A Mixture Model Approach to Assessing Measurement Error in Surveys Using Reinterviews

Simon Hoellerbauer*

August 9, 2023

Abstract

Researchers are often unsure about the quality of the data collected by third-party actors, such as survey firms. They are reliant on survey firms to provide them with estimates of data quality and to identify observations that are problematic, potentially because they have been falsified or poorly collected. This may be because of the inability to measure data quality effectively at scale and the difficulty with communicating which observations may be the source of measurement error. To address these issues, I propose the QualMix model, a mixture modeling approach to deriving 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 reinterviews, a common form of data quality assessment used to detect falsification and data collection problems during enumeration. Through simulation based on real-world data, I demonstrate that the model successfully identifies incorrect observations and recovers latent enumerator and survey data quality. I further demonstrate the model's utility by applying it to reinterview data from a large survey fielded in Malawi, using it to identify significant variation in data quality across observations generated by different enumerators.

Word Count: 7,021

*Postdoctoral Fellow of Data Science and Society, Vassar College. Email: shoellerbauer@vassar.edu

Statement of Significance

This paper presents a novel way to assess survey data quality in situations in which data producers like survey organizations have two sets of responses, ostensibly from the same respondents, for at least a portion of the overall survey respondents. This situation exists in the case of reinterviews. In this context, it provides a framework for using reinterviews to assess survey data quality systematically, potentially identify falsification, and summarize measurement error issues in a survey. The approach can also allow reinterview data to be used to estimate the quality of individual observations and of interviewers, in the context of interviewer-administered modes. The approach is very flexible and could be adapted to fit different contexts, such as panel data. Because this approach requires two sets of data and is best expanded with survey paradata, it will be most useful for data producers rather than researchers working with publicly available data.

1 Introduction

Researchers are usually neither the primary collectors of their data nor observers of the data collection process and so may be unsure about the quality of their data. Data quality issues can induce measurement error, which can bias analyses and lead researchers to draw incorrect conclusions. As such, data producers like survey organizations and researchers working with survey firms often want straightforward ways to evaluate and identify such issues.

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; Blasius and Thiessen 2016, 2012; Alwin 2011; Madans et al. 2011). Much of this work, however, looks at individual survey items or at aggregate levels. Furthermore, there is little agreement about how to assess survey data quality (Tourangeau et al. 2021). In this paper, I propose QualMix, a general approach to assessing survey data **qual**ity using **mix**ture models in situations in which researchers have two sets of information, ostensibly from the same respondents. The QualMix model has the potential to be useful in a variety of situations due to its flexibility, but the clearest use case and original inspiration is streamlining the evaluation of measurement error when doing reinterviews.

I first summarize existing approaches to assessing 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: reinterviews. I use a simulation study to show that QualMix can successfully estimate survey and enumerator data quality. Finally, I use the model in a real-world context, estimating survey and enumerator data quality for a large survey carried out in Malawi.

This serves to show that the model gives informative estimates of survey quality in the

context of reinterviews. The simulations demonstrate that the method works well even when only a small proportion (5%) of responses is chosen for reinterviewing. QualMix is not meant to replace other approaches to estimating survey response quality. For example, this method may not be optimal for assessing issues of data quality due to lack of concept or construct validity or for detecting within-item reliability. Nevertheless, QualMix streamlines and makes less arbitrary a data quality assessment step that is already part of the workflow of survey firms, fleshing out the type of data quality assessments that they can provide to the public and to researchers. In addition, it provides respondent-level (and potentially enumerator-level) summary assessments that can be incorporated into future analysis.

2 Reinterviews, Measurement Error, Reliability, and Falsification

Data quality refers to "the amount of error in the data" (Biemer 2011, 5). Researchers, policy makers, and data producers alike place great importance on survey data quality. Data quality issues can lead to misleading point estimates and can also impact statistical analyses (Asher 1974; Duncan and Hill 1985; Bound et al. 2001). High-quality data implies less measurement error — mismatches between respondents' "true" responses and collected responses Groves (1989, *iv*). There are many sources of such errors with surveys including respondent satisficing, mode effects, implementer policies, poorly-thought-out questions, survey data fabrication, and low-quality enumerators. Survey researchers have developed a wide array of tools and strategy for assessing different sources of measurement error. One tool that survey designers often turn to in order to assess the possibility of measurement error in particular, including data falsification, are reinterviews.

2.1 Reinterviews

When doing reinterviews—also called backchecks, recontacts, callbacks or field audits— a subset of respondents is re-interviewed some time after the original data collection. Reinterviews 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 reinterviews in their "Minimum Must Dos" (IPA 2018). The World Bank states that "[b]ack checks [reinterviews] 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 reinterviews 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 reinterviews can be used to "evaluate field work" and "estimate error components in a survey model" (280-281).

This approach is usually used early during enumeration, to help assess the general "health" of the data. It can be used to assess the reliability of survey data and items – does the data match the response the respondent should have selected, given the truth? – in addition to identifying potential falsification of data. Reinterviews, like many data quality assessment tools, rest on the idea of repeated measures — measure the same item in similar ways and check for deviations to see whether the data that have been collected are reliable: whether there is "agreement between two efforts to assess the underlying value using maximally similar measures" (Campbell and Fiske 1959, 83).

Reinterviews present a way to make comparisons between two sets of data that should match. One way to assess data reliability using reinterviews is to calculate the gross difference rate (GDR) for individual survey items. The GDR is defined as $1 - p_a$, where p_a "is the proportion of respondents giving the same answer in both interviews" (Tourangeau 2021, 966). However, while the initial comparisons are done on the level of the respondent, the GDR is an item-level measure and sheds no light on the reliability of the data coming

from any individual respondent.

Although reinterviews can help shed light on data quality, the survey literature primarily discusses re-interviews in the context of finding falsified data, starting with Crespi (1945). Falsified data can cause bias in multivariate analysis (Schnell 1991; Schräpler and Wagner 2005; Ahmed et al. 2014; Finn and Ranchhod 2017; Sarracino and Mikucka 2017). By definition, fabricated data would be unreliable data, as there is no chance of repeat measurement. Catching fabricated data *can*, but does not have to, involve repeated measures.

If data cannot be confirmed through reinterviews, there is a chance that they may have been fabricated. It is also possible that enumerators did a poor job of collecting or entering the collected information.

2.2 Complements and Alternatives to Reinterviews

The strength of reinterviews is the actual recollection of data – if the data collected matches, then data producers can have confidence that measurement error may be limited. Mechanistically, reinterviews are also straightforward to carry out. However, they can add significant costs and presuppose that individuals truthfully recall the answers that they provided during the initial survey attempt (Bredl et al. 2013).

When it comes to data falsification, random re-interviews as a way to identify data falsification by enumerators are not always efficient (Schreiner et al. 1988; Krejsa et al. 1999; Bredl et al. 2013). Random sampling might lead to too many "good" enumerators being chosen for reinterviewing. As such, survey analysts and statisticians have proposed a series of methods for detecting interviewer falsification without using reinterviews, 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). Olbrich et al.

(2023) propose using multilevel models to identify interviewer fraud. Each of these methods shows promise for identifying observations that may be fraudulent. However, it can be difficult to integrate these approaches into a broader data quality assessment.

Reinterviews are also not ideal for assessing the reliability of single survey items or to assess consistency over time. There are more sophisticated data quality assessment procedures that use statistical models to assess survey data reliability. Multitrait-multimethod (MTMM) approaches (Campbell and Fiske 1959) are suitable for cross-sectional data and use "*multiple indicators* measured within the same interview, using different methods or different types of questions for a given concept" (Alwin 2011, 266). Another alternative for longitudinal data is the quasi-Markov simplex (QMS) method (Alwin 2007, 2021). QMS methods use responses at three or more time points to separate out genuine change over time from error. QMS methods produces estimates of reliability via the ratio of estimated true response variance by overall variance at each point in time.

Latent Markov chains (LMC) are also useful for longitudinal data (Langeheine and van de Pol 2002). LMC attempt to separate respondents into latent groups based on their responses to the same categorical variable over a period of time. Similarly, latent class analysis (LCA) presupposes the existence of latent classes (Biemer 2011). Responses to survey items measuring the same concept are manifestations of these latent classes and LCA groups similar responses together to determine the classes. LCA can be used to assess measurement error because it makes it possible to estimate probabilities that respondents belong to a class that indicates a certain category on the desired latent variable but who responded "incorrectly" to an indicator for that latent category. LCA models generally require three or more indicators to be identified, although this can be relaxed with the addition of grouping variables (Kreuter et al. 2008, 724). LCA models are well-suited to identifying flawed survey items (Kreuter et al. 2008), although they can "underestimate error rates" (Yan et al. 2012, 1017). Finally, it is possible to account for measurement error using confirmatory factor analysis (Harrington 2009).

Scholars have also proposed non-retest 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). Cohen and Warner (2021) focus on identifying observations of such low quality that they should be dropped, using 141 different potential indicators of quality and 36 different machine learning models.

2.3 Motivating QualMix

MTMM, LMC, LCA, and GDR focus on assessing the quality of single survey items. While important, it can become too time intensive to scale these approaches to a whole survey. In addition, as some methods – such as MTMM and certain applications of LCA – require multiple questions within the same survey, they are impractical for assessing data quality for a whole survey. This disincentivizes researchers and survey data producers from using them (Madans et al. 2011, 2). Therefore, we need approaches for estimating general data quality in surveys that are more practical to implement on a larger scale. The benefit of reinterviews is that they promise a snapshot of the overall health of a survey and allow data producers to estimate the reliability of entire vectors of respondent information. While they can contribute to the costs of a survey, so can the additional questions or waves of a survey necessary for MTMM, LMC, and LCA.

