error; whether any lack of data quality is an enumerator's *responsibility* is another question that would require further inquiry.

4 Application: Backchecks

Backchecks can be useful to identify data quality issues, including data falsification (Crespi 1945; Schreiner et al. 1988; Forsman and Schreiner 1991; Murphy et al. 2016). Helpfully, backchecks already produce data like those in Table 1. However, there has not been a

[10] See Appendix D for a discussion of the assumptions necessary for these estimates to be valid.

clear way to *analyze* backcheck data, nor how to use them to express confidence in a survey or the interviewers in a systematic, widely applicable way.

In this section, I apply the framework presented above to re-interview data to derive measures of survey and enumerator quality. Parameter estimates from the associated statistical model can be used to assess survey and enumerator data quality. Survey companies and independent researchers can use this method to quickly get a sense of how well the survey has been implemented. In addition, researchers can use the quality estimates derived from these data in analysis with the overall survey. I use simulated data to demonstrate the use and effectiveness of the proposed approach. I then apply the model to real data as an empirical demonstration.

4.1 Simulation Study

I conduct a simulation study to assess QualMix's sensitivity to real world conditions, varying the percent of each enumerator's respondents backchecked and average match proportion. Under different conditions defined by these parameters, I check 1) if the model assesses overall survey quality accurately, and 2) whether the model does a "good" job identifying enumerators associated with lower quality data. Survey administrators are often wary of doing more backchecks due to their costs. Varying the backcheck rate assesses how the model performs with varying number of observations per enumerator — can survey administrators cut costs yet still be confident in how well the model assesses quality? Also, not all survey processes will go equally smoothly. Varying the average match proportion (representing the proportion of high-quality data) allows me to assess how data quality impacts performance.

4.1.1 Set-Up

For the simulation tests, I varied the following parameters:

- Percent of Respondents Backchecked (%B) $\in \{0.05, 0.10, 0.15, 0.2\}$

- Average Match Proportion ($\text{logit}^{-1}(\beta_0)) \in \{0.8, 0.9, 0.95\}$[11]

This results in twelve different parameter combinations. Simulation proceeds by deciding on an "overall" survey quality and then on how enumerators are better or worse than this overall quality. Subsequently, I generate original data–backcheck data pairs. Note that even for "matching" observation pairs, there was some probability that some of the variable values were incorrect in the backcheck data. Please see Appendix E for a full description of the simulation process and the simulation parameters not varied during the study.

To increase the external validity of the simulation exercise, I creating artificial dissimilarities in existing survey data. The survey used for the basis of the simulation was carried out from October to January 2019 in 128 Malawian markets. Table 2 shows the variables included in the simulation and their type.

Variable	**Type (For Difference Vector)**
whether respondent is female or not	binary
respondent's age	numeric, 18-86
level of education attained by the respondent	ordered, seventeen levels
the respondent's household income	numeric, 0-500, in tens of thousands of Malawian kwacha
how frequently the respondent sells in the market	ordered, eight levels
the respondent's stall's primary activity	categorical, fifty-two levels
how respondent's profits this year compare to profits last year	ordered, five levels
how many vendors out of ten always paid the market tax according to the respondent	numeric, 0-10
a numeric variable that is a function of several of the other variables on this list	numeric, created during each iteration. See Appendix E.2 for more information

Table 2: Variables Used in Simulation.

[11] The β_e's do not vary within or between parameter sets. What varies is the observations chosen as matches and non-matches, and the number of errors in match variables.

These are all variables that could be used for backchecking. Age, education, and sell frequency were three of the variables used during the actual backchecking done for this survey. Variables like name and phone number would normally be used for backchecking, but for privacy reasons I omitted these variables in the simulations. These variables all represent values that should not change between the original survey and the backcheck. As such, in this simulation, we are assessing data reliability.

I fit the model in a Bayesian framework using `Stan` (Stan Development Team 2020). See Appendix E.1 for more information on the model fitting, including the priors used.

4.1.2 Assessing the Model's Ability to Assess Survey Quality

In this section, I show results for how well the model assesses survey and enumerator data quality.[12,13] Because of the questions chosen for this simulation, we are assessing data collection accuracy. In order to calculate the error, I use the true proportion of matches for the survey and for each enumerator, respectively, as true values of Q_S and Q_e. I then use the estimators for these quantities proposed in Section 3.3.1 and calculate the error. Figure 2 shows the error in survey quality under different parameter combinations. We can see the errors are around 0 for all parameter combinations, with the mean error decreasing slightly as the average match proportion goes up. Keeping the overall match probability constant, the mean error is perhaps somewhat lower when only 5% of observations are backchecked, although the credible intervals are much larger. After that, the error does not change much as the backcheck percentage increases, with credible intervals becoming smaller due to the larger number of observations exposed to the model.

Table 3 shows summary statistics for the bias in enumerator quality across all thirty-

[12] An important first step is to assess whether the model was able to identify two clearly separable distributions. The median of the posterior of the Jensen-Shannon Distance for all possible parameter combinations is between .57 and .63 — it grows as the average match proportion grows, which makes sense due to how the data are simulated. The two distributions become farther apart the fewer errors there actually are. See Appendix C.1 for a figure of all JSD estimates.

[13] Here, I only show how well the model allows us to estimate survey and enumerator data quality. In Appendix F, I show that the model also performs well when identifying low-quality observations.

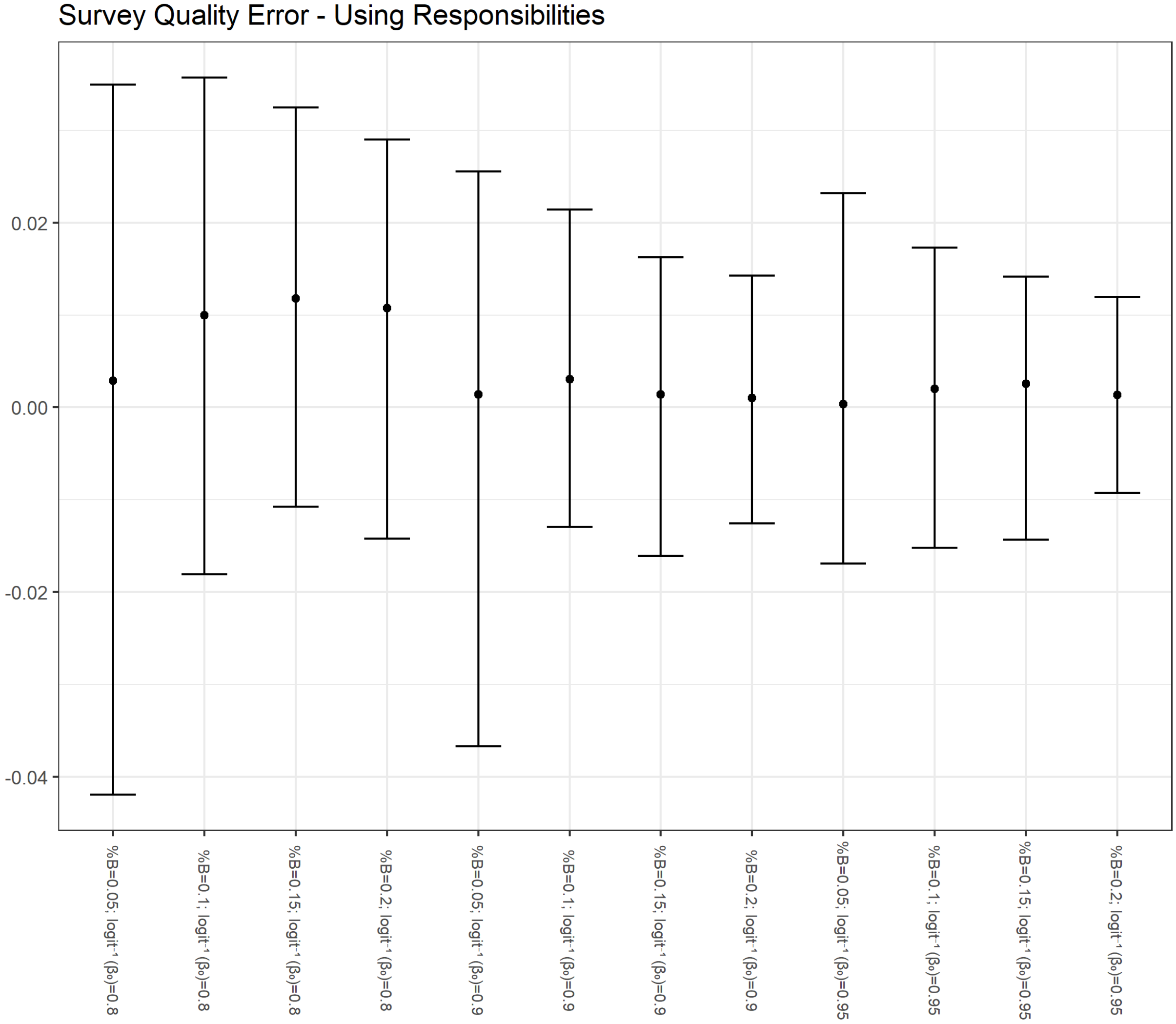


Figure 2: Mean of the posterior of the error (Empirical Bias) of overall survey quality for different parameter combinations with 95% credible intervals.

five enumerators in the simulation. The table demonstrates that enumerator quality bias is mostly clustered tightly around 0. A few enumerators display somewhat large negative bias - the largest absolute bias in any parameter combination is 0.0842. Figure 9 in Appendix G shows that this bias corresponds to one of the three enumerators with the consistently highest negative bias. These enumerators are also the three "worst" enumerators with respect to β_E (in other words, they deviate the most negatively from the average match proportion). There are two reasons why their quality estimates may be negatively biased. First, because the probability of a match is generally low for these enumerators, the probability of a match making it into the backchecks is lower. Second, these enumerators will make more errors in match observations because of their lack of quality (see Appendix E for why this is the case) — the model struggles somewhat to pick this up because they already are very "bad." In other words, the match and non-match observations of these enumerators may be similar. Thus, the model assesses these enumerators as worse than their actual match proportion.

Simulation Parameters	Mean	SD	Min	Max
%B=0.05; $\text{logit}^{-1}(\beta_0)$ =0.8	0.0020	0.0198	-0.0616	0.0381
%B=0.1; $\text{logit}^{-1}(\beta_0)$ =0.8	0.0104	0.0231	-0.0568	0.0410
%B=0.15; $\text{logit}^{-1}(\beta_0)$ =0.8	0.0112	0.0219	-0.0564	0.0428
%B=0.2; $\text{logit}^{-1}(\beta_0)$ =0.8	0.0107	0.0227	-0.0601	0.0323
%B=0.05; $\text{logit}^{-1}(\beta_0)$ =0.9	0.0003	0.0215	-0.0822	0.0286
%B=0.1; $\text{logit}^{-1}(\beta_0)$ =0.9	0.0032	0.0162	-0.0680	0.0299
%B=0.15; $\text{logit}^{-1}(\beta_0)$ =0.9	0.0006	0.0206	-0.0842	0.0198
%B=0.2; $\text{logit}^{-1}(\beta_0)$ =0.9	0.0017	0.0210	-0.0791	0.0207
%B=0.05; $\text{logit}^{-1}(\beta_0)$ =0.95	0.0017	0.0150	-0.0511	0.0326
%B=0.1; $\text{logit}^{-1}(\beta_0)$ =0.95	0.0022	0.0122	-0.0576	0.0160
%B=0.15; $\text{logit}^{-1}(\beta_0)$ =0.95	0.0019	0.0102	-0.0473	0.0145
%B=0.2; $\text{logit}^{-1}(\beta_0)$ =0.95	0.0020	0.0128	-0.0619	0.0169

Table 3: Enumerator Quality Bias Summarized by Simulation Parameters. Calculated across enumerator-specific biases.

Furthermore, Table 3 shows that the mean of the bias across enumerators is lowest when the average match proportion is 0.9, but highest when the average match proportion is 0.8. The mean of the bias when the average match proportion is 0.95 is similar overall to the mean when it is 0.9. The standard deviation, on the other hand, decreases monotonically with the average match proportion. This makes sense: as overall quality increases, it becomes easier to identify poorly performing enumerators. There are fluctuations within levels of average match proportion, but generally speaking the percent backchecked does not seem to drastically impact the mean bias across enumerators. This mirrors the trend seen in the overall quality estimates and supports the conclusion that, in order to assess enumerator quality, survey firms would not need to increase the percent backchecked.

4.2 Real World Application

I next apply the QualMix model to a real world case: the survey used as the starting point for the simulation, carried out in Malawi between October and January 2019. This time, I use the actual backcheck data. Of the 12,370 respondents, 657 (5.3% of the sample) were re-contacted by telephone in November and December 2018. Backchecks were not stratified by enumerator. Fifty-one enumerators were used for the study, but only forty-five had a respondent recontacted (the other six interviewed very few respondents; one of the forty-five interviewed only one respondent, who was then randomly chosen for a backcheck). Figure 3 shows the proportion of each enumerator's respondents who were randomly chosen for the backcheck process. The one enumerator with a 100% backcheck rate is omitted from the figure to make it easier to interpret.

Six variables were chosen for all backchecks:[14]

1. respondent's age
2. respondent's education
3. how often respondent sells at the market

[14] About 20% of respondents were asked a longer version of the original survey; these respondents were also asked more questions during the actual backcheck process.