Up to now, it has not been clear how to analyze the data produced by reinterview comparisons. Partly, this is because quality control at major survey firms is often proprietary and not open to 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 helpful Stata (StataCorp 2019) package *bcstats* (White 2016) to help with analyzing re-interviews, but it does not offer a simple way of summarizing differences between the original data and the re-interview data or generating uncertainty about whether two sets of

information match.

The QualMix model I propose below builds on the kind of repeated measurements on which reinterviews depend: repeated measurements of *sets* of questions. The QualMix model provides an answer to researchers and data producers looking to analyze reinterview data more systematically. QualMix can help identify observations with error, be it due to falsification or data collection mistakes, and can help assess the possible scale of such errors in a survey. As reinterviews are not the only scenario in which data producers will have repeated sets of measurements by design – for example, such a situation exists in panel surveys – I first set up the QualMix generally before applying it concretely to reinterviews.

3 QualMix: Survey Data Quality and Mixture Models

QualMix uses parametric clustering to provide a way to assess the possibility of measurement error in *observations* when two sets of data for certain respondents exist by estimating the probability that two sets of response vectors are the same. The QualMix model can be used to assess overall data quality *and* to detect falsified data when applied to reinterviews and can also generate uncertainty estimates about the quality of individual observations. The method is inspired by the Fellegi and Sunter (FS) probabilistic record linkage (PRL) model (1969; see also Enamorado et al. 2018). The QualMix model is similar to the canonical FS model because both are mixture models, although the FS model is a mixture of a series of independent categoricals, and QualMix is a mixture of multinomials (the categorical distribution is to the multinomial what the Bernoulli is to the binomial). In contrast to PRL, QualMix's aim is to identify potential non-matches where identifiers for two sets of responses *do* exist. In addition, I expand the model by incorporating different information in ways that do not apply for PRL. See Appendix C for a discussion of approximating QualMix using existing PRL implementations.

	Survey Questions			
	Last Name (*String*)	Monthly Income (*Ordered*)	Occupation (*Categorical*)	Age (*Continuous*)
Response Set $\boldsymbol{R_a}$				
$\boldsymbol{r_{a1}}$	Melzer	[$250, $500) (2)	Market Vendor	65
$\boldsymbol{r_{a2}}$	Karlsen	<$250 (1)	Market Vendor	21
Response Set $\boldsymbol{R_b}$				
$\boldsymbol{r_{b1}}$	Beier	<$250 (1)	Tax Collector	57
$\boldsymbol{r_{b2}}$	Karls	<$250 (1)	Business Owner	31
Agreement Vectors				
$\boldsymbol{\gamma}_1$	Complete Disagreement	Complete Disagreement	Complete Disagreement	Similar
$\boldsymbol{\gamma}_2$	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

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 $\boldsymbol{r_{1}}_i$ and $\boldsymbol{r_{2}}_i$ 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 $\boldsymbol{r_{a}}_{k,i}$ and $\boldsymbol{r_{b}}_{k,i}$. If we define information about the agreement or disagreement between $\boldsymbol{r_{a}}_{k,i}$ and $\boldsymbol{r_{b}}_{k,i}$ 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.

Turning the comparison information into L agreement-levels requires pre-specified decision rules, which may be different for different variable types (See App. A for an in-depth description of the decision rules used in the simulations and applications in this

paper). Table 1 presents a concrete hypothetical example of the general approach, with examples of four different variable types and $L = 3$: "Complete Agreement," "Similar," and "Complete Disagreement." Please see Appendix A for an explanation of how decisions were made for Table 1.

3.2 QualMix Model

If data quality issues – i.e. more measurement error caused by data falsification or data collection problems – 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. The key assumption that the high-quality data should see more agreements because high-quality data is more reliable and less likely to have been fabricated. 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 or respondent forgetfulness), nor will agreement-summary vectors for sets of low-quality responses consist of *only* complete disagreements (it is possible that some fabricated information could match the truth, for example, by chance).

Because of the possibility of both incorrect similarities and agreements, 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. 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 $\boldsymbol{r}_{\boldsymbol{a}i}$ and $\boldsymbol{r}_{\boldsymbol{b}i}$. However, such a decision rule is 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? How do we conceptualize our uncertainty about their quality? How certain are we that the data may have been fabricated or not?

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 $\boldsymbol{R_b}$ are high quality or not, and $\boldsymbol{\pi}_m$ is an L length vector of the agreement-level probabilities for distribution q.

The benefit of this approach is that it puts agreement-summary vectors *in context* – how they compare to other agreement summary vectors. If one question is often in disagreement because of an issue with the survey collection platform in the initial round, for example, the model can learn this and does not consider this as very informative about quality.

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 on some of variables of interest, yet still count as high-quality. This model is similar to LCA models, which are also finite mixture models, although whereas for LCA the inputs are outcomes of different categorical items, here the inputs are the agreement-summary vectors.

The observed-data likelihood for this model is

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

If all $\boldsymbol{\nu}_i$ seem to come from the high-quality distribution, then the estimated λ will be close to 1; it will be close to 0 if all seem to come from the low-quality distribution.

We can also calculate the observation-specific probability that observation i represents a high-quality observation using the posterior probability of coming from the high-quality component. If focusing purely on data frabrication, we can also think of this quantity as our estimate of the probability that an observation has *not* been fabricated. Intuitively, this is the amount that the observation i contributes to the likelihood when $Q_i = 1$ divided by observation i's total contribution to the likelihood:

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

Section 3.3 discusses how these posterior probabilities can be used as a measure of the data quality of observation i.

This model is flexible: we can also incorporate respondent-level characteristics or survey metadata into the model (See Appendix B for an expanded discussion of this and other extensions to the general model). For example, in the case of re-interviews, we should incorporate information on interviewers, if the survey mode is interviewer-implemented. Then the extended model becomes:

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

$\text{logit}^{-1}(\beta_0)$ here represents the overall probability of an observation being high-quality, and the intercepts by enumerator (β_e) represent the deviations from this probability. We can, but do not have to, assume a distribution for the β_e's. If we consider them random instead of fixed parameters, a natural choice for their distribution would be the normal distribution; we would then have to estimate their common variance. λ_e represents the

probability that $\boldsymbol{r}_{a_i}$ and $\boldsymbol{r}_{b_i}$ — $\forall\, i \in I_e$ — come from the high-quality distribution, i.e. that observations associated with enumerator e are of high quality. The benefit of this approach is that it allows for the probability that an observation is high quality to vary by enumerator; a limitation is that it requires an interviewer ID variable, which may not always be available in public-facing data sets.

An inherent risk with any unsupervised learning approach is that the model may over-fit 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 D for recommendations on diagnosing issues.

3.3 Quantities of Interest: Assessing Survey Data Quality

The general approach is a test-retest measure. As such, it is best situated to assess questions of reliability and falsification. The posterior probability of being a high-quality/not-fabricated observation ξ_i represents the probability that $\boldsymbol{r_{ai}}$ and $\boldsymbol{r_{bi}}$ are the same — they vary from 0 to 1. ξ_i can thus be interpreted as a measure of how reliable the data associated with unit i is. Without further assumptions, ξ_i cannot tell us whether data were fabricated or not; it only contains information how likely the two sets of responses for respondent i are the same. A low ξ_i can mean fabricated data or data collected with many errors; both scenarios represent low-quality data and indicate considerably measurement error.

We can then designate the *mean* of $\boldsymbol{\xi}$ as a summary of the level of measurement error in the sample.

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

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.

We can use a variety of statistical approaches to estimate the parameters in the model, including the EM algorithm or Bayesian MCMC, which I use in the applications in this paper. Estimates of $\hat{\boldsymbol{\xi}}$ from these approaches represent our confidence in the quality of an individual observation; $\hat{Q_S}$ uses these estimates to estimate data quality for the whole survey ($\hat{\lambda}$ and $\hat{Q_S}$ will be identical when using the EM algorithm, as $\hat{Q_S}$ is the M-Step for λ; in a Bayesian framework, the two are highly correlated but not identical). 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. 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 in either R_a or R_b. Although I do not examine this use directly in the paper, if we use questions that may be different, such as attitudinal questions, which produce less reliable responses than factual questions (Tourangeau 2021, 982), we are assessing both the stability of respondents' opinions or preferences *and* potential data errors. Appendix B.3 discusses how different questions types can change interpretation of the data quality estimates derived from the model.

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 Being High-Quality Observation:

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

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 high-quality observation remains the same. As above, we use the average of the posterior probabilities to assess the data quality of an enumerator's observations – i.e. with how much measurement error an enumerator is associated. 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 recorded incorrectly. If we want to estimate the probability of a high-quality observation by enumerator, we *must* incorporate enumerator information; averaging ξ_i across respondents assigned to an enumerator would result in biased estimates (Vermunt 2010; Bakk et al. 2013). As above, we can estimate these quantities using a variety of statistical approaches.

Q_e is also bounded by 0 and 1. We can use it to assess the quality of the data associated with each enumerator and compare 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 potential level of measurement error associated with an enumerator; whether any error is an enumerator's *responsibility* would require further inquiry.

4 Application: Reinterviews

Reinterviews are useful for identifying data quality issues and produce data like those in Table 1 (Crespi 1945; Schreiner et al. 1988; Forsman and Schreiner 1991; Murphy et al. 2016). In this section, I apply QualMix to re-interview data to derive estimates of survey and enumerator data quality, demonstrating how data producers can use QualMix to assess a survey's implementation. Under certain conditions, we can use these estimates to make

statements about the *original data*. See Appendix E for a discussion of the assumptions necessary for interpreting these quality statements as being about the original data. I first 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 use simulations to assess QualMix's sensitivity to real-world conditions, varying the percent of each enumerator's respondents reinterviewed and the average match proportion (the average of the enumerator-specific proportion of high-quality data, with less measurement error). I check if the model assesses overall survey quality accurately, and whether the model does a "good" job identifying enumerators associated with lower quality data. Survey administrators are often wary of doing more reinterviews due to their costs. Varying the reinterview 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 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:

- Percent of Respondents Reinterviewed (%R) $\in \{0.05, 0.10, 0.15, 0.2\}$
- Average Match Proportion ($\text{logit}^{-1}(\beta_0)$) $\in \{0.8, 0.9, 0.95\}$

This results in twelve parameter combinations. Simulation proceeds by deciding on an "overall" survey quality and how enumerators are better or worse than this overall quality, then generating original data–reinterview data pairs. Note that even for "matching" observation pairs, there was some probability that some of the variable values were incorrect in the reinterview data. Appendix F gives a full description of the simulation

process.

To increase the external validity of the simulation exercise, I create artificial dissimilarities in existing survey data. The survey used for the basis of the simulation is the Endline Market Vendor Survey of the Tax Decentralisation Project, a survey of market vendors in 128 markets in 8 districts in Malawi fielded from October 2018 to January 2019 (Martin et al. 2020). Table 2 shows the variables included in the simulation and their type. See Appendix K for more information on the survey.

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 F.2 for more information

Table 2: Variables Used in Simulation.

Age, education, and sell frequency were three of the variables used during the actual reinterviewing done for this survey. Variables like name and phone number are normally used for reinterviewing but are omitted for privacy reasons in the simulations. The chosen variables all represent values that should not change between the original survey and the reinterview. As such, in this simulation, we can be confident that we are assessing potential

measurement error – mistakes induced by data falsification or mistakes made during enumeration.

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

4.1.2 Assessing the Model's Ability to Assess Survey Data Quality

An important first step is assessing the model's ability to identify two separable distributions. The median of the posterior of the Jensen-Shannon Distance for all possible parameter combinations is between .57 and .64 — it grows as the average match proportion grows. This makes sense; the two distributions become farther apart the fewer errors there actually are. Appendix D.1 includes the figure of all JSD estimates.

In the paper, I show how well the model allows us to estimate survey and enumerator data quality; Appendix G shows that the model also performs well when identifying low-quality observations. I use the estimators for these quantities proposed in Section 3.3.1 and calculate the error. When calculating 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. Figure 1 shows the error in survey quality under different parameter combinations. We can see mean error is around 0 for all parameter combinations. Keeping the overall match probability constant, the mean error is perhaps somewhat lower when 5% of observations are reinterviewed, although the credible intervals are much larger. After that, the error does not change much as the reinterview percentage increases, with credible intervals becoming smaller due to the larger number of observations exposed to the model.

4.1.3 Assessing the Model's Ability to Assess Enumerator Data Quality

Table 3 shows summary statistics for the bias in enumerator data quality across all enumerators in the simulation. The table demonstrates that enumerator data quality bias is clustered around 0. A few enumerators display large negative bias – the largest absolute

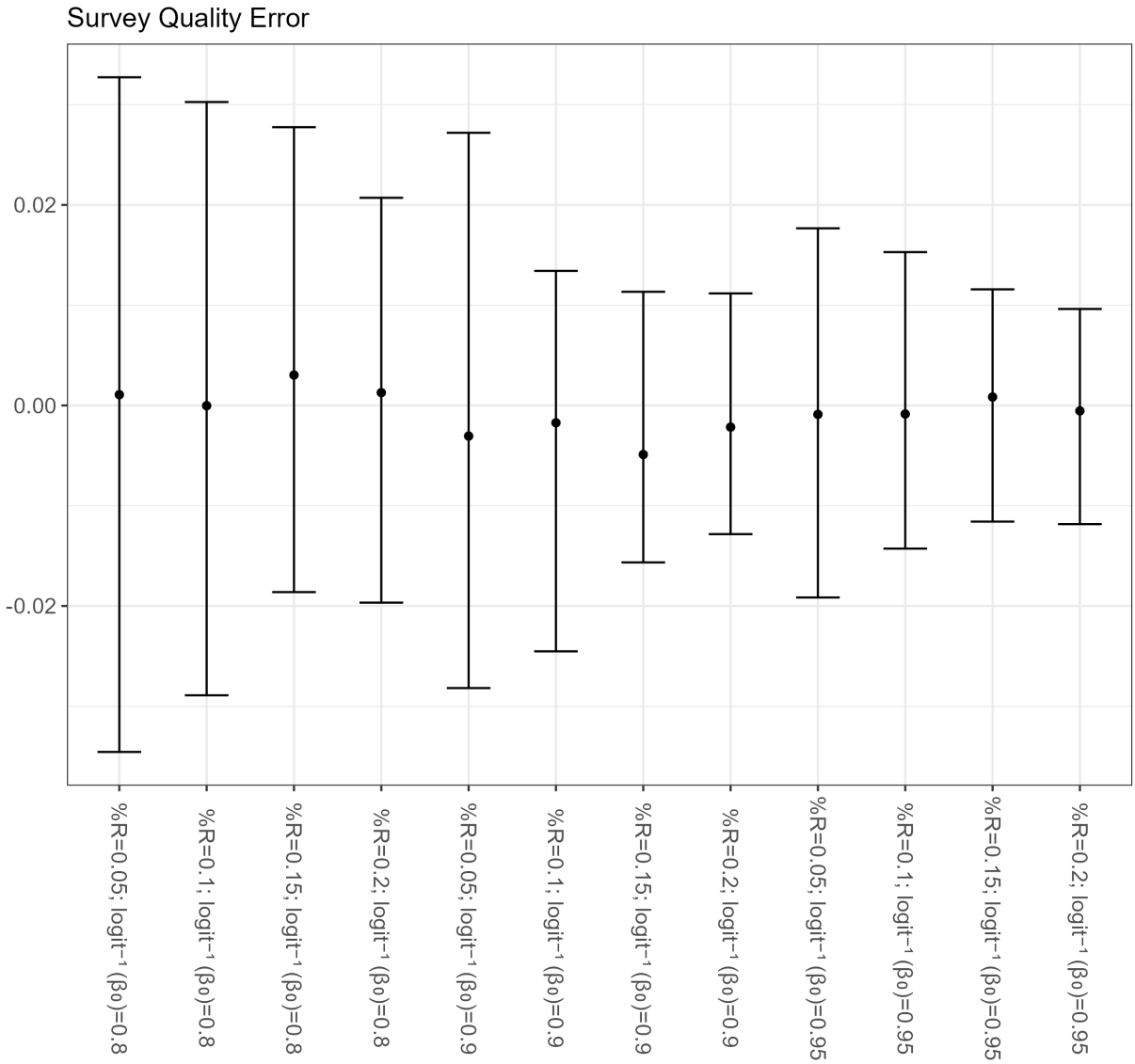


Figure 1: Mean of the posterior of the error (Empirical Bias) of overall survey quality for different parameter combinations with 95% credible intervals.

bias in any parameter combination is 0.108. Figure H1 in Appendix H 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). Their quality estimates may be negatively biased because the probability of a match is generally low for these enumerators and so the probability of a match making it into the reinterviews is lower. Also, these enumerators will make more errors in match observations because of their lack of quality (Appendix F explains why this is the case) — the model struggles to pick this up because they already are "bad" and their high-quality and low-quality observations may be similar. Thus, the model assesses these enumerators as worse than their actual match proportion.

Simulation Parameters	Mean	SD	Min	Max
%R=0.05; $\text{logit}^{-1}(\beta_0)$ =0.8	0.0007	0.0253	-0.0815	0.0458
%R=0.1; $\text{logit}^{-1}(\beta_0)$ =0.8	0.0009	0.0262	-0.0777	0.0333
%R=0.15; $\text{logit}^{-1}(\beta_0)$ =0.8	0.0036	0.0251	-0.0818	0.0289
%R=0.2; $\text{logit}^{-1}(\beta_0)$ =0.8	0.0011	0.0293	-0.0951	0.0258
%R=0.05; $\text{logit}^{-1}(\beta_0)$ =0.9	-0.0023	0.0229	-0.0999	0.0271
%R=0.1; $\text{logit}^{-1}(\beta_0)$ =0.9	-0.0032	0.0212	-0.1052	0.0187
%R=0.15; $\text{logit}^{-1}(\beta_0)$ =0.9	-0.0037	0.0203	-0.0900	0.0140
%R=0.2; $\text{logit}^{-1}(\beta_0)$ =0.9	-0.0011	0.0225	-0.1075	0.0213
%R=0.05; $\text{logit}^{-1}(\beta_0)$ =0.95	-0.0012	0.0163	-0.0785	0.0182
%R=0.1; $\text{logit}^{-1}(\beta_0)$ =0.95	-0.0009	0.0163	-0.0789	0.0229
%R=0.15; $\text{logit}^{-1}(\beta_0)$ =0.95	0.0005	0.0138	-0.0611	0.0201
%R=0.2; $\text{logit}^{-1}(\beta_0)$ =0.95	-0.0004	0.0103	-0.0512	0.0128

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

Table 3 shows that the means of the bias across enumerators is smallest when the average match proportion is 0.95, but highest when the average match proportion is 0.9.

The means of the bias when the average match proportion is 0.95 is similar overall to the means when it is 0.80. Overall, however, the means do not vary considerably. The standard deviation, on the other hand, decreases 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 the percent reinterviewed does not seem to 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 data quality, survey firms would not need to increase the percent reinterviewed.