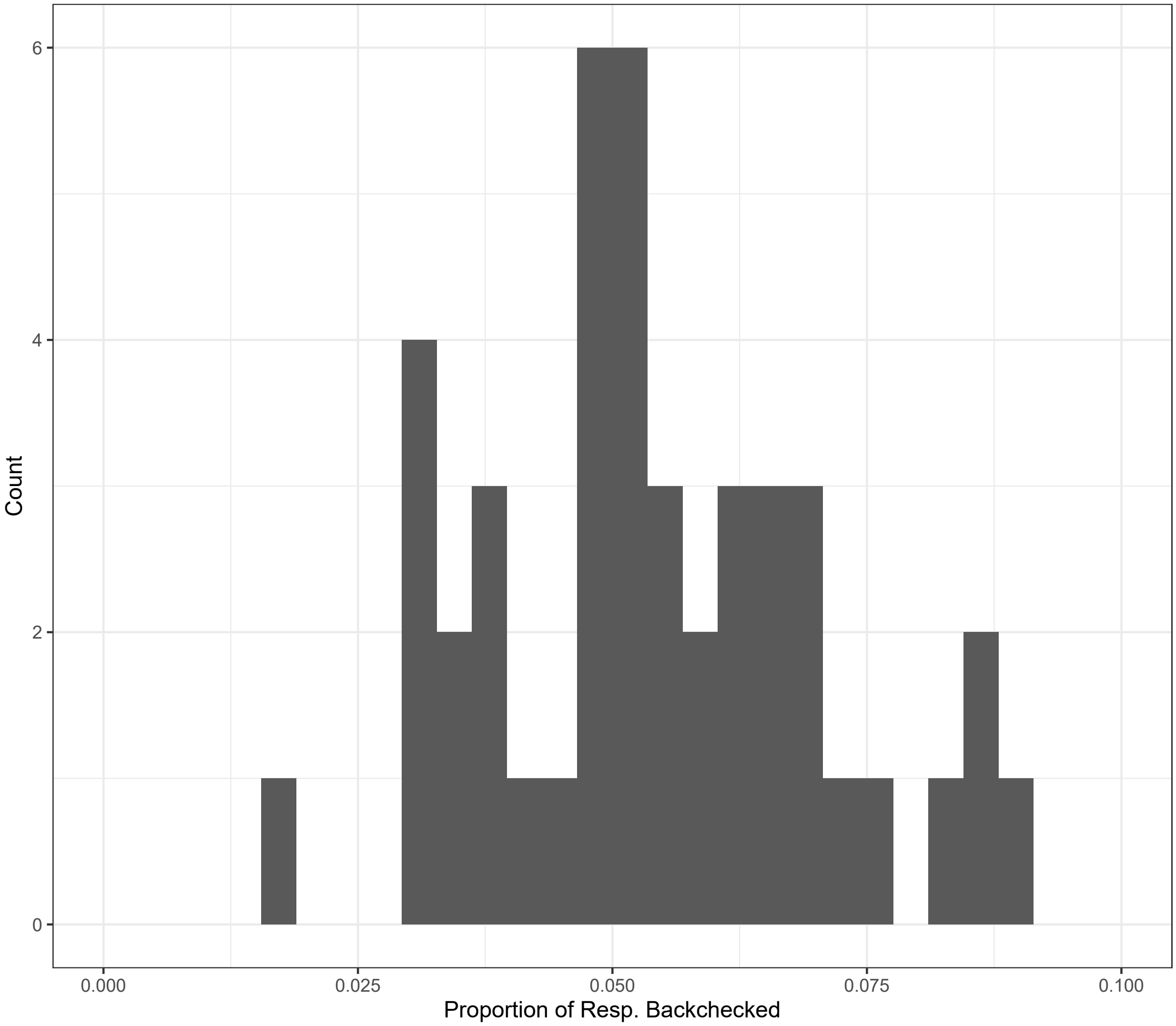


Figure 3: Histogram of Proportion of an Enumerator's Respondents Chosen for the Backcheck.

4. what the respondent sells/offers
5. whether the respondent showed the enumerator a receipt
6. respondent's satisfaction with developments in the market.

These should not have changed between original enumeration and the backcheck, besides perhaps satisfaction with developments. As such, we can use these questions to assess the reliability of the survey. I create $\boldsymbol{\nu}$ for each backcheck pair. I categorize NA values as disagreements. I use the model described in Appendix E.1 without alterations to derive estimates of enumerator and survey quality. I fit the model using `Stan` using the same priors as in the simulation.[15]

4.2.1 Results

Figure 4 shows the two estimated multinomials, demonstrating that the model was able to identify distinct distributions over agreement categories.[16]

It is important to note that the low-quality distribution puts almost all of the probability into the "Complete Disagreement" category. Of the 657 backcheck observations, 85 had no correct values - these were all cases where a different person answered the phone than the one interviewed for the survey (most often the individual who answered the phone did not know the person originally interviewed) or where no one answered the phone. The survey company considered these as failed backchecks, but it is important to take these cases into account - after all, it is possible that the original observations were fabricated.[17] Because these represent 12% of backchecked respondents, it was straightforward for the model to identify all of these observations as belonging to the same cluster. The match distribution, however, clearly still contains some "Complete Disagreement" values. The 95% credible intervals for the expected value of the categories are shown in Table 4.

[15] I use `cmdstanr` to fit the models in this section (Stan Developers and their Assignees 2021).

[16] The 95% credible interval for the Jensen-Shannon Distance for these two distributions is [0.714, 0.760].

[17] If this were a random process, then we would expect the distribution of such backcheck failures to be uniform among enumerators. Figure 5 clearly shows that this is not the case. See Appendix H for an analysis of dropping these failed backchecks. Once these observations are dropped, it becomes more difficult to detect two distinct distributions.

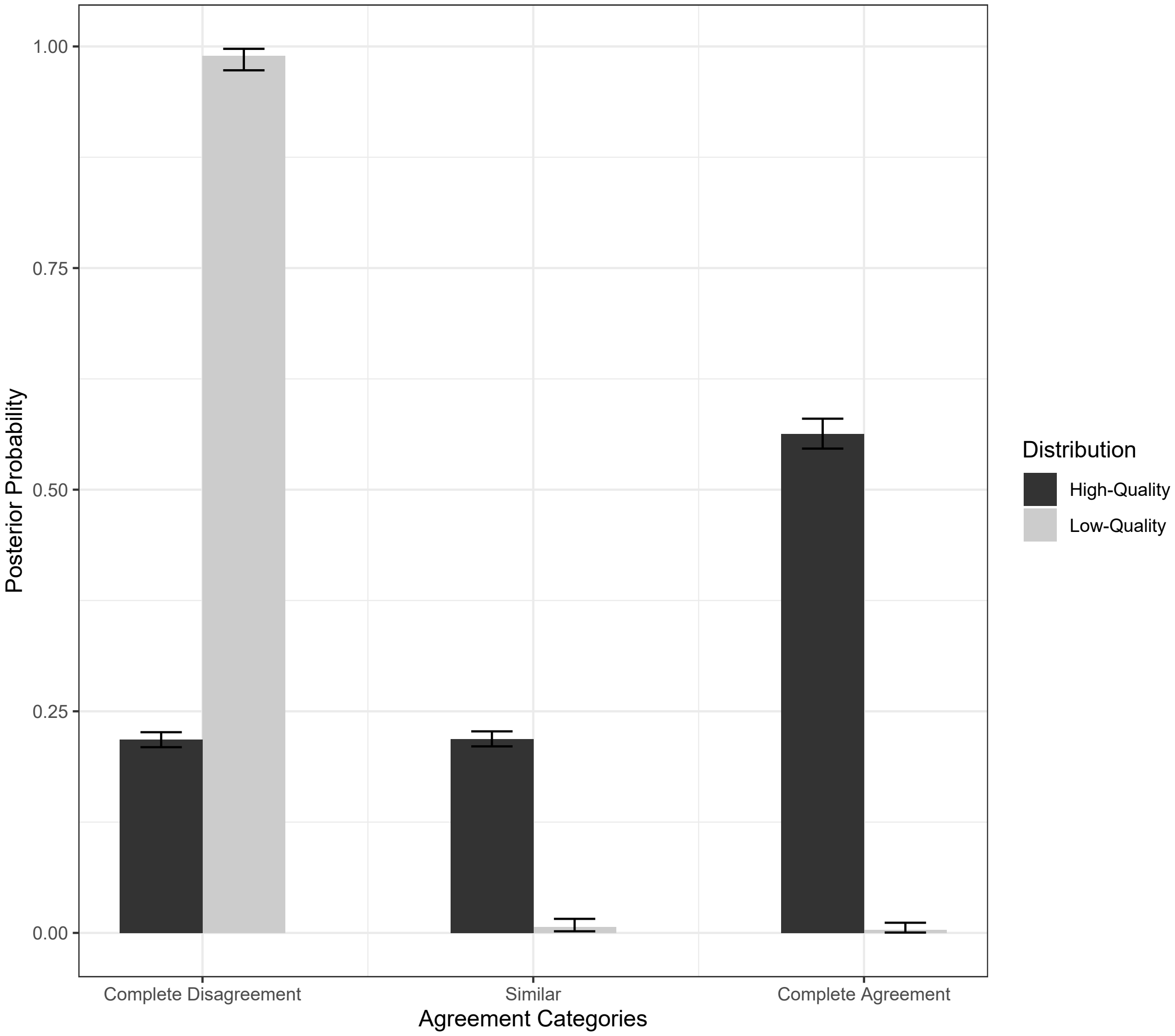


Figure 4: Median of posterior distributions of $\hat{\boldsymbol{\pi}}_0$ (non-match) and $\hat{\boldsymbol{\pi}}_1$ (match) along with 95% credible intervals.

Category	2.5%	Median	97.5%
Complete Disagreement	1.257	1.308	1.358
Similar	1.264	1.313	1.364
Agreement	3.278	3.378	3.480

Table 4: 95% Credible Intervals for Expected Values of the Match Distribution

The variable with the largest number of inconsistencies between the backcheck and the original data was the one which asked if respondents had shown the enumerator a receipt for paying the daily market tax. There are several possible explanations for this fact. First, respondents could be suffering from social desirability bias to not say no; more respondents in the backcheck said that they showed a receipt than in the original data. Second, it is possible that vendors *did* show a receipt, but that enumerators reported that they did not, to make the survey go quicker — if a respondent showed the enumerator a receipt, the enumerator was directed to take a photo of it, which could have taken time. The backcheck itself unfortunately does not offer evidence one way or another, which demonstrates that survey implementers still need to assess quality actively at the time of enumeration as well.

An added value of this analysis is that it identifies what our model considers "high quality." Survey administrators and researchers must decide whether they are satisfied with the high-quality and low-quality distributions.[18] Even if they are *not* satisfied with a high-quality distribution, however, the model and approached defined here still have utility, as they identify common patterns in the data.[19] If a high-quality distribution is unsatisfactory, that is a sign in and of itself that something might have gone wrong during data collection.

We can derive enumerator data quality estimates using Eq. 1, shown in Figure 5. There is considerable variation in enumerator data quality, with some estimates of data quality being low: eight have data quality estimates of .75 or lower, with two below .5. Importantly, while these enumerator data quality estimates may be *correlated* with enumerator quality – that is, enumerator expertise and ability – they should not be directly interpreted as enumerator quality. An enumerator handed a malfunctioning tablet that misrecords data would be associated with poor data quality, but this has nothing to do with the enumerator's ability to do their job. Additionally, because of how the survey company performed the survey[20]

[18] The model does allow users to specify desired quality distributions.

[19] Note that while it will be possible and perhaps beneficial to compare high-quality and low-quality distributions between surveys, quality estimates will not necessarily be directly comparable unless the quality distributions are the same.

[20] The implementing organization sent enumerator teams to specific parts of the country. Thus, enumerator data quality may be confounded by regional data collection issues.

and the backcheck, we cannot say that the enumerators were responsible for flawed data, but we can say that some are associated with many more backcheck failures: there are clear within-enumerator data patterns that the model helps us see.

How can we be sure, however, that these enumerator data quality estimates actually represent something akin to real world quality and not just variations in backcheck performance? As a validation exercise, I examine the relationship between the receipt variable and enumerator data quality. As mentioned above, the receipt variable is one where quality can impact data collection quality starkly. For this exercise, I use enumerator data quality as a proxy for enumerator quality. More experienced enumerators would be more likely to get someone to show them a tax receipt because they may be better at getting a respondent's trust and less likely to rush through a survey. Using a simple binomial logit regression model using all 12,370 observations with enumerator data quality is a sole predictor, I find that quality is associated with the probability that a respondent showed an enumerator a receipt. Increasing enumerator data quality from 0.5 to .75 increases the probability that an enumerator reported being shown a receipt by .107; increasing enumerator data quality from .75 to 1 increases it by a further .140.[21] As Figure 5 shows, this level of variation in enumerator data quality is observed in the data, underscoring the vast differences in how well enumerators were able to solicit receipts.

As such, these enumerator data quality estimates help identify enumerators who may have produced problematic data. Survey administrators and researchers can then investigate potential causes of these issues, and can even see if certain enumerator characteristics (more experienced versus less experienced, for example) correlate with these quality estimates. Survey administrators can also see if estimated enumerator quality is associated with certain survey measures. In addition, they can act on this information, by down-weighting observations or enumerators about whose quality they are uncertain. It is also possible to run this model in real time, consistently updating it with data from the field. Researchers

[21] Median of the posterior predictive distribution. See Appendix I for more information on the data, model, and fitting process used for this analysis.

do not need to wait until the end of enumeration to apply this model to their data. In such ways, the model more systematically facilitates the identification, investigation, and solving of data quality issues than deterministic backcheck comparisons.