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 2018 and January 2019. This time, I use the actual reinterview data. Of the 12,370 respondents, 657 (5.3% of the sample) were re-contacted by telephone in November and December 2018. Reinterviews 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 few respondents; one of the forty-five interviewed only one respondent). Figure 2 shows the proportion of each enumerator's respondents chosen for the reinterview process. The enumerator with 100% reinterview rate is omitted from the figure to make it easier to interpret.

Six variables were chosen for all reinterviews:

1. respondent's age
2. respondent's education
3. how often respondent sells at the market
4. what the respondent sells/offers
5. whether the respondent showed the enumerator a receipt
6. respondent's satisfaction with developments in the market.

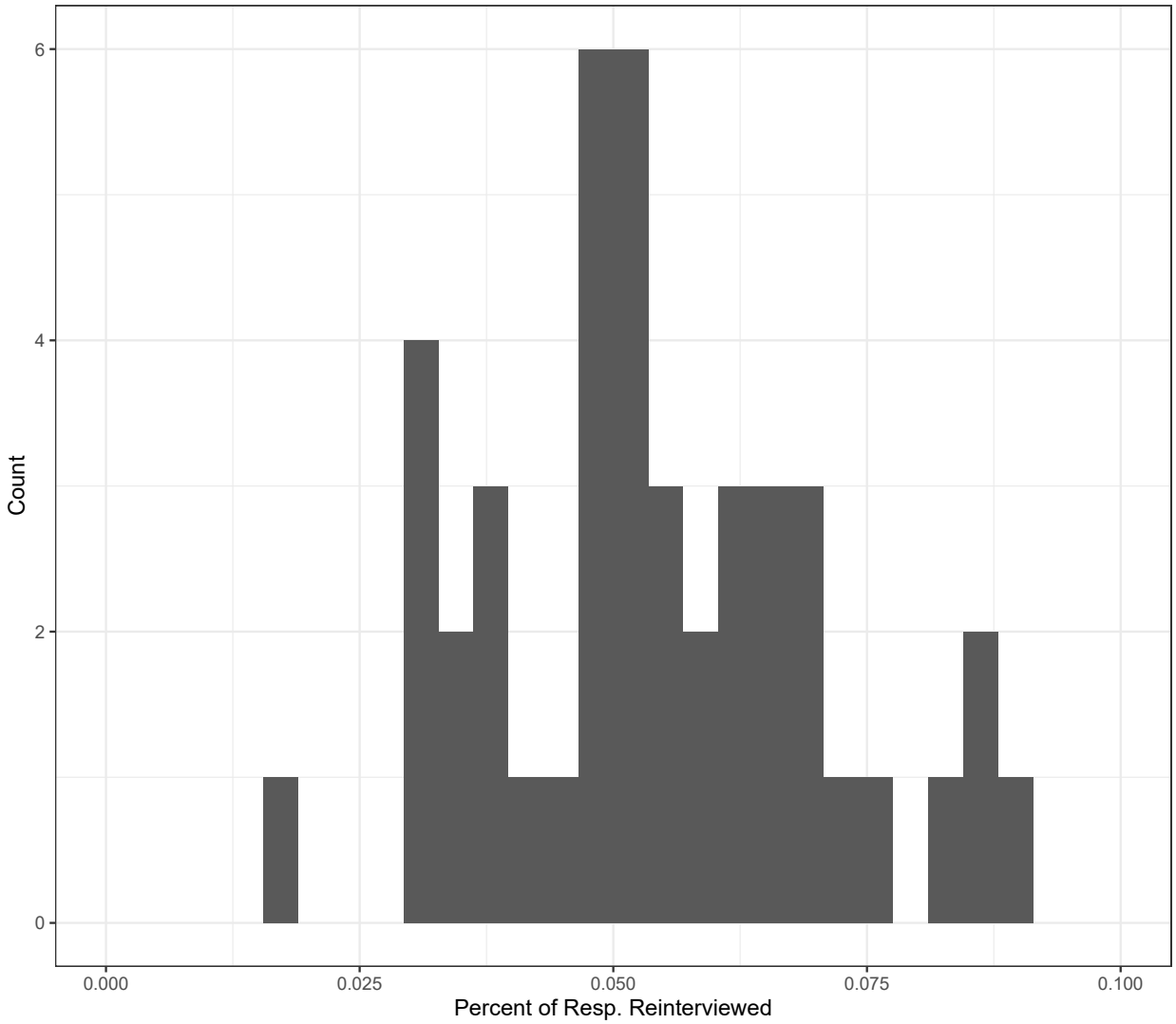


Figure 2: Histogram of Proportion of an Enumerator's Respondents Chosen for Reinterview.

These should not have changed between original enumeration and the reinterview, 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 reinterview pair (see Appendix A for more details on this process, including examples from the application). I categorize NA values as disagreements. I fit the model described in Appendix F.1 using `Stan` to derive estimates of enumerator and survey quality, employing the same priors as in the simulation (Stan Developers and their Assignees 2022).

4.2.1 Results

Figure 3 shows the two estimated multinomials; the model was able to identify distinct distributions over agreement categories. The 95% credible interval for the Jensen-Shannon Distance for these two distributions is [0.704, 0.768].

The low-quality distribution puts almost all of the probability into the "Complete Disagreement" category. Of the 657 reinterview 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 reinterviews, but it is important to take these cases into account – it is possible that the original observations were fabricated. If this were a random process, then we would expect the distribution of such reinterview failures to be uniform among enumerators. Figure 4 clearly shows that this is not the case (see Appendix I for an analysis of dropping these failed reinterviews; it becomes more difficult to detect two distinct distributions). Because these represent 12.9% of reinterviewed respondents, it was straightforward for the model to identify all of these observations as belonging to the same cluster. The high-quality distribution, however, still contains some "Complete Disagreement" values. The 95% credible intervals for the expected value of the categories are shown in Table 4.

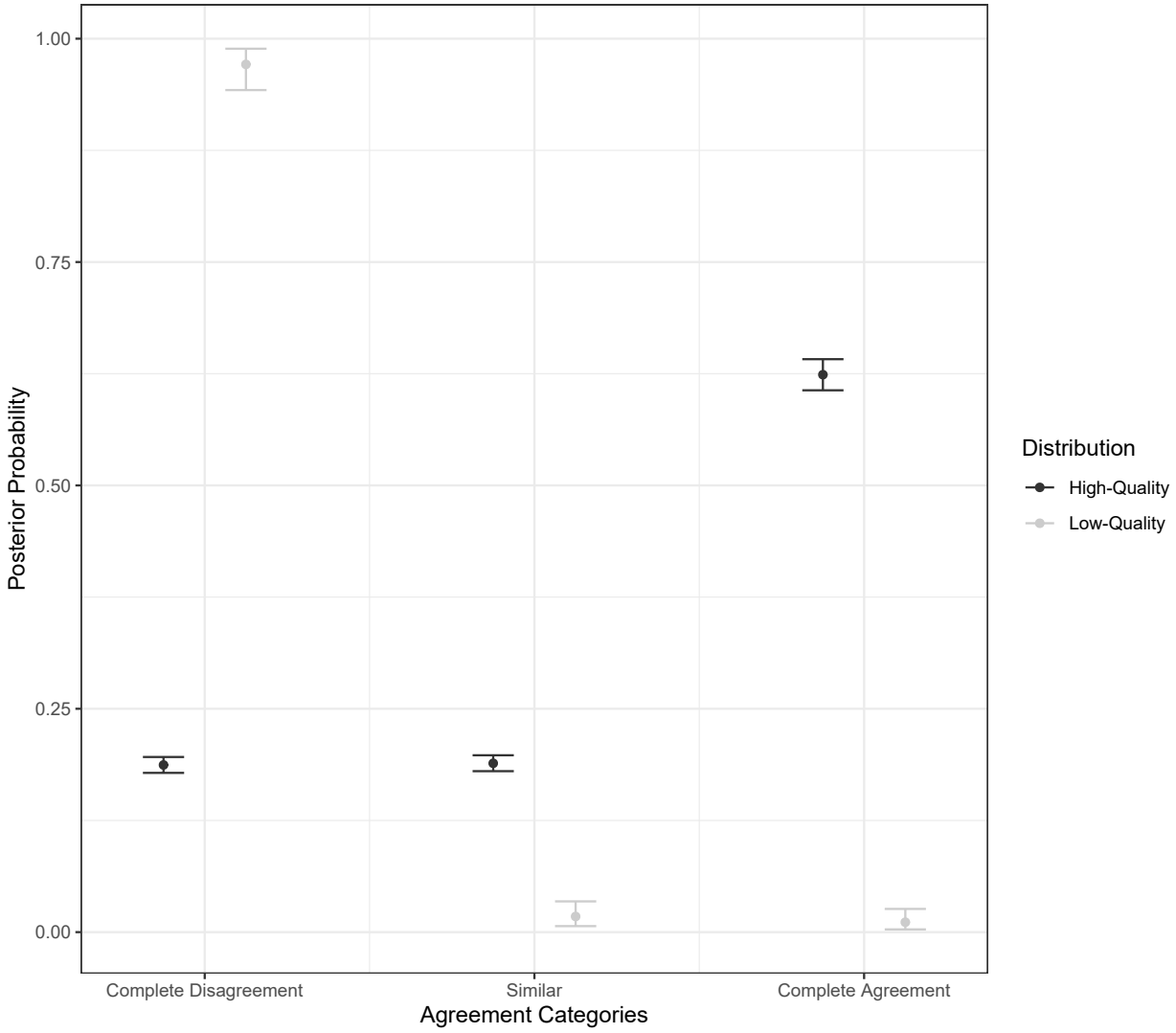


Figure 3: Median of posterior distributions of $\hat{\boldsymbol{\pi}}_0$ (low quality) and $\hat{\boldsymbol{\pi}}_1$ (high quality) along with 95% credible intervals.

Category	2.5%	Median	97.5%
Complete Disagreement	1.070	1.123	1.176
Similar	1.082	1.134	1.187
Agreement	3.640	3.743	3.846

Table 4: 95% Credible Intervals for Expected Values of the High-Quality Distribution

The variable with the largest number of inconsistencies between the reinterview and the original data asks whether respondents showed the enumerator a daily market tax receipt. It is possible that respondents could be suffering from social desirability bias to not say no; more respondents in the reinterview said that they showed a receipt than in the original data. At the same time, vendors perhaps *did* show a receipt, but 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 took additional time. The reinterview itself does not offer evidence one way or another, which demonstrates that implementers 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." Data producers must decide whether they are satisfied with the high-quality and low-quality distributions. Even if they are *not* satisfied with a high-quality distribution, however, QualMix still has utility, as it identifies common patterns in the reinterview data. If a high-quality distribution is unsatisfactory, that is a sign in and of itself that something might have gone wrong during data collection. It is important to 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.

We can derive enumerator data quality estimates using Eq. 1, shown in Figure 4. There is considerable variation in enumerator data quality, with some estimates of data quality being low: ten have data quality estimates of .75 or lower, with two below .5. While these enumerator data quality estimates may be *correlated* with enumerator quality 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 and the reinterview – the implementing organization sent enumerator teams to specific parts of the country; thus, enumerator data

quality may be confounded by regional data collection issues – we cannot say that the enumerators were responsible for flawed data, but we can say that some are associated with poorer data quality.

How can we be sure that these enumerator data quality estimates represent something akin to real-world quality and not just variations in reinterview 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 enumerator quality can impact data quality. Better enumerators may be better at getting a respondent's trust and less likely to rush through a survey and thus more likely to get a respondent to show them a receipt. Using a simple logit regression using all 12,370 observations with enumerator data quality (proxying for enumerator quality) as 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 median posterior predictive probability that an enumerator reported being shown a receipt by .098; increasing enumerator data quality from .75 to 1 increases it by a further .122 (see Appendix J for more information about this validation exercise). As Figure 4 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.

Thus, these enumerator data quality estimates help identify enumerators who may have produced problematic data that may be rife with measurement error. Data producers can 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. Data producers can also see if estimated enumerator data quality is associated with certain survey measures. Data producers do not need to wait until the end of enumeration to apply this model to their data; it is possible to run this model in real time, updating it with data from the field. In such ways, the model systematically facilitates the identification, investigation, and solving of data quality issues using

reinterview comparisons.

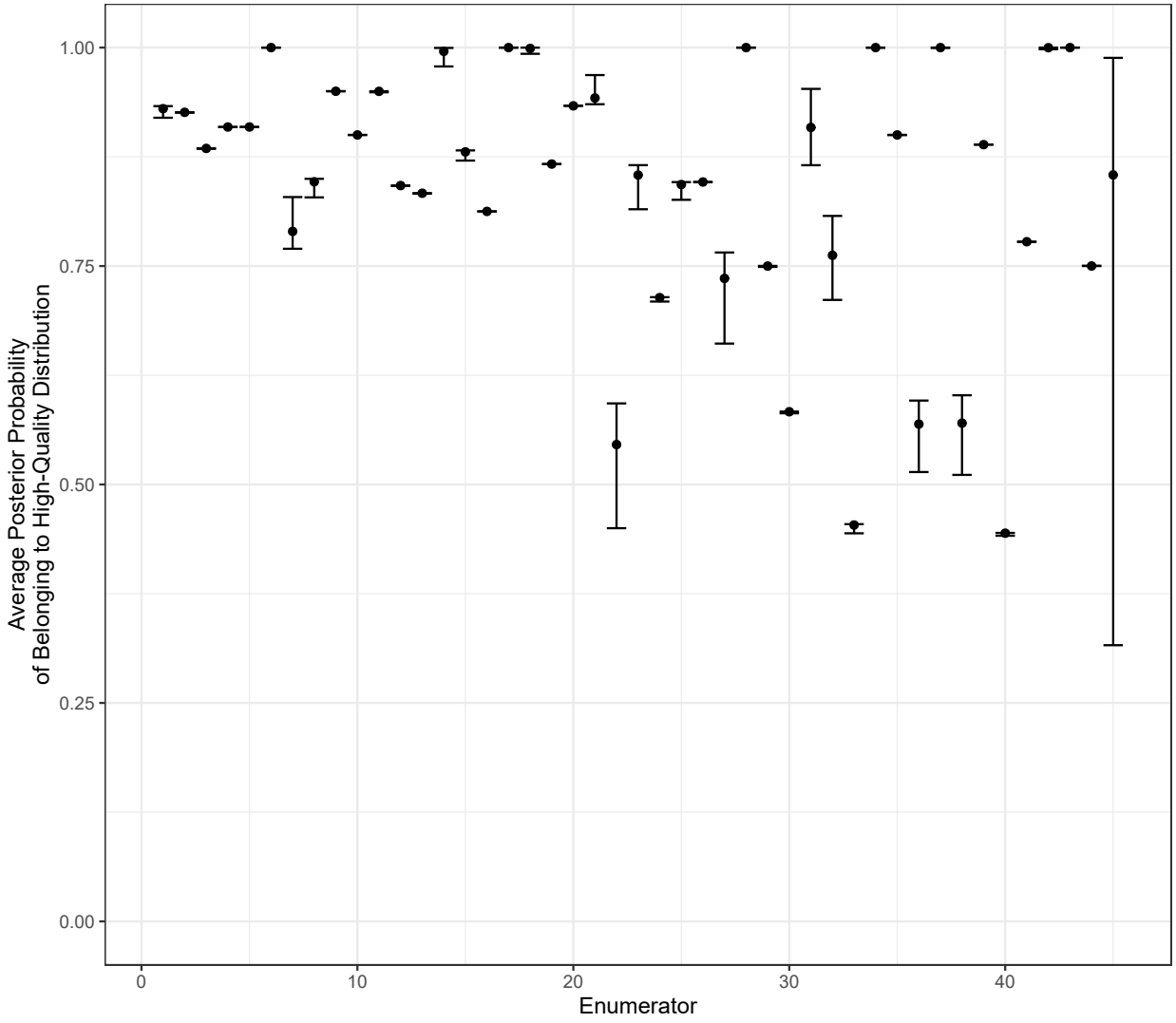


Figure 4: Median of the Posterior of the Average Posterior Probability of Being High Quality for all 45 Enumerators. Error bars show 95% credible intervals.

Data producers interested in applying this model to their own data can find the replication archive for this project on GitHub: `https://github.com/hoellers/QualMix`. The replication archive includes the code used to produce all analysis and figures in this document and the supplementary appendix, as well as an example workflow for using the model to evaluate survey and enumerator data quality.

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 agreements between the two sets of records as data – a data structure commonly found in reinterviews. The simulations demonstrate that the model effectively identifies problematic observations and assesses survey and enumerator data quality. It also shows that data producers can get sufficient estimates of quality by reinterviewing 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 distributions. In order to make this situation less likely, data producers should use more than six variables for reinterviewing. Reinterviewers should ask all reinterview questions, regardless of whether a name matches — it is possible that a *name* is incorrect, but that the other values are correct. Finally, data producers need to be careful about confounding enumerator data quality and region of enumeration, regardless of quality assessment used.

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. The model represents a straightforward way for data producers to synthesize what reinterviews are saying about the quality of the data, which can form part of a data quality statement – in addition to other data quality assessments – for researchers and the general public.

Data producers are not limited to only evaluating survey reinterviews with this model; other applicable scenarios include estimating the uncertainty that the correct respondents have been recontacted in a panel survey, for example, or assessing respondent-level item stability and response consistency by asking similar items in multiple ways throughout a

survey, akin to MTMM or LCA. This latter approach may be useful for surveys that cannot afford re-interviews, although it does imply additional respondent burden and can extend surveys.

With this model, data producers can estimate the level of measurement error in a survey, as well as the data quality associated with enumerators implementing the survey. It does have some limitations, however, and data producers need to be aware of the assumptions that must be met in order to interpret parameter estimates in this way. The model requires two sets of data and is best expanded with survey paradata; it may not be as useful for researchers working with publicly released data that they did not help collect. This method also presupposes that there is no non-random significant change in the re-interview data, either due to change in mode, respondent forgetfulness, and time between original and re-interview surveys. Data producers also need to consider carefully how to establish the agreement-summary vectors on which the model is fit – including which survey items to use – as this can have large impacts on the quality estimates.

The model has uses beyond estimating survey data quality. In particular, it offers avenues for dealing with measurement quality issues once they have been discovered. Instead of dropping data, wasting resources, or ignoring data quality concerns because of resource constraints, researchers could use estimates from QualMix to weight observations by their posterior probability that they are high quality, upweighting observations about whose quality we are more certain, and downweighting those about whose quality we are less certain. This has the potential to reduce bias in parameter estimates, ultimately resulting in more accurate analyses.

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.

— (2021), "Developing Reliable Measures: An Approach to Evaluating the Quality of Survey Measurement Using Longitudinal Designs," in *Measurement Error in Longitudinal Data*, eds. Cernat, A. and Sakshaug, J. W., Oxford: Oxford University Press, pp. 113–154.