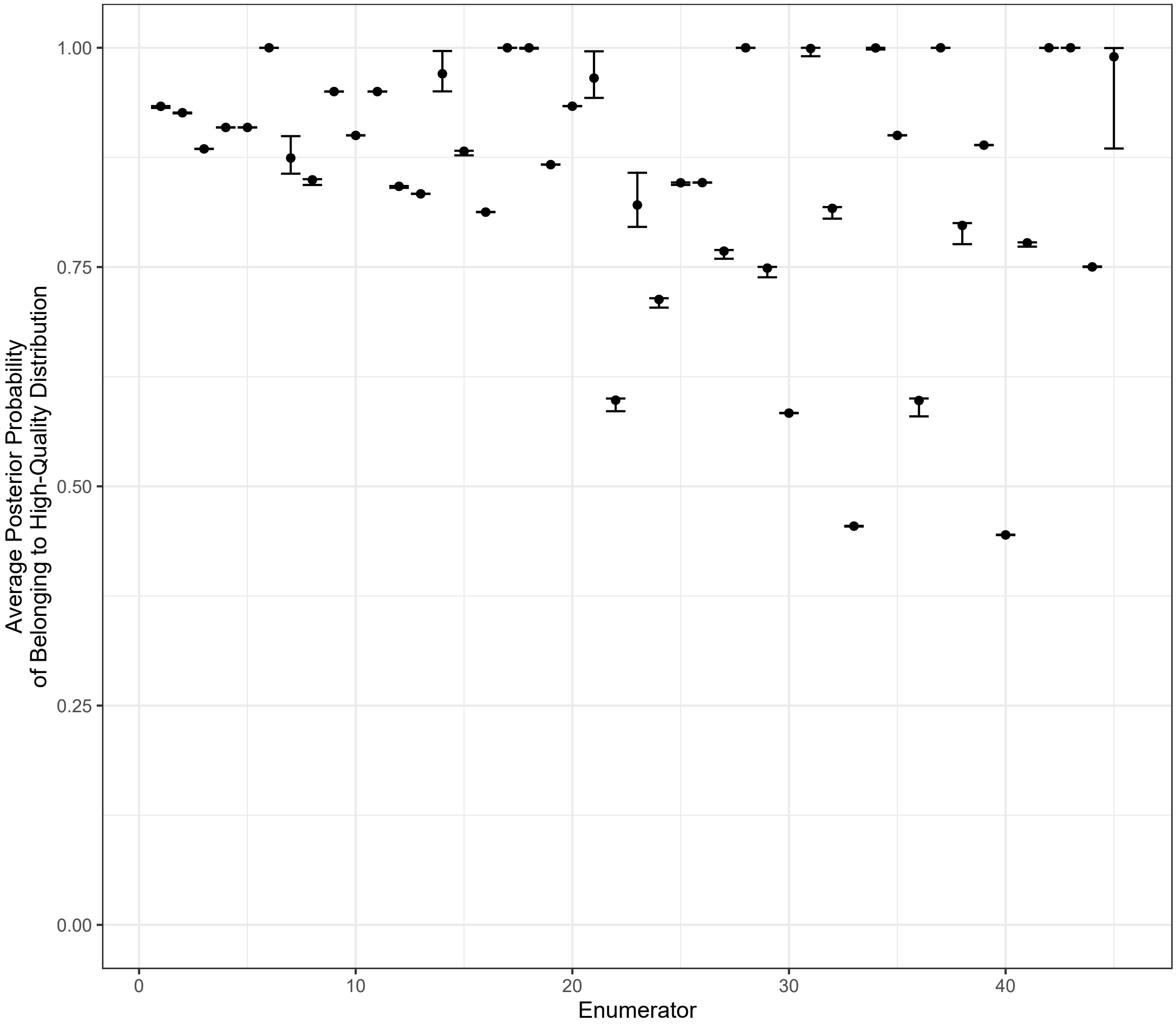


Figure 5: Median of the Posterior of the Average Posterior Probability of a Match for all 45 Enumerators. Error bars show 95% credible intervals.

5 Conclusion

In this paper, I describe the QualMix model to help assess data quality when two sets of responses exist for the same individual to the same questions. I suggest a mixture model approach that uses the number of inconsistencies and agreements between the two sets of records as data. The model allows us state how confident we can be that there is limited measurement error. With this model, survey administrators can estimate the overall data quality of a survey, as well as the data quality associated with enumerators implementing the survey.

The simulations demonstrate that the model effectively identifies problematic observations and assesses survey and enumerator data quality. It also shows that survey implementers can get sufficient estimates of quality by backchecking only 5% of respondents. The empirical application demonstrates how to apply the model to real world data. It also shows what can happen when the model struggles to separate out the two multinomials, or when there is greater uncertainty about one of the distributions. In order to make this situation less likely, survey implementers should use more than six variables for backchecking. Backcheckers should ask all backcheck questions, regardless of whether a name matches — after all, it is possible that a *name* is incorrect, but that the other values are correct. Finally, in order to not confound variation in region of enumeration and enumerator identity, enumerators be sent to all regions of enumeration.

While the approach I present here should not replace existing quality control measures (Cohen and Warner 2021), it can be incorporated into existing quality control suites. The model is flexible. It can be adapted to allow a more fine-grained analysis to assess different kinds of survey quality. It can also easily incorporate other information, such as on enumerators. Researchers are not limited to only evaluating survey backchecks with this model; other applicable scenarios include estimating the uncertainty that the correct respondents have been recontacted in a panel survey, for example.

Finally, the model could be used more expansively than just estimating survey quality. In particular, it offers avenues for dealing with measurement quality issues once they have been discovered. Generally, when researchers think that the quality of their data is poor, they have one of three options: 1) dropping data, 2) ignoring data concerns, and 3) directly modeling measurement error. The problem with the first solution is that it can be very costly to drop data. Dropping data also has analytic implications—fewer observations generally means lower statistical power and greater uncertainty about parameter estimates. Researchers could try to get new, better quality data, but this is once again expensive. Some researchers choose to ignore seemingly non-serious data quality concerns, for exactly these reasons. Yet, this has unknown implications for any subsequent analyses. Finally, while directly modeling measurement error is a sound strategy, researchers first have to develop a model, without knowing truth. An extension of the QualMix model would be using the estimated posterior probability of a match as a weight in subsequent analysis. We know that measurement error can induce bias in regression analysis. The aim of this process is to upweight observations about whose quality we are more certain, and to downweight those about whose quality we are less certain, reducing bias in parameter estimates. This can help researchers save money by not having to seek out more observations, if they decide, using this model, that they cannot fully trust some of the data they have collected.

References

Ahmed, B., Ahmad, A., Herekar, A. A., Uqaili, U. L., Effendi, J., Alvi, S. Z., Herekar, A. D., and Steiner, T. J. (2014), "Fraud in a population-based study of headache: prevention, detection and correction," *The Journal of Headache and Pain*, 15, 1–5.

Alwin, D. F. (2007), *Margins of Error: A Study of Reliability in Survey Measurement*, Hoboken, NJ: John Wiley & Sons, Inc.

— (2011), "Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach," in *Question Evaluation Methods: Contributing to the Science of Data Quality*, eds. Madans, J., Miller, K., Maitland, A., and Willis, G., Hoboken, NJ: John Wiley & Sons, Inc., pp. 265–293.

— (2016), "Survey Data Quality and Measurement Precision," in *The SAGE Handbook of Survey Methodology*, eds. Wolf, C., Joye, D., Smith, T. W., and chih Fu, Y., Thousand Oaks, CA: SAGE, pp. 527–557.

Birnbaum, B., Borriello, G., Flaxman, A. D., DeRenzi, B., and Karlin, A. R. (2013), "Using Behavioral Data to Identify Interviewer Fabrication in Surveys," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pp. 2911–2920.

Blasius, J. and Thiessen, V. (2012), *Assessing the Quality of Survey Data*, Thousand Oaks, CA: SAGE Publications.

Bohrnstedt, G. W. (2010), "Measurement Models for Survey Research," in *Handbook for Survey Research*, eds. Marsden, P. V. and Wright, J. D., Bingley, UK: Emerald Group Publishing, Ltd., pp. 347–404.

Bredl, S., Storfinger, N., and Menold, N. (2013), "A Literature Review of Methods to Detect Fabricated Survey Data," in *Interviewers' Deviations in Surveys - Impact, Reasons, De-*

tection and Prevention, eds. Winker, P., Menold, N., and Porst, R., Frankfurt am Main: Peter Lang, pp. 3–24.

Castorena, O., Cohen, M. J., Lupu, N., and Zechmeister, E. J. (2021), "How Worried Should We Be? The Implications of Fabricated Survey Data for Political Science," Working Paper. Version: July 26, 2021. `https://www.noamlupu.com/fabrication.pdf`.

Cohen, M. J. and Warner, Z. (2021), "How to Get Better Survey Data More Efficiently," *Political Analysis*, 29, 121–138.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003), "A Comparison of String Distance Metrics for Name-Matching Tasks," in *Proceedings of the Workshop on Information Integration on the Web*, International Joint Conference on Artificial Intelligence (IJCAI), pp. 73–78.

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

De Haas, S. and Winker, P. (2014), "Identification of partial falsifications in survey data," *Statistical Journal of the IAOS*, 30, 271–281.

— (2016), "Detecting Fraudulent Interviewers by Improved Clustering Methods – The Case of Falsifications of Answers to Parts of a Questionnaire," *Journal of Official Statistics*, 32, 643–660.

DeDeo, S., Hawkins, R. X. D., Klingenstein, S., and Hitchcock, T. (2013), "Bootstrap Methods for the Empirical Study of Decision-Making and Information Flows in Social Systems," *entropy*, 15, 2246–2276.

DIME, W. B. (n.d.), "Back Checks," DIME Wiki. https://dimewiki.worldbank.org/wiki/Back_Checks, accessed: 2020-10-10.

Drost, H.-G. (2018), "Philentropy: Information Theory and Distance Quantification with R," *Journal of Open Source Software*, 3.

Enamorado, T., Fifield, B., and Imai, K. (2018), "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records," *American Political Science Review*, 1–19.

Endres, D. M. and Schindelin, J. E. (2003), "A New Metric for Probability Distributions," *IEEE Transactions on Information Theory*, 49, 1858–1860.

Fellegi, I. P. and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

Finn, A. and Ranchhod, V. (2017), "Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South African Survey," *The World Bank Economic Review*, 31, 129–157.

Forsman, G. and Schreiner, I. (1991), "The Design and Analysis of Reinterview: An Overview," in *Measurement Errors in Surveys*, eds. Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S., Chichester: Wiley, pp. 279–301.

Gibson, M. (n.d.), "Data quality checks," Abdul Latif Jameel Poverty Action Lab (J-PAL). `https://www.povertyactionlab.org/resource/data-quality-checks#:~:text=The%20Abdul%20Latif%20Jameel%20Poverty%20Action%20Lab%20(J%2DPAL),is%20informed%20by%20scientific%20evidence.&text=They%20set%20their%20own%20research,%2C%20policy%20outreach%2C%20and%20training.`

Imai, K. and Tingley, D. (2012), "A Statistical Method for Empirical Testing of Competing Theories," *American Journal of Political Science*, 56, 218–236.

IPA (2018), "IPA's Research Protocols," https://www.poverty-action.org/researchers/research-resources/research-protocols, accessed: 2020-10-10.

Jaro, M. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Krejsa, E. A., Davis, M. C., and Hill, J. M. (1999), "Evaluation of the Quality Assurance Falsification Interview used in teh Census 2000 Dress Rehearsal," in *Proceedings of the Survey Research Method Section*, American Statistical Association, pp. 635–640.

Kuriakose, N. and 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., and Singer, P. (2011), "Using Statistical Models for Sample Design of a Reinterview Program," *Journal of Official Statistics*, 27, 433–450.

Lin, J. (1991), "Divergence Measures Based on the Shannon Entropy," *IEEE Transactions on Information Theory*, 37, 145–151.

Madans, J., Miller, K., Maitland, A., and Willis, G. (2011), *Question Evaluation Methods: Contributing to the Science of Data Quality*, Hoboken, NJ: John Wiley & Sons.

McLaughlan, G. and Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley & Sons.

Murphy, J., Biemer, P., Stringer, C., Thissen, R., Day, O., and Hsieh, Y. P. (2016), "Interviewer falsification: Current and best practices for prevention, detection, and mitigation," *Statistical Journal of the IAOS*, 32, 313–326.

Nielsen, F. (2011), "A family of statistical symmetric divergences based on Jensen's inequality," eprint arXiv:1009.4004v2 [cs.CV].

of Survey Quality, T. S. H. (2016), "Another Look at Survey Data Quality," in *The SAGE Handbook of Survey Methodology*, eds. Wolf, C., Joye, D., Smith, T. W., and chih Fu, Y., Thousand Oaks, CA: SAGE, pp. 613–629.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rosmansyah, Y., Santoso, I., Hardi, A. B., Putri, A., and Sutikno, S. (2019), “Detection of Interviewer Falsification in Statistics Indonesia’s Mobile Survey,” *International Journal on Electrical Engineering and Informatics*, 11, 474–484.

Sarracino, F. and Mikucka, M. (2017), “Bias and efficiency loss in regression estimates due to duplicated observations: a Monte Carlo Simulation,” *Survey Research Methods*, 11, 17–44.

Schnell, R. (1991), “Der Einfluß gefälschter Interviews auf Survey-Ergebnisse,” *Zeitschrift für Soziologie*, 20, 25–35.

Schräpler, J.-P. and Wagner, G. G. (2005), “Characteristics and impact of faked interviews in surveys - An analysis of genuine fakes in the raw data of SOEP,” *Allgemeines Statistisches Archiv*, 89, 7–20.

Schreiner, I., Pennie, K., and Newbrough, J. (1988), “Interviewer falsification in census bureau surveys,” in *Proceedings of the Survey Research Method Section*, American Statistical Association, pp. 491–496.

Stan Developers and their Assignees (2021), *CmdStanR*, r Package Version 0.4.0 (Not published on CRAN).

Stan Development Team (2020), *Stan Modeling Language Users Guide and Reference Manual*, 2.27.

StataCorp (2019), *Stata Statistical Software: Release 16*, StataCorp LLC, College Station, TX.

Team, S. D. (2020), *RStan: the R interface to Stan*, r package version 2.21.2.

Tourangeau, R. (2021), “Survey Reliability: Models, Methods, and Findings,” *Journal of Survey Statistics and Methodology*, 9, 961–991.

Tourangeau, R., Sun, H., and Yan, T. (2021), “Comparing Methods for Assessing Reliability,” *Journal of Survey Statistics and Methodology*, 9, 651–673.

White, M. (2016), *BCSTATS: Stata module to analyze back check (field audit) data and compare it to the original survey*, Statistical Software Components S458173, Boston College Department of Economics.

Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Appendix

A Forming Agreement Summary Vectors

In this appendix, I describe decisions rules for different types of variables. These explain how Table 1 was filled. These decision rules are also the ones used in the simulation and in the real-world application presented in the paper.

For the string variable (*Last Name*), I use the Jaro-Winkler string comparator (Jaro 1989; Winkler 1990; Cohen et al. 2003). The Jaro-Winkler string comparator is a metric that turns the similarity between two strings into a number between 1 (most similar) to 0 (most different). Winkler (1990) suggests cutoffs of .94 for "complete agreement" and .88 for "similar." The Jaro-Winkler values for the Melzer:Beier and Karlsen:Karls comparisons are .7 and .943, respectively. Using the cutoffs suggested by Winkler, we can therefore say that Melzer and Beier are in "complete disagreement" and Karls and Karlsen are in "complete agreement."