Asher, H. B. (1974), "Consequences of Measurement Error in Survey Data," *American Journal of Political Science*, 18, 469–485.

Bakk, Z., Tekle, F. B., and Vermunt, J. K. (2013), "Estimating the Association Between Latent Class Membership and External Variables Using Bias-Adjusted Three-Step Approaches," *Sociological Methodology*, 43, 272–311.

Biemer, P. P. (2011), *Latent Class Analysis of Survey Error*, Hoboken, NJ: John Wiley & Sons, Inc.

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.

— (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.

Bound, J., Brown, C., and Mathiowetz, N. (2001), "Measurement Error in Survey Data," in *Handbook of Econometrics, Volume 5*, eds. Heckman, J. J. and Leamer, E., Amsterdam: Elsevier Science B.V., pp. 3707–3843.

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

Campbell, D. T. and Fiske, D. W. (1959), "Convergent and Discriminant Validation By the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56, 81–104.

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.

Duncan, G. J. and Hill, D. H. (1985), "An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data," *Journal of Labor Economics*, 3, 508–532.

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.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York: Wiley.

Harrington, D. (2009), "Confirmatory Factor Analysis," in *Handbook of Structural Equation Modeling*, ed. Hoyle, R. H., New York: Oxford University Press, 2nd ed., pp. 261–273.

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.

Kreuter, F., Yan, T., and Tourangeau, R. (2008), "Good item or bad–can latent class analysis tell?: the utility of latent class analysis for the evaluation of survey questions," *Journal of the Roay Statistical Society: Series A*, 171, 723–738.

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.

Langeheine, R. and van de Pol, F. (2002), "Latent Markov Chains," in *Applied Latent Class Analysis*, eds. Hagenaars, J. A. and McCutcheon, A. L., New York, NY: Cambridge University Press, pp. 304–341.

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.

Martin, L., Seim, B., and Hoellerbauer, S. (2020), *DRG learning, evaluation, and research (DRG-LER) activity : impact evaluation of USAID/Malawi local government accountability and performance (LGAP) activity : final report.*

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].

Olbrich, L., Kosyakova, Y., Sakshaug, J. W., and Schanhäuser, S. (2023), "Detecting Interviewer Fraud Using Multilevel Models," *Journal of Survey Statistics and Methodology*, 00, 1–22.

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 (2022), *CmdStanR*, r Package Version 0.5.2 (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.

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.

Vermunt, J. K. (2010), "Latent Class Modeling with Covariates: Two Improved Three-Step Approaches," *Political Analysis*, 18, 450–469.

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.

Yan, T., Kreuter, F., and Tourangeau, R. (2012), "Latent class analysis of response inconsistencies across modes of data collection," *Social Science Research*, 41, 1017–1027.

Supplementary Appendix for "A Mixture Model Approach to Assessing Measurement Error in Surveys Using Reinterviews"

Please note: All code and data for analysis in the main paper and in the appendix can be found in this project's GitHub repo: `https://github.com/hoellers/QualMix`.

A Forming Agreement-Summary Vectors

In this appendix, I describe decisions rules for different types of variables. These decisions rules can involve somewhat arbitrary decisions, such as cut-offs for different comparison values. As with any form of analysis, data producers assessing data quality using these tools should be up front about the decisions made during the creation of agreement vectors. By establishing and applying the decisions rules, it is possible to automate the creation of agreement vectors; this does not have to be a manual process.

The paragraphs below explain the decisions rules used in this analysis, using Table 1 as an example. These decision rules are also the ones used in the simulation and in the real-world application presented in the paper. An example of the latter can be found in Appendix A.1.

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" (Enamorado et al. 2018). 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." This cut-off can be changed – if researchers want to only consider exactly identical strings as being in "complete agreement," then they can set the cuff-off to 1. If the strings to be compared are several sentences long instead of names or words, researchers and data producers could turn to

natural-language-process models that compare the similarity of two sets of text.

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.[1]. Similar to the string distance, the higher the cut-offs, the more similar two numbers or ordered factor levels have to be for them to count as "similar" or "complete agreement."

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

[1] An alternative is to use the percent max range measure only for ordered variables with more than some number l of levels. A fitting l value would be 8, as when there are 8 or fewer possible values, a difference of 1 level would still be considered a total disagreement, using a .88 cutoff. For ordered variables with 8 or fewer levels, one could consider a difference of one level as "similar," and differences of greater than one level as "complete disagreement." In the simulation and real-world application, this is the approach used for ordered variables with 8 or fewer categories

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 using their subject-matter knowledge, if possible, and consider levels within such groupings as similar. In the example in Table 1, 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 86, 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."[2]

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.

A.1 Example from Real-World Application

Table A1 shows three examples of original (OG) and reinterview (RE) data points for three respondents used for the real-world application (Section 4.2). Underneath each variable name is the variable type used for the creation of the agreement-summary vectors. Although type of product/service was encoded as a categorical variable in the data (with fifty-four possible categories), I used the category label as a string for creating agreement vectors because some categories that were similar had very similar labels. An alternative would have been to go through all categories and decide which other categories would be similar or complete agreements for each of the fifty-four categories.

Table A2 shows the comparison values (comparators used can be found in italics in the

[2]An alternative would be to treat numbers as strings use the string distance or something similar that more accurately captures the possibility of data entry errors. In such a context, 21 and 31 could be treated as "similar" because the first digit differs by one – it is possible that whoever entered the data mistyped.

Table A1: Three Example Original and Reinterview Survey Pairs

Resp	**Survey**	Age (*Continuous*)	Education (*Ordered*)	How Often Sells at Market (*Ordered*)	Type of Product/Service (*String*)	Receipt Shown (*Categorical*)	Satisfaction with Developments (*Ordered*)
18	OG	21	Standard 8	1-3 days a week	Retail - Cooked food and snacks	No Recipt	Very Dissatisfied
18	RE	NA	NA	NA	NA	NA	NA
24	OG	38	Standard 7	1-3 days a week	Retail - Clothes/shoes	No Recipt	Somewhat Satisfied
24	RE	35	Standard 8	4-6 days a week	Retail - Clothes/shoes	No Receipt	Somewhat Dissatisfied
44	OG	31	Form 1	1-3 days a week	Retail - Clothes/shoes	No Receipt	Very Dissatisfied
44	RE	31	Standard 8	4-6 days a week	Retail - Plastics	Receipt Available	Somewhat Satisfied

column headers) for these three observations. The difference in levels comparator is the one mentioned above – for an ordered variables, it calculates the difference between the levels of the two observations in each pair. I use this comparator for sell frequency and satisfaction with development, as these have 8 and 4 levels, respectively. I use percent max range for education as this variable has 19 levels. For all variables, I treat "Refused to Answer" responses as missing data. The exact comparator produces only two values: different or same.

Table A2: Comparison Values for Three Example Pairs

Resp	Age *Percent Max Range*	Education *Percent Max Range*	How Often Sells at Market *Difference in Levels*	Type of Product/Service *Jaro-Winkler String Distance*	Receipt Shown *Exact*	Satisfaction with Developments *Difference in Levels*
18	NA	NA	NA	NA	NA	NA
24	0.956	0.923	1	1.000	Same	1
44	1.000	0.923	1	0.856	Different	2

Table A3 depicts the agreement vectors produced from the comparisons in Tables A1 and A2. Table A4 shows the resulting agreement-summary vectors; these are the inputs to the QualMix model.

Table A3: Agreement Vectors for Three Example Pairs

Resp	Age	Education	How Often Sells at Market	Type of Product/Service	Receipt Shown	Satisfaction with Developments
18	Complete Disagreement	Complete Disagreement	Complete Disagreement	Complete Disagreement	Complete Disagreement	Complete Disagreement
24	Complete Agreement	Similar	Similar	Complete Agreement	Complete Agreement	Similar
44	Complete Agreement	Similar	Similar	Complete Disagreement	Complete Disagreement	Complete Disagreement

Table A4: Agreement-Summary Vectors for Three Example Pairs

Resp	Complete Disagreement	Similar	Complete Agreement
18	6	0	0
24	0	3	3
44	3	2	1

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 reinterviews, 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 data producers 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.

In the case of reinterviews, 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 such 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-reinterview 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 reinterviewing. However, this may not be the case. The model can be adapted to have random intercepts by enumerators and reinterviewers. This would mean that $\lambda_{eb} = \text{logit}^{-1}(\beta_0 + \beta_e + \beta_b)$. This would allow data producers to see the impact of both field enumerators and reinterviewing enumerators on data quality.

This example should make it clear that data producers could include respondent-level variables in the regression on the latent Q_i as well. 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.[3] This would also be useful if data producers 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

The K questions chosen for comparison can impact how we interpret quality estimates from the QualMix model. Using the reinterviewing 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.).[4] 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

[3]Without incorporating additional enumerator or respondent characteristics, this model would then be identical to the Fellegi-Sunter probabilistic record linkage model, although the set-up would be different as the goal is not to create agreement vectors for all combinations of observations in the two response sets.

[4]Tourangeau (2021) similarly separates questions into demographic, behavioral, and attitudinal questions, with decreasing *reliability* expectations for each of them (982).

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: it would still serve to assess data quality.

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 might often fall into this category. Ideally, 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?