For the ordered categorical variable (*Monthly Income*), I use the *percent of max range* measure. More specifically, I use the numeric ordering behind the categories in the following equation: for two numbers a and b, Percent of Max Range $= 1 - \frac{|a-b|}{\max\{\max(\boldsymbol{V}_a)-\min(\boldsymbol{V}_a), \max(\boldsymbol{V}_b)-\min(\boldsymbol{V}_b)\}}$, where $\boldsymbol{V}_a$ and $\boldsymbol{V}_b$ represent the vectors of observed values from which a and b were drawn. This measure will also be between 0 (most different) and 1 (most similar). The logic behind this measure is that small differences when the range is large are more likely to be random than similarly sized differences when the range is small. I use cutoffs of .94 and .88, for continuity with the Jaro-Winkler approach for strings. In the Table 1 example, we can imagine that there are six categories (<$250, [$250 - 500),...,>$1,500). Then the percent of max range values for the [$250, $500): < $250 comparison is $1 - \frac{|2-1|}{5} = 0.8$ and for the <$250:<$250 comparison is $1 - \frac{|1-1|}{5} = 1$. This suggests complete disagreement for the first comparison and complete agreement for the second comparison.

Comparing categorical values is in some ways more straightforward. As there is no natural ordering, different values represent disagreements. Nevertheless, depending on the application, certain categories could be more similar than others. For example, in Table 1, $\boldsymbol{r_{a2}}$ and $\boldsymbol{r_{ab}}$ have "Market Vendor" and "Business Owner" as recorded responses for *Occupation*. Market vendors may see themselves as business owners, and so two different responses *of this type* could come from the same individual. Therefore, a researcher applying this method could group similar levels of a categorical variable together, if possible, and consider levels within such groupings as similar. In the example here, I demonstrate such a strategy; this results in agreement vector entries for *Occupation* of "complete disagreement" for the Market Vendor:Tax Collector comparison and "similar" for the Market Vendor:Business Owner comparison.

For the continuous variable *Age*, I once again use the percent of max range measure. Suppose the observed maximum for the age variable is 18, and the observed minimum is 18. Then, the comparison value for the 65:57 and 21:31 comparisons are .882 and .853, respectively. Using the same cutoffs as before, this results in the *Age* entries for the two agreement vectors to be "similar" and "complete disagreement."

We can then add up how many of each of the three agreement-levels there are in each agreement vector to form the agreement summary vector.

B Possible Extensions to QualMix Model

B.1 Incorporating Respondent-Level Characteristics or Survey Metadata

Quality probability λ does not have to be the same for each observation. In fact, it is possible to regress the latent cluster membership Q_i (high-quality vs. low-quality) on additional data (Imai and Tingley 2012). In the case of backchecks, we can incorporate information on enumerators into the model, for example. It may also make sense to incorporate metadata into the model in this way, as additional information such as, for example, differences in completion time or survey location may help differentiate between matching observation sets. For example, if survey implementers are concerned that data quality may be different in different regions — data collection may be more difficult in some places than others — the model can have region specific λ's.

For example, in the case of backchecks, we will have two sets of responses to the same K questions for a subset of the sample: the first set is the originally collected data; the second set corresponds to the information collected during the re-interviews. The goal of applying the model will be to see how well these responses match.

In the case of re-interviews, however, we have additional information that we can incorporate: the original data enumerators. In the original model λ characterizes the overall probability that $\boldsymbol{r}_{a_i}$ and $\boldsymbol{r}_{b_i}$ match. It is possible, however, to form a simple logistic regression using the latent Q_i as the outcome. This allows us to see how enumerators affect the probability of the survey-backcheck pair being high quality. We will use a random intercept by enumerator in this regression, which also means that we must now index λ by e, the enumerator. Thus, the extended model becomes

$$\begin{aligned}
\boldsymbol{\nu}_i|Q_i = q &\overset{\text{i.i.d}}{\sim} \text{Multinomial}(\boldsymbol{\pi}_q) \\
Q_i &\overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\lambda_{e_i}) \\
\lambda_e &= \text{logit}^{-1}(\beta_0 + \beta_e) \\
\beta_e &\sim \mathcal{N}(0, \sigma_e)
\end{aligned}$$

$\text{logit}^{-1}(\beta_0)$ in this context represents the overall probability of a match, and the intercepts by enumerator (β_e) represent the deviations from this probability. λ_e represents the probability that $\boldsymbol{r}_{a_i}$ and $\boldsymbol{r}_{b_i}$ — $i \in I_e$ — match, i.e. that observations associated with enumerator e are of high quality. In short, we now have E different λ's. The benefit of this approach is that it allows for match probability to vary by enumerator.

Up to now we have been assuming that there are no data quality concerns from the individuals doing the backchecking. However, this may not be the case. The model can be adapted to have random intercepts by enumerators and backcheckers. This would mean that $\lambda_{eb} = \text{logit}^{-1}(\beta_0 + \beta_e + \beta_b)$. This would allow survey administrators to see the impact of both field enumerators and backchecking enumerats on data quality.

This example should make it clear that survey implementers could include respondent-level variables in the regression on the latent Q_i as. We know that respondent characteristics can interact with enumerator characteristics, such as gender, which could affect data quality. A further extension could include having interviewer intercepts vary by interviewer characteristics.

B.2 Different Agreement Categories

It is possible to generalize QualMix to allow for different agreement categories for each for the K response questions. Each would then have L_k levels. Then, $\gamma_{ik}|Q_i = q \sim \text{Categorical}(\boldsymbol{\pi}_{qk})$, where $\boldsymbol{\pi}_{qk}$ is a vector of the probabilities of the L_k categories for question k. This would be useful if survey implementers were interested in these probabilities for each question – for example if they wanted to see if different questions had different probabilities π_{qkl}, that is, the probability of disagreement category l for question k in the match and non-match distributions. This would be of interest in panel surveys, for example, where some questions, such as age, are *expected* to disagree more between response sets — if there is *no* variation, it would represent a problem. A slightly simpler version would be to separate the variables with different levels into separate agreement vectors, each stemming from independent multinomials. This would lose the ability to say something about individual questions, but would result in fewer parameters. However, the herein described formulation is more parsimonious and is therefore easier to fit.

B.3 Including Multiple Questions Types to Assess Different Data Quality Issues

This model can be used to assess different aspects of data quality, depending on the K questions chosen for comparison. Using the question typology drawn up by the Abdul Latif Jameel Poverty Action Lab (J-PAL), we can conceive of three main types of questions in this context, which lead to different interpretations of the parameter estimates (Gibson n.d.). The first are questions that are factual in nature — for example, questions about age, gender, first name, last name, and occupation, among others. The responses to these kinds of questions should rarely change, regardless of repetition, and so the parameter estimates drawn from a model fit with agreement summary vectors drawn from these questions will, at the survey and at the respondent level, indicate our uncertainty about whether the information has been accurately collected. These questions help assess the possibility of the wrong person having been re-contacted, hints at data falsification, or indicates shoddy interviewer work.

The second kind of question are ones with responses that are not expected to change between repetition, but which could indicate that enumerators and other survey staff took shortcuts. The goal here is not so much to detect falsification, but to assess issues with the execution of the survey. $\hat{Q}_S$ would represent our confidence in how well the survey was administered.

The final type of question is one that may — but does not have to — change depending on survey context and where there may be slightly more variation over time, such as attitudinal questions. Items used to analyze research questions directly would fall into this category. Using the method described in this paper on these questions would allow one to assess how reliable crucial outcomes are — can we believe that the information we collected represents respondents' true opinions or preferences?

All three types can detect falsification of data, if it exists. However, they lead, in the absence of gross falsification, to different assessments of survey quality, and it is crucial for researchers to realize the implications of the kinds of questions they choose as input to the model. For example, the four variables in Table 1 — last name, monthly income, occupation, and age — all represent information that should not, given a reasonably short time between when questions were asked, provide different information. The number of disagreements between $\boldsymbol{r}_{a_1}$ and $\boldsymbol{r}_{b_1}$ would seem to indicate that these two responses do not come from the same individual, although researchers expected them too. The differences between $\boldsymbol{r}_{a_2}$ and $\boldsymbol{r}_{b_2}$ also hint at issues with data collection; the Karls*en* versus Karls and 21 vs 31 can both indicate typographic errors.

This suggests fitting three different models if we are interested in all three kinds of questions. We can, however, also include questions of all three kinds in one model. Once we do, however, we combine the various sources that would be identified via the separate types of questions.

It is also possible to include information on question type in the model, for more flexibility. If there are J sets of questions, we split K into J K_j's, each representing the number of questions asked of each type. $\boldsymbol{\nu}_{ji}$ becomes the agreement summary vector for questions set j, each with L_j agreement levels. We can either estimate J separate models, or assuming the question sets are independent, we can characterize the joint probability for all J questions sets for response vector i given its match status — $\Pr(\boldsymbol{\nu}_{1i,\ldots,\boldsymbol{\nu}_{Ji}}|M_i)$ — as $\prod_{j=1}^{J}\prod_{l_j=1}^{L_j}\pi_{1l_j}^{\nu_{jil_j}}$, and then fit one, more complex model. The benefit of this approach is that it allows different probabilities of agreement levels for each kind of question.

In either case, $\hat{Q_S}$ would be estimate of the overall quality of the survey, combining the three different types of data quality issues.

B.4 Identifying Falsifying Enumerators

The purpose of the QualMix model is not exclusively identify falsifying enumerators. It does allow researchers and survey practitioners to identify falsifying enumerators and estimate the probability that an enumerator is falsifying observations. However, the model in its simplest form may not be the most efficient way to do so. In order to identify fabrications (not just problematic enumerators) wholesale, it will be more effective to oversample enumerators based on various factors, including not only the quality assessments the mixture model procedure provides, but also incorporating metadata and data collected from respondents in the initial survey. The latter could potentially be done without the need for backchecks of any kind, for example. The benefit of the approach I advance in this paper is that a survey company could use some method to oversample suspicious enumerators but still use the scope of the backcheck data to assess general survey and enumerator quality. A balanced approach would involved weighted observations from oversampled enumerators so that the total of every enumerator's observations count equally for assessing their quality.

C Diagnosing Issues with Model

An inherent risk with any unsupervised learning approach is that the model may overfit and find patterns in the data that may not exist in reality. In this case, except in situations of grievance incompetence or fabrication, that would most likely mean characterizing true matches as non-matches, as one can expect that there would be more matches than non-matches. A possible cause of such a scenario would be if the two estimated multinomial distributions end up being very similar.[22]. This would result in estimated responsibilities "heaping" around .5 in a histogram. I suggest three strategies for detecting such issues. First, looking at $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_1$. Second, plotting a histogram of the estimated responsibilities. Third, seeing how similar the two estimated multinomial distributions are using the Jensen-Shannon Distance, the square root of the Jensen-Shannon Divergence. The Jensen-Shannon Distance is bounded by 0 below and 1 above, with 0 indicating that two distributions are the exact same (Lin 1991; Endres and Schindelin 2003; Nielsen 2011; DeDeo et al. 2013).

C.1 Evaluating the Difference Between Match and Non-Match Distributions in the Simulation

Figure 6 shows the Jensen-Shannon Distance (JSD). We can see that for all parameter combinations it is around .5. Given that the JSD is bounded by 0 and 1, where 0 means identical distributions, this is key evidence that the component distributions of the mixture are sufficiently different.[23]

[22] While the motivation behind the model is to identify matches and non-matches, what the model actual does is identify clusters of similar $\boldsymbol{\nu}_i$

[23] I use the `philentropy` package to calculate the Jensen-Shannon Distance (Drost 2018).

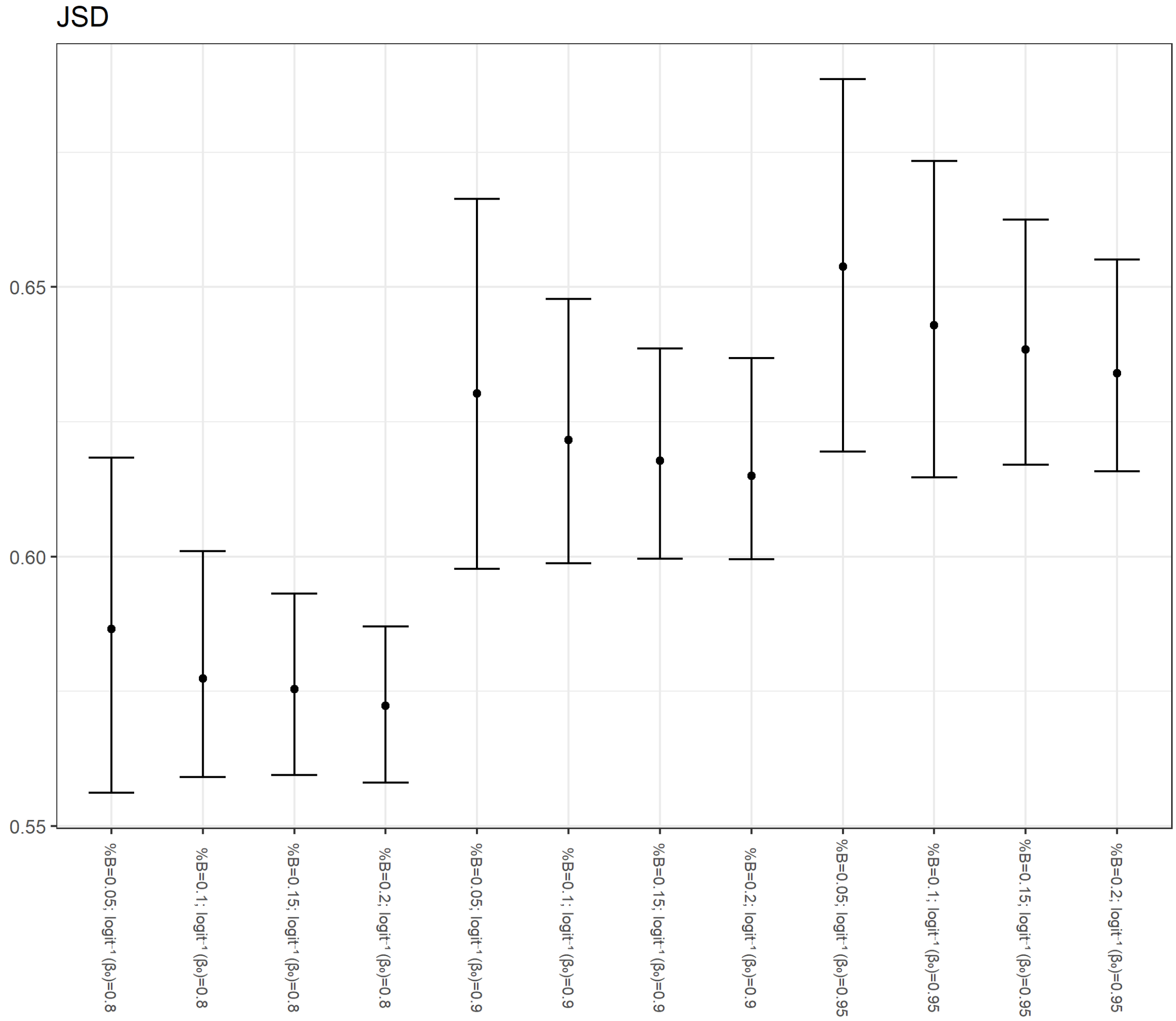


Figure 6: Median of the posterior of the Jensen-Shannon Distance for different parameter combinations with 95% credible intervals.

D Assumptions for Applying Model to Backchecks

In the context of backchecks, we do not have access to two sets of responses for all respondents and will instead extrapolating from a sample of the respondents. In order for these parameters to represent the quantities we want, we need to make the following assumptions:

1. The re-interview values are correct. If the backcheck process creates artificial non-matches — if, for example, backchecking relies on phone numbers, and the respondent provided an incorrect or temporary phone number — then the survey quality and enumerator quality estimates will be flawed.
2. Enumerators are not aware *which* questions are selected for re-interviewing. If they are, they could make sure to ask respondents those questions and then not be as careful with other questions.
3. Backchecking is performed randomly. This is important because when it comes to backchecks, we will have two sets of responses for only a subset of respondents. If the backcheck sample is non-random, the quality assessments derived from this method will be biased.

E General Simulation Process

I simulate two data sets with mistakes in the first data set in the following way. Simulation parameters are in italics, with the fixed values used in the simulation in the paper in bold. The simulation process is slightly different if the probability of a mistake is stratified by enumerator. Steps involved only if this is the case are marked "[enumerator]" The simulation process is slightly different for the backcheck case. Steps involved only in the backcheck simulation are marked "[backcheck]".

1. Start with an existing survey data set ($\boldsymbol{R_b}$). This will be the "correct" set.
2. Choose a baseline *proportion of mistake*: λ (actually the inverse logit function applied to β_0).
3. [enumerator] If a specific *number of enumerators* is desired, first drop all observations from enumerators who have fewer respondents than some user-determined number (**35**). This step ensures that randomly selected enumerators do not have low numbers of respondents. This is purely for stability's sake. Randomly sample the requested number of enumerators from all remaining enumerators.
4. [enumerator] Draw enumerator intercepts β_e from $\mathcal{N}(0, \sigma_{beta_e})$, with some user-determined *variance* (**1**). Calculate each enumerator's match proportion λ_e by combining β_0 and β_e.
5. Decide which observations will be matches and which will be non-matches. This is random with respect to the observations picked, but the proportion of matches is fixed.
6. Create a copy of the original data set, $\boldsymbol{R_a}$. Replace non-match observations in $\boldsymbol{R_a}$ with observations chosen at random from all match observations (there will then be duplicates).
7. Next, induce small mistakes in match observations $\boldsymbol{R_a}$ via the following steps:

   (a) Decide the maximum possible number of variables that can be changed for any one observation (as a *proportion of variables*) (**.7**).

(b) Decide for each observation how many variables will be changed by drawing from either

- [enumerator] a binomial distribution where the number of trials is the maximum decided in the previous step and where π is the inverse of the probability of a match (i.e. "lower" quality enumerators will have a higher number of variables changed). We set a *lower bound for this probability* (**.1**) to represent the fact that humans are not infallible (even the best enumerators will make some mistakes).[24]
- A discrete distribution where the categories are the numbers 0 to the maximum number of variables possible, with $\pi_i = \frac{(\text{Max. \# of Vars.}+1)-i}{\sum_{i=0}^{\text{Max \# of Vars.}} i+1}$, $i = 0, ..., \text{Max \# of Vars.}$

(c) Randomly choose the variables that will be scrambled by selecting the number of variables determined in the previous step from all possible variables with equal probability.

(d) For each observation, set the number of variables that will be scrambled and which will be perturbed. A fixed *proportion of variables* is chosen (i.e. this does not vary by observation) (**.5**) from the variables picked in the previous step.

(e) To scramble, replace the chosen variables with incorrect ones from an observation from $\boldsymbol{R_b}$ chosen at random.

(f) To perturb, insert small mistakes into the existing response value. Different mistakes are possible:

- For ordered factor variables, replace the current value with an adjacent one. For unordered factor variables, do nothing.
- For numbers, with equal probability: transpose two digits at random, insert

[24] Note that this makes the simulation not quite match the model, which is simpler. In fact, it should make it *harder* for the model to correctly identify mistakes and assess overall and enumerator quality because this will directly impact ν_i's.

a typo (replace a digit with a numerically adjacent number), or delete a digit at random. With very small probability (.05) change the sign of the variable.

- For characters, with equal probability: transpose two letters at random, insert a typo (replace a letter with a keyboard adjacent letter), or delete a letter at random.

8. [backcheck] Sample a portion of observations to reflect backchecking (*backcheck portion*). Retain the entire incorrect survey as well. [enumerator] Stratify by enumerator.

Data Preparation

To prepare the data for the simulation, I drop all observations with NA values in these variables. I also randomly choose thirty-five enumerators from all enumerators with more than 150 observations.[25] This results in 9,973 total observations. For each set of parameters, I then simulate fifty original data–backcheck data pairs. For each simulated original data–backcheck data pair, I then calculate agreement summary vectors for all original data–backcheck data observation pairs in the manner described in Section 3.3 and Appendix A.

[25] There are forty such enumerators, from an original fifty-one. The chosen enumerators all have between 171 and 352 respondents.

E.1 Simulation Model Specification

I fit the QualMix model to the agreement summary vectors using `Stan`'s `R` interface `rstan` (Team 2020). I use following model specifications:

$$\begin{aligned}
\boldsymbol{\nu}_i &\overset{\text{i.i.d}}{\sim} \text{MixMulti}(\lambda_{e_i}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_0) \\
\lambda_e &= \text{logit}^{-1}(\beta_0 + \beta_e) \\
\beta_e &\sim \mathcal{N}(0, \sigma_e) \\
\beta_0 &\sim \mathcal{N}(\mu_{\beta_0}, \sigma_{\beta_0}) \\
\beta_e &\sim \mathcal{N}(0, \sigma_e)\sigma_e \qquad\qquad \sim \text{Gamma}(1,1) \\
\boldsymbol{\pi}_1 &\sim \text{Dir}(1,2,3) \\
\boldsymbol{\pi}_0 &\sim \text{Dir}(3,2,1) \\
\mu_{\beta_0} &\sim \mathcal{N}(0,1) \\
\sigma_{\beta_0} &\sim \text{Gamma}(1,1)
\end{aligned}$$

I run each fit of the model (on each of the 100 simulated datasets for each of the 12 parameter combinations) for 1500 iterations each on four chains (for a total of 3000 post-warm-up samples from the posterior in each iteration).[26]

When fitting an unsupervised mixture model, the labels are generally not identified – the model cannot by itself decide to which distribution (i.e. high- or low-quality) $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_0$ correspond. To identify the model, I force $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_0$ to follow an ordering — the probabilities in $\boldsymbol{\pi}_1$ *must* be in ascending order, while in $\boldsymbol{\pi}_0$ they must be in descending order, so that high-quality records tend to have a higher probability of similar entries in the agreement summary vector, and vice-versa for low-quality records.

[26] R-hat values for all parameters were all 1 or very close to 1.

E.2 Simulating Effects of Measurement Error

To simulate outcome y (which also becomes a backcheck variable) I use the following data-generating process during *each* simulation:

$$
\begin{aligned}
\mu_i = {} & 1.45 - .25 * \text{Age} - 1.3 * \text{Always Pay Fee} + 2.35 * \text{Female} + 0.67 * \text{Household Income} + \\
& 0.3 * \text{Profit vs Last Year: My profits are lower today} + \\
& 1 * \text{Profit vs Last Year: My profits are about the same} + \\
& 2 * \text{Profit vs Last Year: My profits are higher today} + \\
& 3 * \text{Profit vs Last Year: My profits are much higher today} \\
y_i \sim {} & \mathcal{N}(\mu_i, 5)
\end{aligned}
$$

for all i, where i indexes observations in the original survey. Age takes on values over 18. Always Pay Fee takes on values between 0 and 10. Female is a dummy variable. Household Income takes on values greater than 0. Profit vs Last Year is an ordered categorical variable, with baseline level "My profits are much lower today."

I then insert measurement error by simulating matches and mismatches, as described in the parent section. Next, I fit a correctly specified linear model to the *full* mismatched and measurement-error-containing data set produced during the simulation.[27] Finally, I calculate error as a percentage of the original effect size: $\text{error}_j = (\hat{\beta}_j - \beta_j)/\beta_j$ for all j predictors.

[27] I use `lm()` in `R`.

F Identifying Low-Quality Observations

This appendix assesses how well the model is able to identify low-quality (and potentially falsified) observations. Figure 7 shows the Area Under the Receiver Operating Characteristic Curve for all parameter combinations, calculated out of sample (on the observations not selected for the backcheck in each simulation). The figure demonstrates that the model does exceptionally well at identifying observations that are not the same between $\boldsymbol{R_a}$ and $\boldsymbol{R_b}$, with AUCs very close to 1. There are some minor differences in performance – the change between the lowest median AUC and the highest is only 0.005965. In general, performance improves the higher the average match proportion. Model performance also somewhat improves as the backcheck portion increases, although this is not consistent across overall match proportion.

In this application, we are worried about both false negatives and false positives – determining that two sets of data do not match when they do, and determining that two sets of data match when they in fact do not. Figure 8 shows the false negative and false discovery rate (also known as the false positive rate), considering an observation a match if its responsibility is greater than or equal to .5. As the figure shows, both decrease strongly as the average match proportion increases. In other words, if there are more observations that match, the model makes fewer mistakes with respect to both false positives and false negatives. The figure also shows that, keeping the overall match probability constant, there is little change when the backcheck proportion increases — the false negative rate goes slightly up (which makes sense, because there are more chances to make mistakes), while the false discovery rate generally goes down (which again makes sense, because the model has seen more data). However, these differences are very small in substantive terms. This should be reassuring to researchers and survey implementers, as more backchecks require more resources.

In summary, the model performs well out of sample when it comes to identifying non-matching (i.e. low-quality) and matching (i.e. high-quality) observations. The fact that the model successfully identifies non-matching observation in this context demonstrates its utility for this type of application—assessing data quality and identifying backcheck issues.

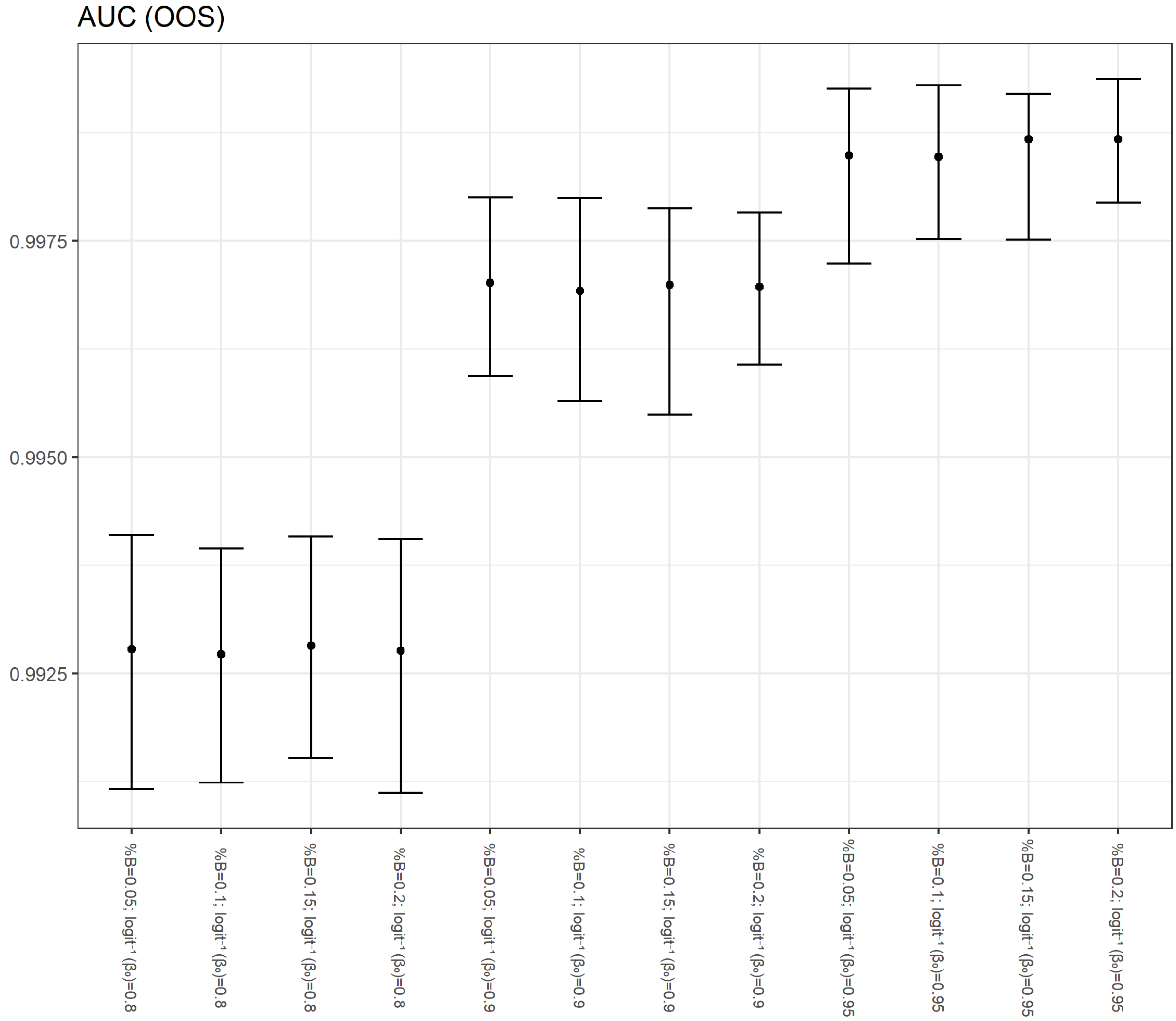


Figure 7: Median of the posterior of the Area Under the Receiver Operating Characteristic Curve for different parameter combinations with 95% credible intervals.