However, we must be careful with these kinds of questions. Because there is no way to separate within-respondent variability due to question type and variability due to error (and these may even be related), for question types where there may be expected variability between response sets, we are combining assessments of stability and data quality. Nevertheless, even for these questions the core assumption that high-quality observations imply more agreements could still hold, even if there are *fewer* agreements overall. Thus, in general, within question types, the proposed data quality estimates still assesses data quality from the perspective of reliability — are two *sets* of measures the same. The interpretation of the ξ_i's and λ will still be the posterior probability of belonging to the high-quality (more agreements) distribution, and the probability that a randomly selected observation is part of the high-quality distribution. However, the $\boldsymbol{\pi}_1$ and $\boldsymbol{pi}_0$ parameters – in other words the high-quality and low-quality distributions – may look very different. This can mean that estimates of data quality from this model will look different if the

model is fit using agreement-summary vectors created using questions of different types.

All three types can detect falsification of data, if it exists, under the assumption that falsified data will result in more disagreements than non-falsified data. However, they lead, in the absence of gross falsification, to different assessments of survey data 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 to. 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 also include questions of all three kinds in one model. Once we do, however, we combine the various sources of variability 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 survey data quality, combining the three different types of data quality issues.

B.4 Identifying Falsifying Enumerators

The purpose of the QualMix model is not exclusively to identify falsifying enumerators – in fact, it cannot not adjudicate the source of poor data quality: whether an observation is fabricated or just poorly collected. It does allow researchers and survey practitioners to identify falsifying enumerators and estimate the probability that an enumerator is falsifying observations, under the assumption that the low-quality distribution represents fabricated data. 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 reinterviews 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 reinterview data to assess general survey and enumerator data 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.

B.5 Using Priors to Encode Expected Quality Distributions

Users of QualMix can encode expected desired quality distributions using priors on $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_0$. For example, if they expect a certain number of agreements in the high-quality distribution, they could put a prior on $\boldsymbol{\pi}_1$ that reflects a correspondingly high probability of seeing complete-agreement values. The most extreme case would be to completely specify the high-quality and low-quality distributions - the model would then try to do its best to place observations in the two pre-specified distributions. This would then no longer be clustering analysis, however.

C Approximating QualMix with Existing PRL Methods

It is possible to approximate one version of QualMix – the one presented in Appendix B.2 *without* incorporating additional characteristics such as enumerator ID – with existing probabilistic record linkage (PRL) methods, such as the `fastLink` package in R, the Python Record Linkage Toolkit library in Python, and `dtalink` in Stata.

Yet, because PRL solves a fundamentally different problem, users of these methods have to be cautious. While in QualMix, λ represents the probability that a randomly selected observation is high quality (reliable because it represents a match between R_a and R_b), in PRL models without modification λ is the probability that a randomly selected observation in R_a will have a match in R_b. In order to get these approaches to approximate this particular instance of QualMix, we would have to *block* – in the PRL terminology – on respondent ID, as this would force the model to compare only within respondent ID. The interpretation of λ would then be roughly comparable to the QualMix model. Estimated $\hat{\xi}_i$'s could then be used to assess survey quality as proposed in Section 3.3.

None of the PRL approaches mentioned here, however, allow for the incorporation of respondent or enumerator characteristics in a regression on λ, which makes them somewhat ill-suited for the types of analyses data producers would like to do. In addition, because the goal of PRL – identifying observations that may match in two datasets when we are not sure whether there are matches – is different from that of the QualMix model – separating data that *should* match into sets that probably do and probably do not – the vignettes and documentation for PRL may not be very helpful for individuals seeking to use QualMix.

D 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.[5]. This would result in estimated posterior probabilities of high quality "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 posterior probabilities of high quality. 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).

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

Figure D1 shows the Jensen-Shannon Distance (JSD). We can see that for all parameter combinations it is higher than 0.5. 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.[6]

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

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

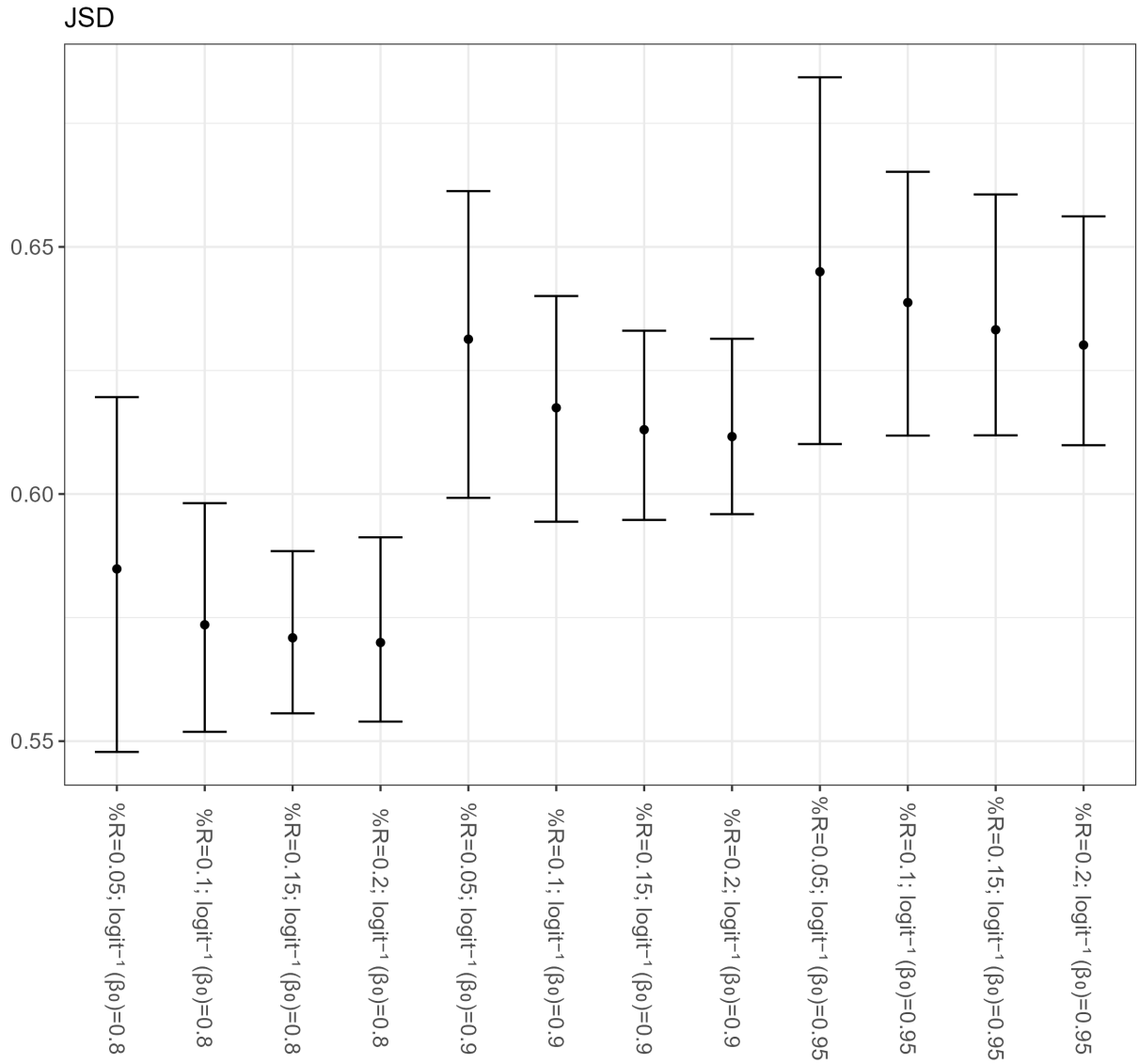


Figure D1: Median of the posterior of the Jensen-Shannon Distance for different parameter combinations with 95% credible intervals.

E Assumptions for Applying Model

In general, for this modeling approach to be valid, we assume that a higher number of agreements between R_a and R_b, on average, indicate higher quality data. In order to interpret model estimates as being about falsification, we additionally assume that falsified data will be less likely to match between R_a and R_b.

E.1 Assumptions for Applying Model to Reinterviews

In the context of reinterviews, 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 generally correct. If the reinterview process creates artificial non-matches — if, for example, reinterviewing relies on phone numbers, and the respondent provided an incorrect or temporary phone number — then the survey quality and enumerator data quality estimates may be flawed.

    - It is not totally necessary for the re-interview data to be the gold standard; there can be random errors in the re-interview data as well because in expectation this would affect the high and low-quality components equally. This assumption is not met, however, if there are systematic errors in the re-interview data.
    - On a similar note, re-interviews can often be done in a different mode than the original survey. This was the case in the empirical demonstration in the paper, where the original survey was face-to-face but the reinterview was done via telephone. A strength of QualMix model is that as long as any mode effects are consistent across the reinterview sample and not too drastic, it will still group the same observations into the high-quality distribution.
    - Survey companies using reinterviews to assess survey data quality already

implicitly make this assumption and also often use different modes when doing re-interviewing and thus already implicitly accept the risk for mode-induced errors.

- If this assumption is not met, then the quality estimate becomes, like with the general model, a statement about both R_a and R_b.

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. Reinterviewing is performed randomly. This is important because when it comes to reinterviews, we will have two sets of responses for only a subset of respondents. If the reinterview sample is non-random, the quality assessments derived from this method will be biased.

F 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 reinterview case. Steps involved only in the reinterview simulation are marked "[reinterview]".

1. Start with an existing survey data set ($\boldsymbol{R_b}$). This will be the "correct" set.
2. Choose a baseline *proportion of matches*: 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 <u>match</u> 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).[7]
- 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.

[7]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 data quality because this will directly impact ν_i's.

For unordered factor variables, do nothing.

- For numbers, with equal probability: transpose two digits at random, insert 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. [reinterview] Sample a portion of observations to reflect reinterviewing (*reinterview 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.[8] This results in 9,973 total observations. For each set of parameters, I then simulate fifty original data–reinterview data pairs. For each simulated original data–reinterview data pair, I then calculate agreement-summary vectors for all original data–reinterview data observation pairs in the manner described in Section 3.3 and Appendix A.