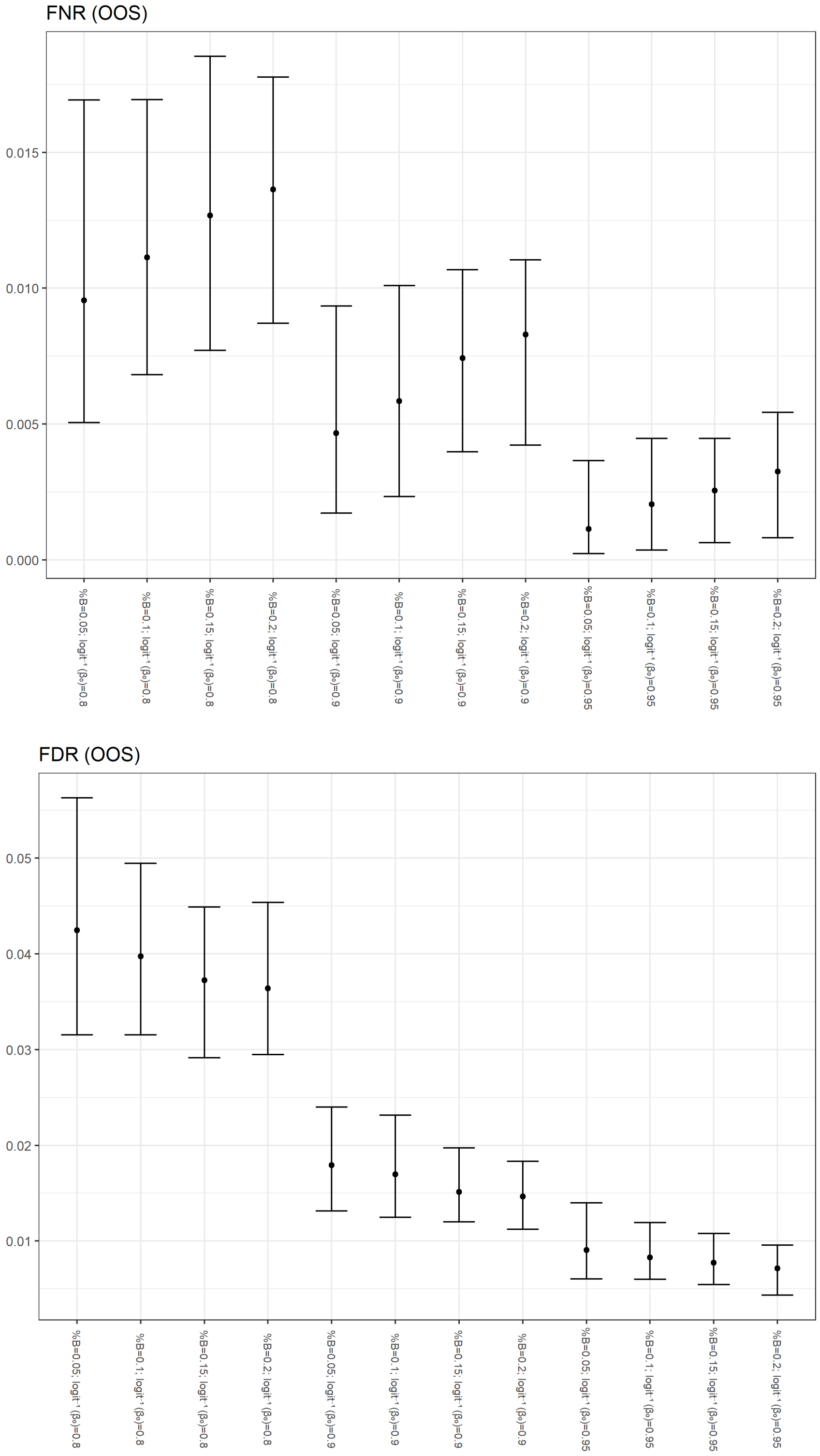


Figure 8: Median of the posterior of the False Negative Rate and the False Discovery Rate for different parameter combinations with 95% credible intervals.

G Enumerator Quality Plot

This plot shows the bias in enumerator quality for each of the thirty-five enumerators in the simulation. Each color represents a different enumerator (there were 35 enumerators used in the simulation), with lines connecting points to make it easier to follow patterns.

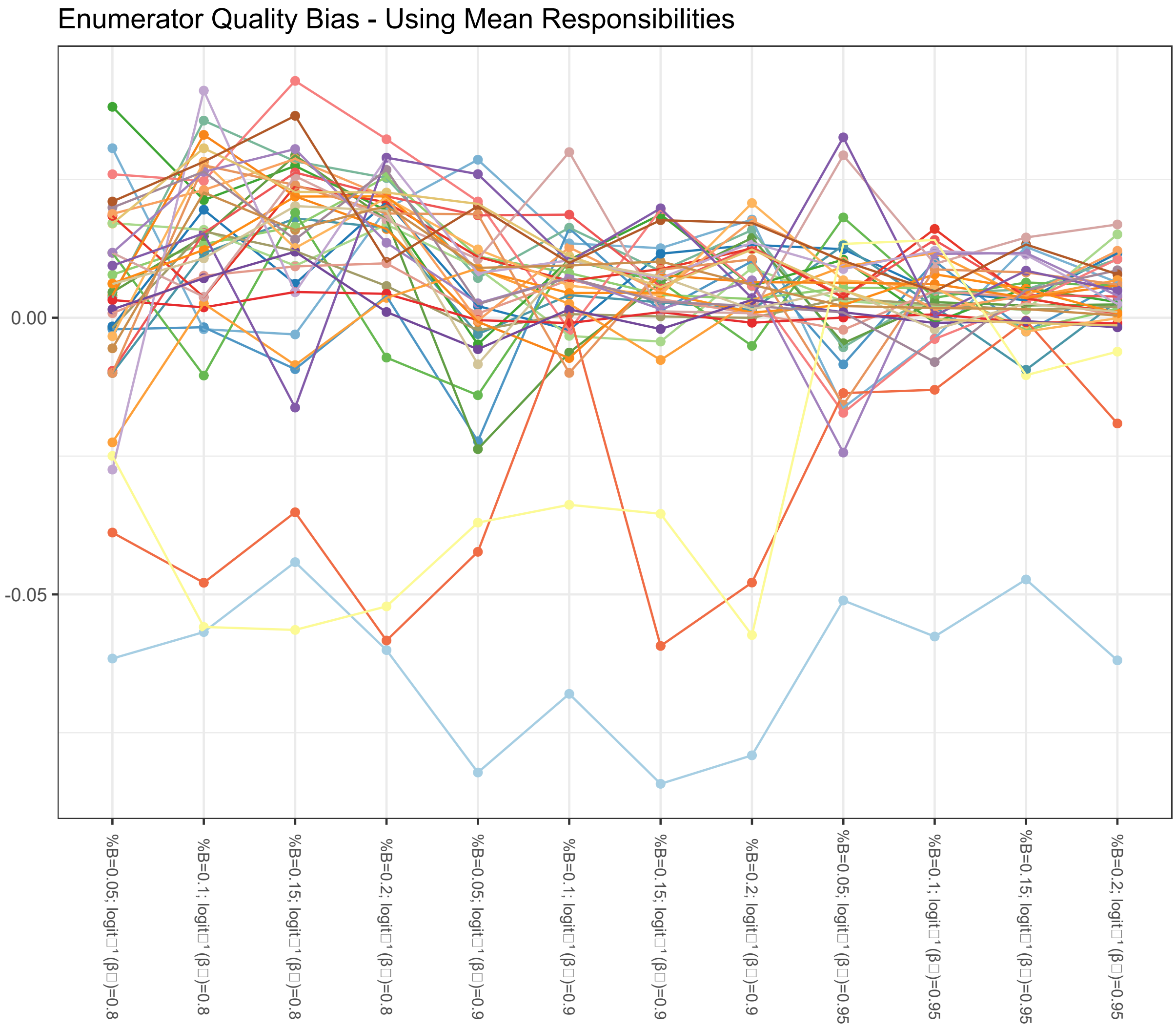


Figure 9: Bias of enumerator quality for different parameter combinations.

H Results Dropping Failed Backchecks

There were some idiosyncrasies with the backchecking process for this survey. Backchecks were done using telephones; although this saves on expense, it is possible that individuals would sell their phones or sim cards in between the time the survey was in the field and when the backchecking occured. Interviewers were directed to ask no further backcheck questions if the name of the person who picked up the phone did not match the one in the survey. It is possible that the *name* was incorrectly collected, but that the respondent was re-contacted successfully.

Because of these idiosyncrasies, I also apply the model to only observations where the full backcheck was completed. When we removed "failed" backchecks, we can see in Figure 10 that the model can still identify two distinct distributions, although there is considerable more uncertainty about the low-quality distribution. This is because fewer observations qualify for this distribution, with most posterior probabilities of a match heaped around 1, as 11 shows. The 95% credible interval for the Jensen-Shannon Distance for this application is [.276, 464]. The large interval comes from the uncertainty around the non-match distribution.

Figure 12 shows updated enumerator quality estimates. Unfortunately there is considerable uncertainty about enumerator quality — this is because there is similar uncertainty about the posterior probability of a match for each respondent, which comes from uncertainty around the two distributions. We see a similar issue with the estimate of overall survey quality, with a 95% credible interval of [0.880, 0.999]. It is important to note that the incomplete backchecks are clearly not missing at random. As such, dropping these observations represents losing valuable information on quality – this also means that these data quality estimates are most likely biased estimates of overall survey quality. It is likely that the model struggles to identify a low-quality and a high-quality distribution in the remaining observations because they are more similar.

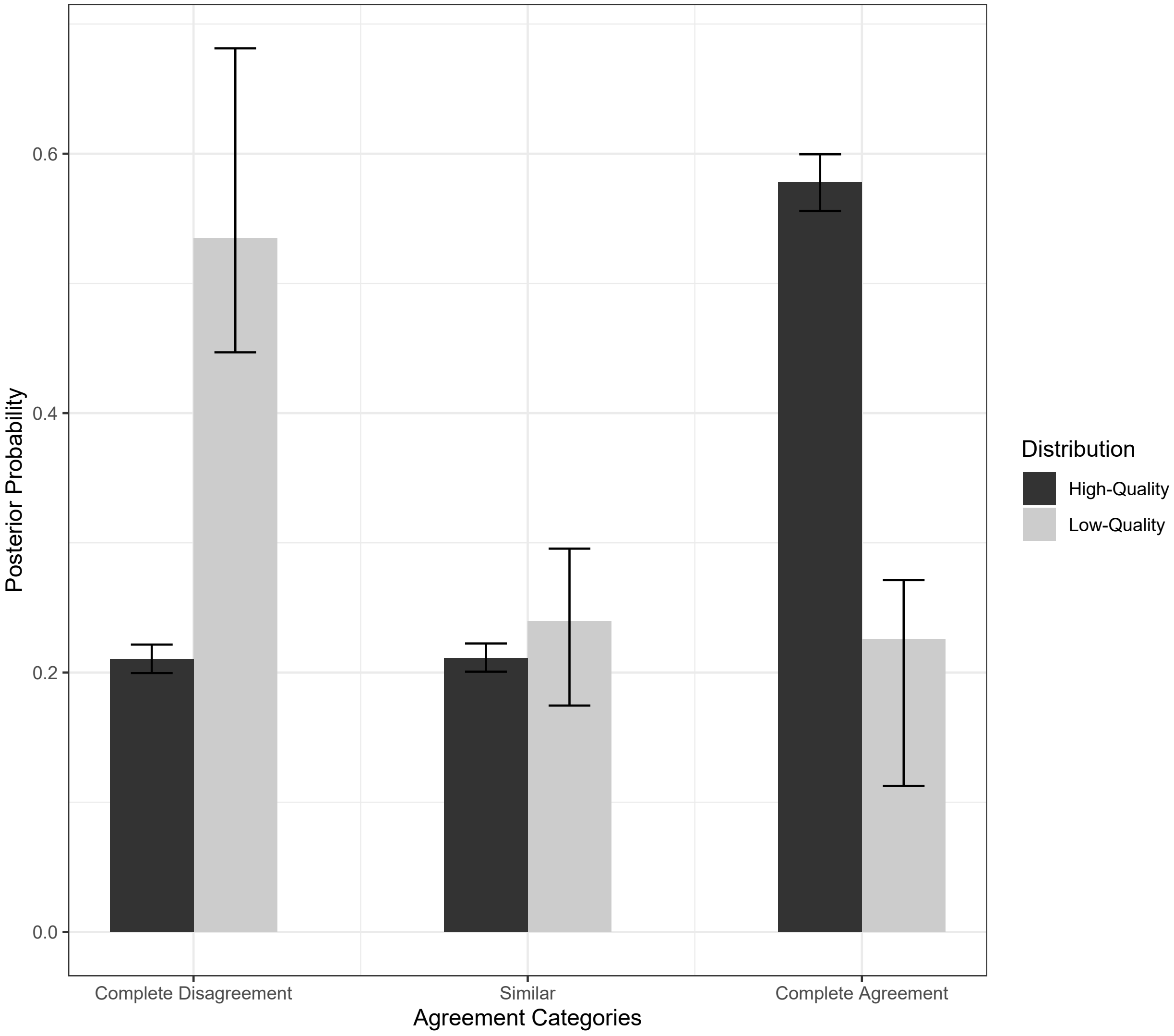


Figure 10: $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_1$ when backcheck pairs where $\nu_0 = 6$ are omitted.

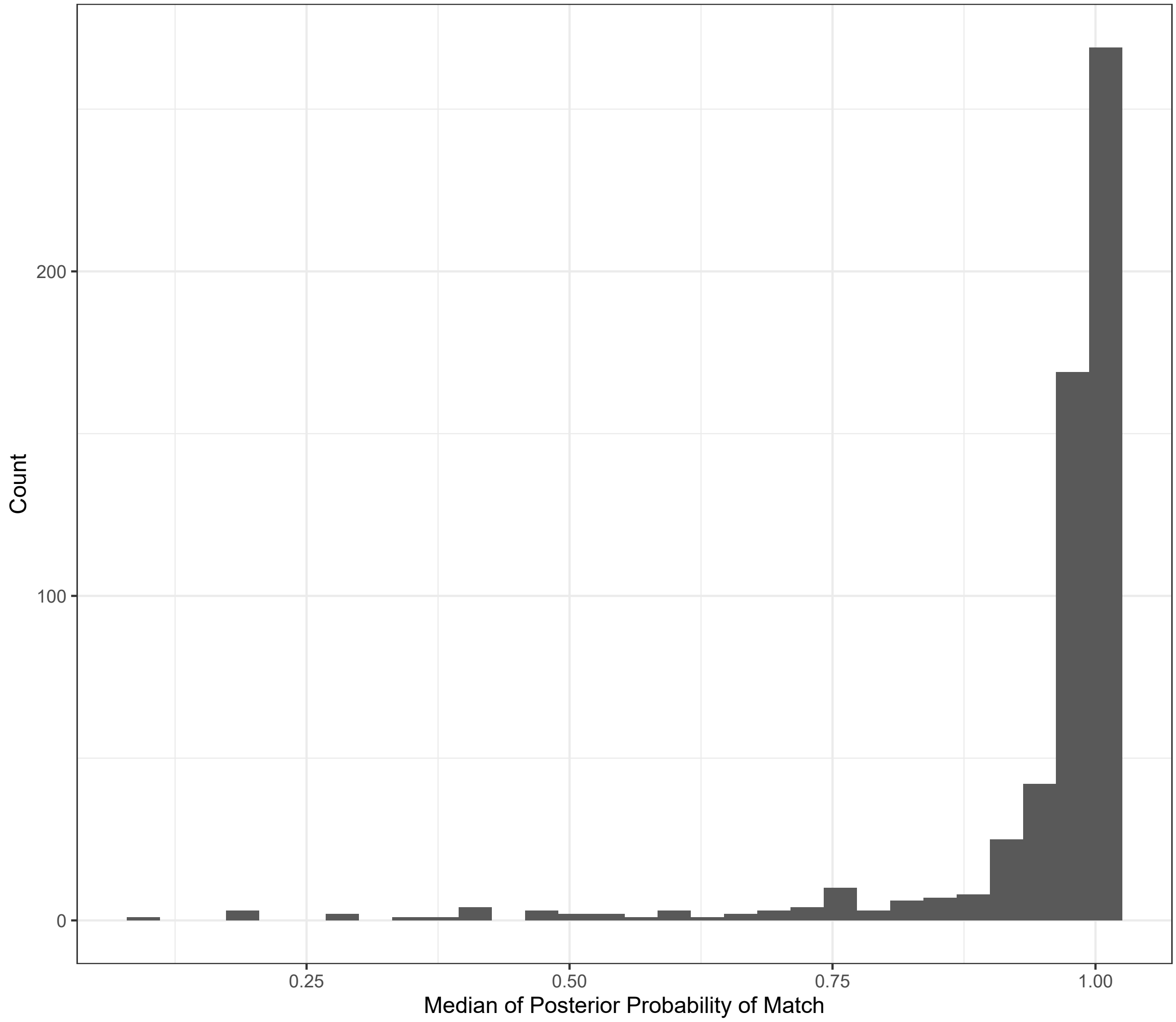


Figure 11: Histogram of the median of the posterior of the posterior probability of a match when backcheck pairs where $\nu_0 = 6$ are omitted.

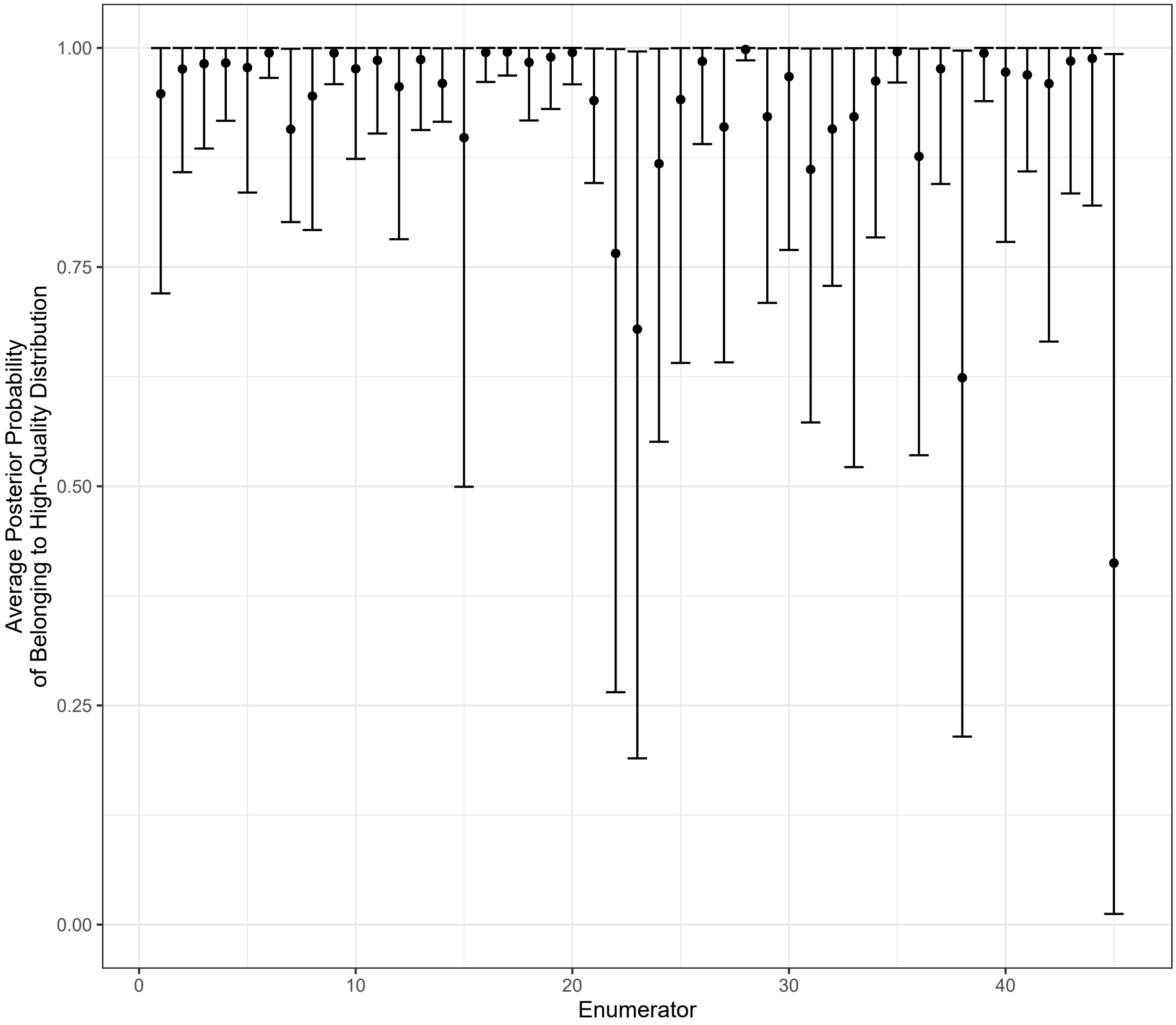


Figure 12: Median of the Posterior of the Average Posterior Probability of a Match for all 45 Enumerators when backcheck pairs where $\nu_0 = 6$ are omitted. Error bars show 95% credible intervals.

I Receipts and Enumerator Quality

In this appendix, I describe the validation exercise. Because I use the product of a Bayesian model as the predictor in this analysis, I incorporate uncertainty about the data into the model fitting process. For completeness, I describe the initial modeling attempt, which failed due to features of the data. I then present the approach used for the results presented in the main body of the paper and present further results.

I.1 Initial Modeling Attempt

Initially, I incorporated enumerator data quality estimates (the distribution of which I refer to as $\hat{Q}_e$) from the main model as a prior on latent true enumerator quality Q_e. For each enumerator e:

$$\begin{aligned}
y_e =& \text{\# of Receipts Reported Shown} \\
n_e =& \text{Total \# of Respondents Interviewed} \\
y_e \sim& \text{Binomial}(n_e, \pi_e) \\
\pi_e =& \text{logit}^{-1}(\beta_0 + \beta_0 * Q_e) \\
Q_e \sim& \text{Beta}(u_e, v_e) \\
\mu_e =& \overline{\hat{Q}_e} \\
\gamma_e =& \hat{\mathbb{V}}(\hat{Q}_e) \\
u_e =& \left(\frac{1-\mu_e}{\gamma_e} - \frac{1}{\mu_e}\right) * \mu_e^2 \\
v_e =& u_e * \left(\frac{1}{\mu_e} - 1\right) \\
\beta_0, \beta_1 \sim& \mathcal{N}(0, 1)
\end{aligned}$$

I chose the Beta distribution for Q_e because its support is [0, 1] — enumerator data quality is similarly constrained to this interval. However, the log-likelihood becomes negative infinity when the value is exactly 1 or 0. Because there are quality estimates that are at or very close to 1, this presented problems for the sampling algorithm (as in the rest of the paper, I used `Stan`), resulting in almost a quarter of transitions being divergent. This led me to a different solution.

I.2 Working Model

As the previous approach did not work, I instead pick 1000 random values from the estimated posterior for each $\hat{Q}_e$. I then fit the following model, where e indexes enumerator and i indicates the sample from the estimated posterior:

$$\begin{aligned}
y_e =& \text{\# of Receipts Reported Shown}\\
n_e =& \text{Total \# of Respondents Interviewed}\\
y_e \sim& \text{Binomial}(n_e, \pi_e)\\
\pi_e =& \text{logit}^{-1}(\beta_0 + \beta_0 * \hat{Q}_{e,i})\\
\beta_0, \beta_1 \sim& \mathcal{N}(0,1)
\end{aligned}$$

I ran each model for 1000 post-warm up iterations on two chains.[28] This results in 1000 model fits, each with 2000 draws from the posteriors of the parameters.[29] I pool the posteriors for β_0 and β_1, separately, across all 1000 model fits. This allows me to incorporate uncertainty about enumerator data quality into this model.

I then use the pooled posteriors for β_0 and β_1 to calculate the predicted probability of showing a receipt for enumerators with quality .5, .75, and 1. I then calculate the change in probability when quality goes from .5 to .75, and from .75 to 1.

[28] Rhat values were all very close to 1, and ess_bulk and ess_tail were all above 200.

[29] There are only 45 observations in this model (The 45 enumerators for whom I was able to derive quality estimates using backchecks). Each model fit took less than a second, so this procedure was not very computationally demanding.

I.3 Results

Figure 13 shows the change in probability of a receipt being reported shown by an enumerator for different changes in enumerator data quality. Figure 14 shows how the probability changes as a function of enumerator data quality. The rug plot at the bottom indicates where estimated enumerator data qualities fall.

Figure 13: Median of the posterior predictive distribution of the difference between the probability at different values of enumerator quality. Error bars show 95% credible intervals.

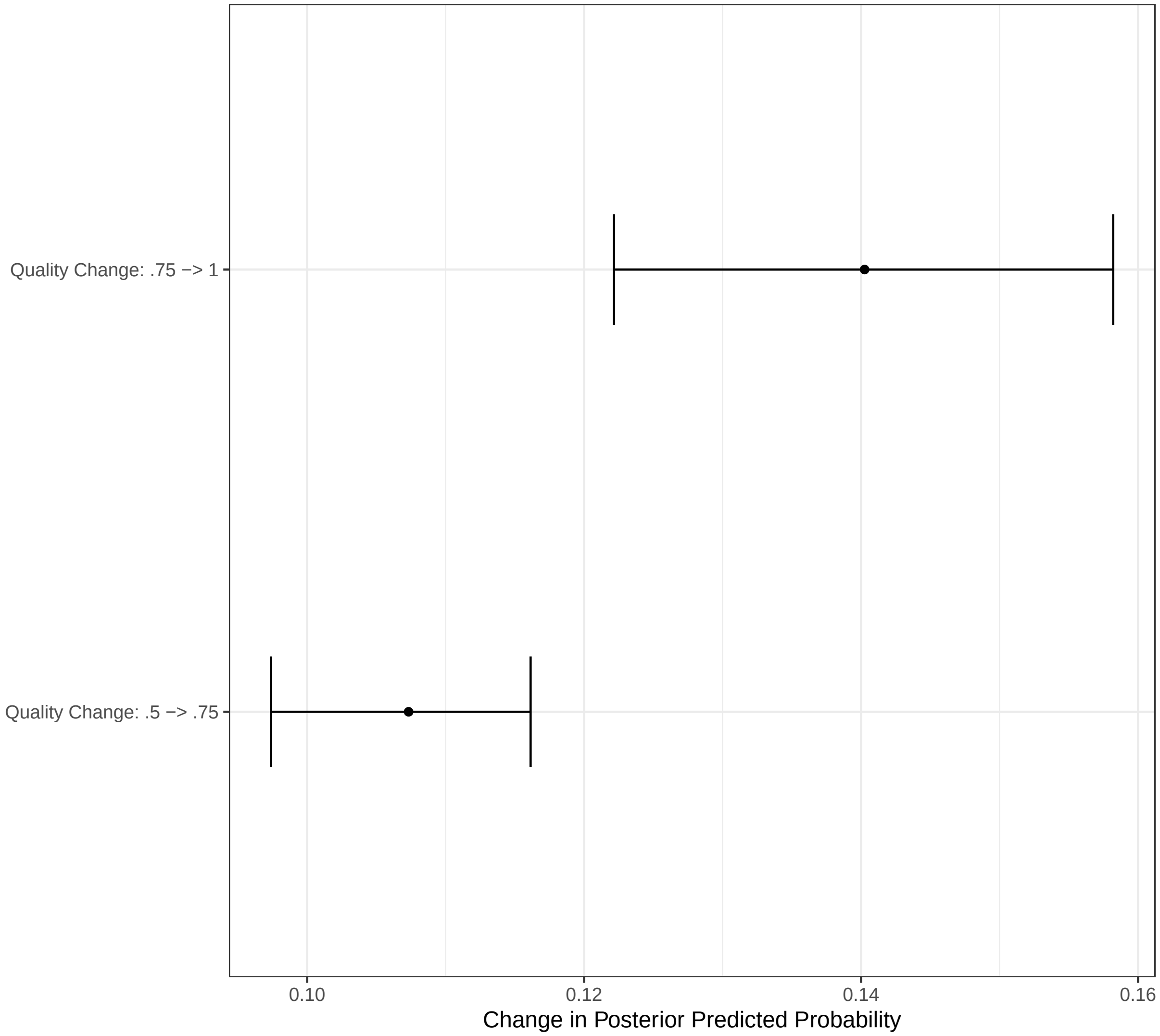
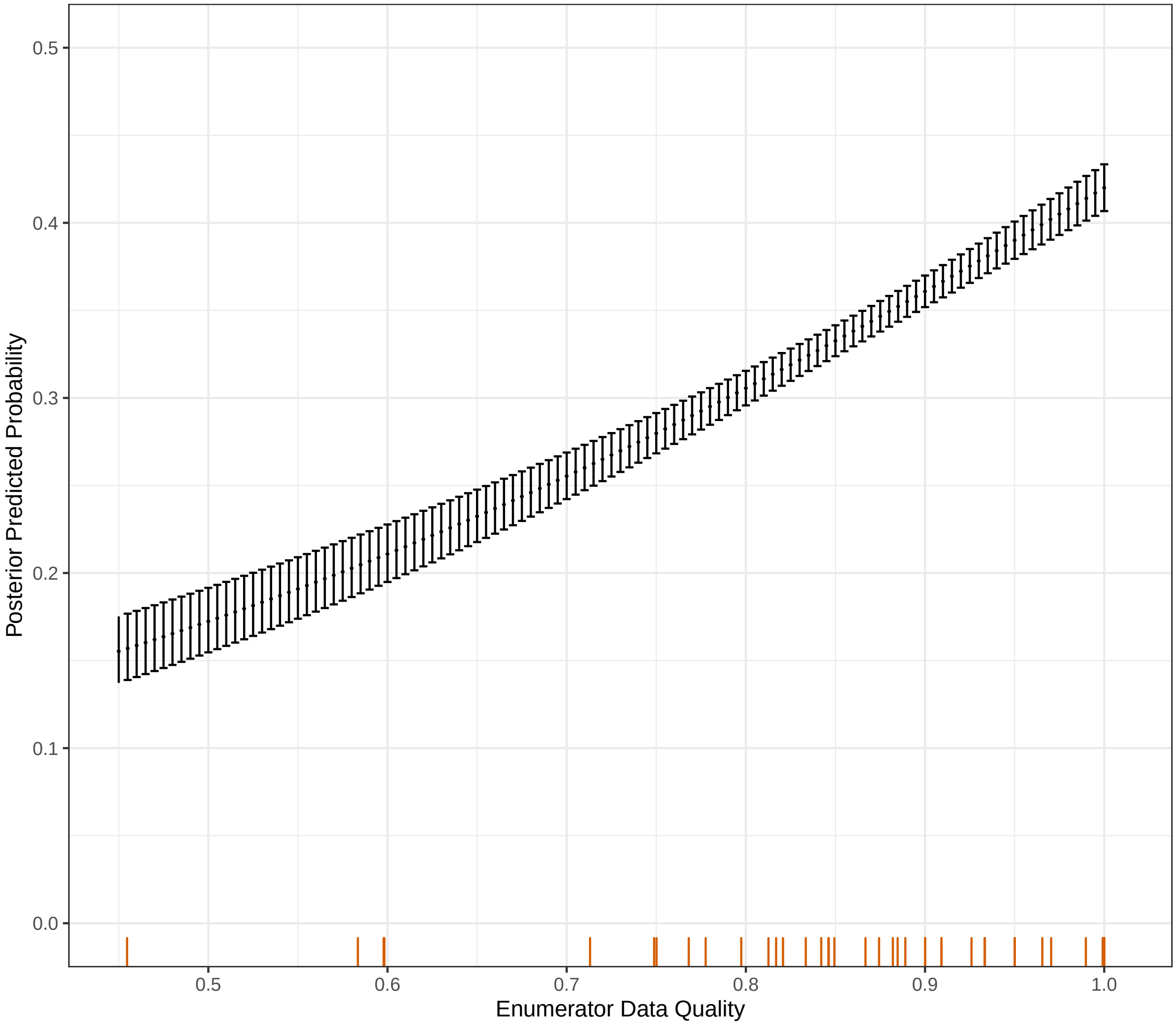


Figure 14: Median of the posterior predictive distribution of the probability of being shown a tax receipt at different values of enumerator quality. Error bars show 95% credible intervals.



J AAPOR Code of Professional Ethics & Practices III.A Information

J.1 Data Collection Strategy

Data collection for the data used in this paper took the form of an interviewer-implemented face-to-face survey

J.2 Who Sponsored the Research and Who Conducted It

The data used in this paper were **not** collected for the purposes of this paper. This research has no sponsor.

The data used in this study come from the endline survey for an impact evaluation study of the USAID/Malawi Local Government Accountability and Performance (LGAP) Activity. The impact evaluation was contracted out to NORC at the University of Chicago. This is associated with the following USAID information: DRG Learning, Evaluation, and Research (LER) Activity; Tasking N030; Contract No. GS-10F-0033M/AID-0AA-M-13-00013.

The LGAP project itself was contracted out to DAI. The associated contract number is AID-OAA-I-14-00061/AID-612-TO-16-00004

The survey data were collected by Innovations for Poverty Action, under contract from NORC at the University of Chicago.

J.3 Measurement Tools/Instruments

This subappendix contains the survey items used and the associated response options in the simulations and in the empirical example, as well as the validation exercise.

J.3.1 Survey Items Used in Simulations

1. What is the respondent's gender? (*not asked; enumerator selected*)

   - Male
   - Female

2. How old are you?
   - [integer value]
3. What is the highest level of education you have **completed**?
   - None
   - Nursery School
   - Standard 1
   - Standard 2
   - Standard 3
   - Standard 4
   - Standard 5
   - Standard 6
   - Standard 7
   - Standard 8
   - Form 1
   - JCE/Form 2
   - Form 3
   - MSCE/Form 4
   - Technical/Private College (non-Degree)
   - Degree
   - Masters
   - PhD

- Other
- Refused to Answer

4. What is your estimated total household monthly income? In other words, how much do the adults in your household earn in total each month from all sources, full- and part-time employment, businesses, investments, and other fees or services?
    - [integer value]

5. How frequently do you sell in this market?
    - Every Day
    - 1-3 days a week
    - 4-6 days a week
    - A few days each month, but not every week
    - Once a month
    - Once every few months
    - Once a year
    - This is my first time
    - Refused to Answer

6. Enumerator: Select the activity that most closely matches the ¡b¿main¡/b¿ service or good provided. (*not asked; enumerator selected*)
    - Retail - Groceries; Retail - Wine,beer,liquor,soft drinks sale; Retail - Agricultural produce (Perishable Fruit/Leafy); Retail - Agricultural produce (Storable Grains/Legumes); Retail - Animal produce(meat,fish); Retail - Cooked food and snacks; Retail - Hardware; Retail - Timber/wood/charcol; Retail - Clothes/shoes; Retail - Motor vehicle spare parts; Retail - Stationery/Printing; Retail - Cosmetics, beauty products; Retail - Electronics and other appliances; Retail - Curios/Handcrafts/Art; Retail - Cell phone units, SIM card retailer; Retail - Bags/Plastic

Bags/Sacks; Retail - Plastics; Retail - Agricultural Goods; Retail - Cooking Oil; Retail - Sale of other products; Service - Milling (incl. Hand milling); Service - Food processing; Service - Canning; Service - Beer brewing; Service - Carpentry, joinery, metal work; Service - Tailoring, knitting, leather products, shoe repair; Service - Mini-bus; Service - Bicycle taxi; Service - Other transport of passenger; Servcie - Transport of products; Service - Storage and warehouse; Service - Hair salon/Barber Shop; Service - Cleaning; Service - Auto repair; Service - Battery charge; Service - Other repair and maintenance services; Service - Collection/Sale of firewood, fetching water; Service - Laundry or ironing; Service - CD Burning; Service - Videoshow/Cinema; Service - Phone Repair; Service - Welding; Catering; Restaurant/Bar/Tavern; Hotel/Guest house; Mobile Money; Financial institution; Real estate; IT services; Nursery(child care)/School; Health clinic; Pharmacy/Herbalist; Arts and sports; Other

7. In general, how do your profits today compare to your profits in [current month] 2017?
   - My profits are much higher today
   - My profits are higher today
   - My profits are about the same
   - My profits are lower today
   - My profits are much lower today
   - Refused to Answer
   - Don't Know

8. Here are 10 tokens that represent all the vendors in this market. Please separate these into three piles. Put here *[indicate location]* the vendors who pay their fee every day they sell in the market. Put here *[indicate location]* the vendors who pay their fee sometimes but not always. Put here *[indicate location]* the vendors who never pay their fees.

- [3 integer values, summing to 10] (*only "number of vendors pay every day" used*)

J.3.2 Survey Items Used In Empirical Application

Four survey items overlap with the simulations and are not shown again here: age, education, frequency of selling, and the stall type.

1. Can you show me the last receipt you received from paying fees?
    - No Receipt
    - Receipt Available
2. In general, how satisfied are you with the developments in THIS market provided by the district government?
    - Very Satisfied
    - Somewhat Satisfied
    - Somewhat Dissatisfied
    - Very Dissatisfied
    - Refused to Answer

All questions were worded the same in the backcheck and in the original survey except for the receipt question. The question used to backcheck the receipt information was:

- Did you show the original interviewer a receipt you received from paying fees?
    - Don't Know
    - Refused to Answer
    - No
    - Yes

J.3.3 Survey Item Used For Validation Exercise

1. Can you show me the last receipt you received from paying fees?
    - No Receipt
    - Receipt Available

J.4 Population Under Study

The population under study was market vendors in 128 markets[30] in eight districts[31] in Malawi from October 2018 to January 2019.

J.5 Method Used to Generate and Recruit the Sample

Enumerator teams visited each of the 128 markets once during the enumeration period. In each market, enumerators sought to recruit 100 respondents using a random walk procedure. Enumerator teams of ten individuals determined the best division of the market to facilitate the random walk. They divided the market into five roughly equal in size sections. Pairs of enumerators were assigned to each section. Each pair then divided their section again. Together, they planned routes that would take them past all market vendors in their half section. This included counting all market vendors in their section. The enumerators then

[30] Mpale, Nthandizi, Ulongwe Market, Kaliyati, Kantwanje, Phalula, Chiyenda Usiku, Kachenga, Mwaye, Balaka Main Market, Mbela, Mwima, Dziwe, Mdeka, Chilobwe, Ntonda, Chikuli, Linjidzi, Lirangwe, Mombo, Checkpoint, Chima, Chinkhoma, Kamboni, Kawamba, Mtunthama, Bua, Chatoloma, Chisemphere, Kasera, Mankhaka, Wimbe, Chiseka, Chulu, Katondo, M'Doni, Mpepa, Santhe, Chamama, Chitenje, Katenje, Mnkhota, Ndonda, Nkhamenya, Chigwirizano, Malingunde, Nathenje, Nsalu, Chinsapo 2, Kamphata, Msundwe, Namitete, Malembo, Mbang'ombe, Mchezi, Nkhoma, Kabudula, Kasiya, Mitundu, Mpingu, Liwonde Central Market, Mpita, Nayuchi, Nsanama, Nselema, Ntaja, Chikweo, Ngokwe, Edingeni, Enukweni, Euthini, Kazuni, Luzi, Mzimba Market, Ekwendeni, Eswazini, Jenda Market, Kafukule, Mpherembe, Mzalangwe, Bulala, Embangweni, Engucwini, Kawonekera, Madede, Monolo, Bwengu Market, Engalaweni, Kapando, Luviri, Mafundeya, Manyamula, Chikuse, Macholowe, Nalikata, Namtombozi, Wendewende, Chimbalanga, Limbuli, Mathambi, Mizimu Trading, Nachimango, Laudadelo, Mbowela, Mpala, Mpholiwa, Sadibwa, Chitakale, Kambenje, Namphungu, Njala, Nkando, Chimwalira, Govala, Malosa, Ngwalangwa, Chingale, Makina, Namadidi, Sakata, Chinseu, Jali, Namasalima, Six Miles, Kachulu, Mayaka, Songani, and Thondwe

[31] Balaka, Blantyre, Kasungu, Lilongwe, Machinga, M'mbelwa, Mulanje, and Zomba.

determined the skip pattern that would result in 10 responses each. If a market vendor refused to participate, enumerators were directed to move on to the next respondent.

Vendors who participated received either 200, 300, or 600 Malawian kwacha in airtime vouchers. There were two versions of the survey. A short survey and a longer version that included many more questions survey. The short survey took roughly 15 minutes to complete. The long survey took up to an hour to complete. 80% of respondents answered the short survey, while the remaining 20% responded to the long survey. Who answered which survey was also determined using a pre-determined skip pattern (to ensure that 2 out of the 10 respondents each enumerator interviewed would respond to the longer survey). Respondents who completed the short survey received 200 MWK worth of airtime. There was a delayed gratification experiment embedded in the long survey. Respondents could receive 300 MWK worth of airtime immediate or 600 MWK worth of airtime at a later point.

J.6 Methods and Modes of Data Collection

Responses were collected face-to-face. Enumerators used tablets to collect respondents' answers. The survey was available in English, Chichewa, Chitumbuka, and Chiyao.

J.7 Dates of Data Collection

Data collection occurred between October 30, 2018 and January 17, 2019. The bulk of data collection was completed by December 15, 2018 (which is why the last backcheck day was December 17, 2018). There were some concerns about incomplete data in one of the markets (Jenda Market), and so it was visited again on January 17, 2019.

J.8 Sample Sizes

12,370 responses were collected during enumeration. Not all markets had 100 respondents, resulting in fewer than 12,800 responses over all.

J.9 How the Data Were Weighted

The data were not weighted.

J.10 How the Data Were Processed and Procedures to Ensure Data Quality

IPA performed high-frequency checks. A backcheck (re-contacts) was also conducted. Data were colllected on tablets using SurveyCTO. Logic checks were built into the survey.

This project represents a new way to assess the quality of data, using re-contact data in this particular case.

J.11 Acknowledging Limitations of Design and Data Collection

This design was chosen because it was infeasible to construct a full sampling frame of all market vendors in these 128 markets in the eight districts in Malawi. It has its drawbacks, in particular putting a lot of the onus on enumerators for the construction of the random sample. As with any design, there is the potential for unmeasured error. The aim of this project is to assess the potential for unmeasured error in data such as these.