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

F.1 Simulation Model Specification

I fit the QualMix model to the agreement-summary vectors using `Stan`'s `R` interface `cmdstanr` (Stan Developers and their Assignees 2022). I use following model specifications:

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

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).[9]

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.

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

F.2 Simulating Effects of Measurement Error

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

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

$$y_i \sim \mathcal{N}(\mu_i, 5)$$

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.[10] 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.

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

G Identifying Low-Quality Observations

This appendix assesses how well the model is able to identify low-quality (and potentially falsified) observations. Figure G1 shows the Area Under the Receiver Operating Characteristic Curve for all parameter combinations, calculated out of sample (on the observations not selected for the reinterview 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 0.00509. In general, performance improves the higher the average match proportion. Model performance also somewhat improves as the reinterview 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(in other words, considering a high-quality observation as low quality, and considering a high-quality observation as low quality). Figure G2 shows the false negative and false discovery rate (also known as the false positive rate), considering an observation a match if its posterior probability of being high quality 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 reinterview 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 data producers, as more reinterviews require more resources.

In summary, the model performs well out of sample when it comes to identifying

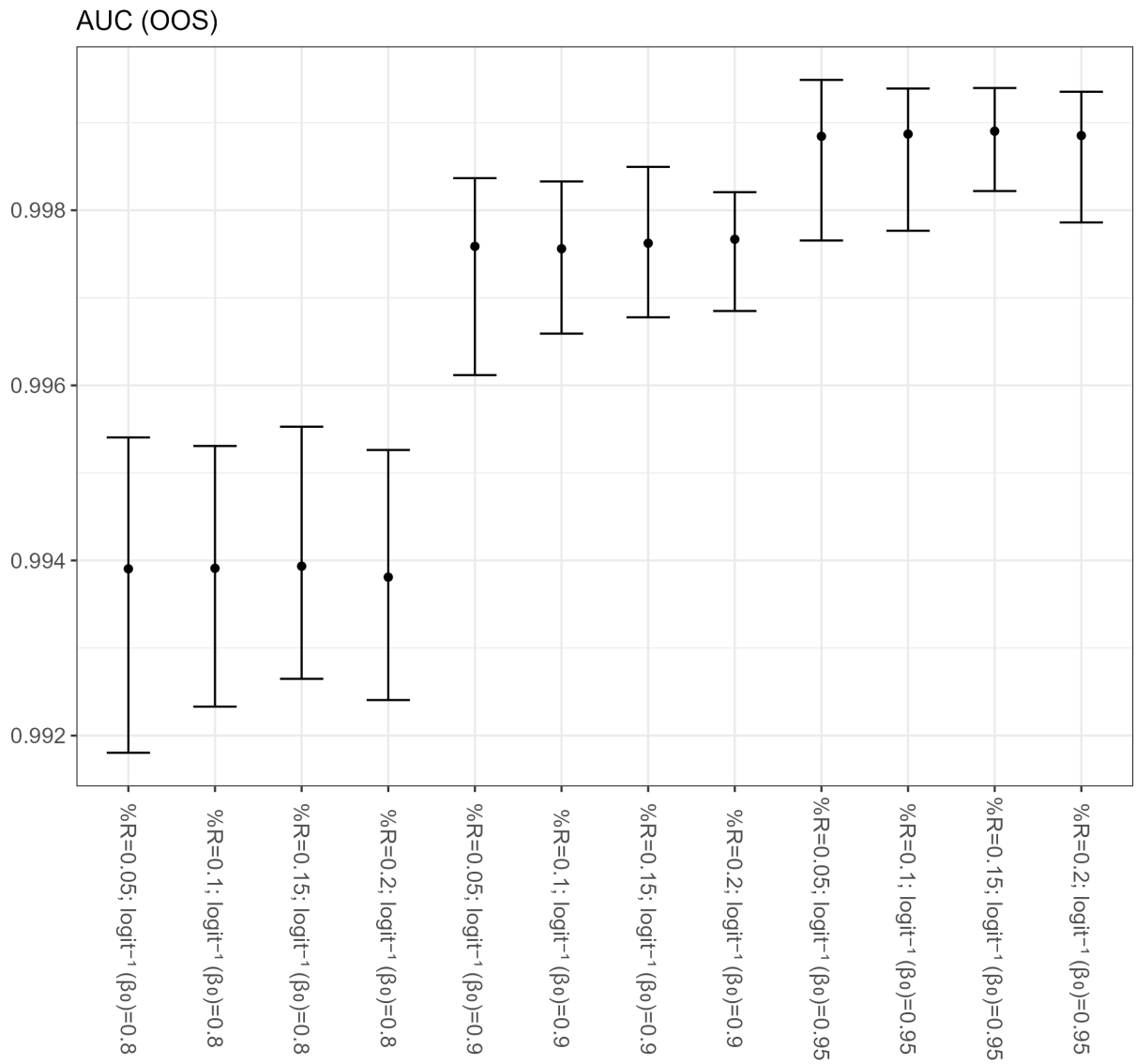


Figure G1: Median of the posterior of the Area Under the Receiver Operating Characteristic Curve for different parameter combinations with 95% credible intervals.

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 reinterview issues.

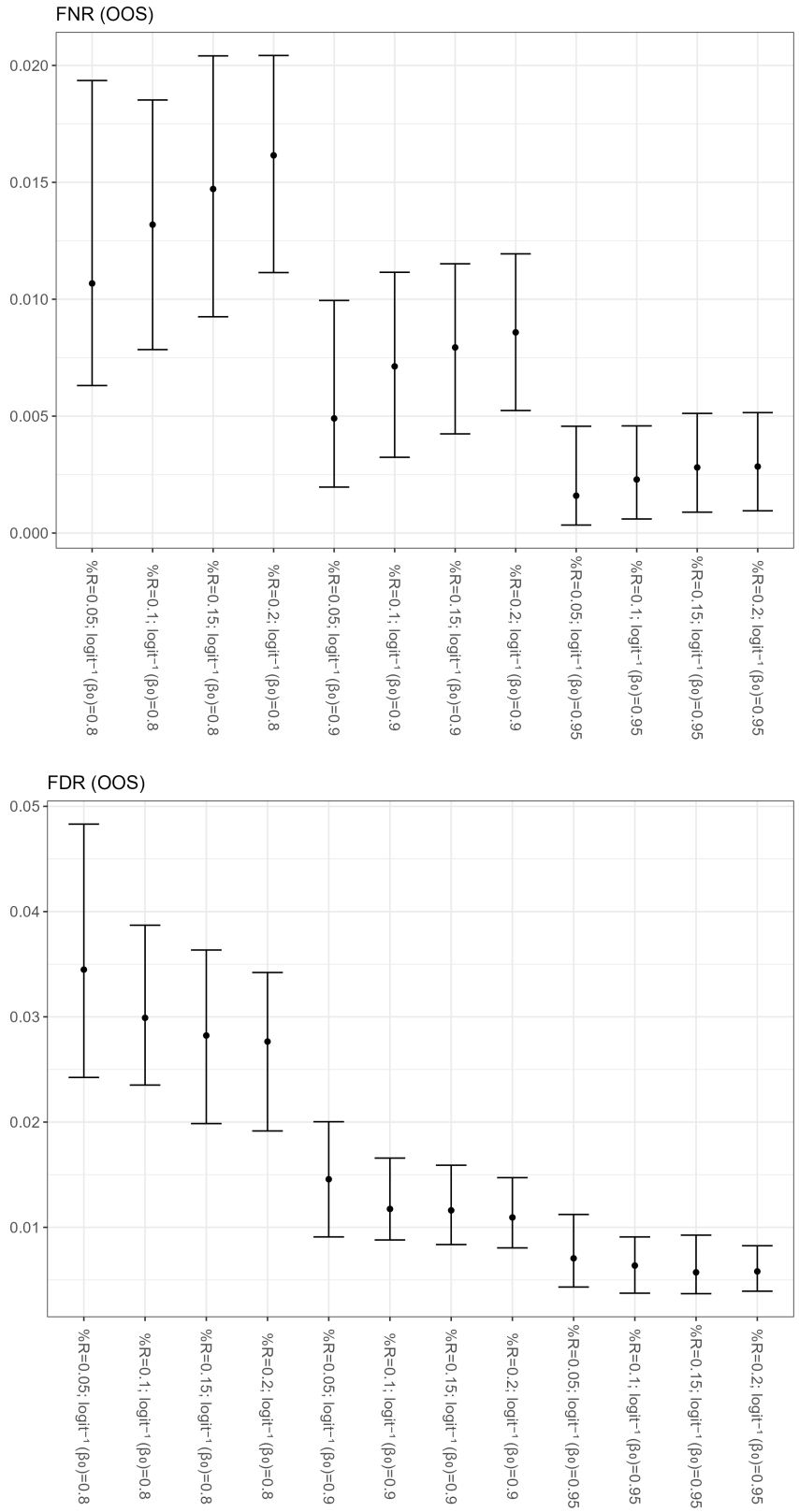


Figure G2: Median of the posterior of the False Negative Rate and the False Discovery Rate for different parameter combinations with 95% credible intervals.

H Enumerator Data Quality Plot

This plot shows the bias in enumerator data 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. The three enumerators with the most consistent negative biases over simulations are shown in the figure; they were also the three enumerators with the lowest probability of having a high-quality observation. Enumerator 1 had a 23.3% probability, enumerator 25 a 34.3% probability, and enumerator 6 a 43.0% probability.

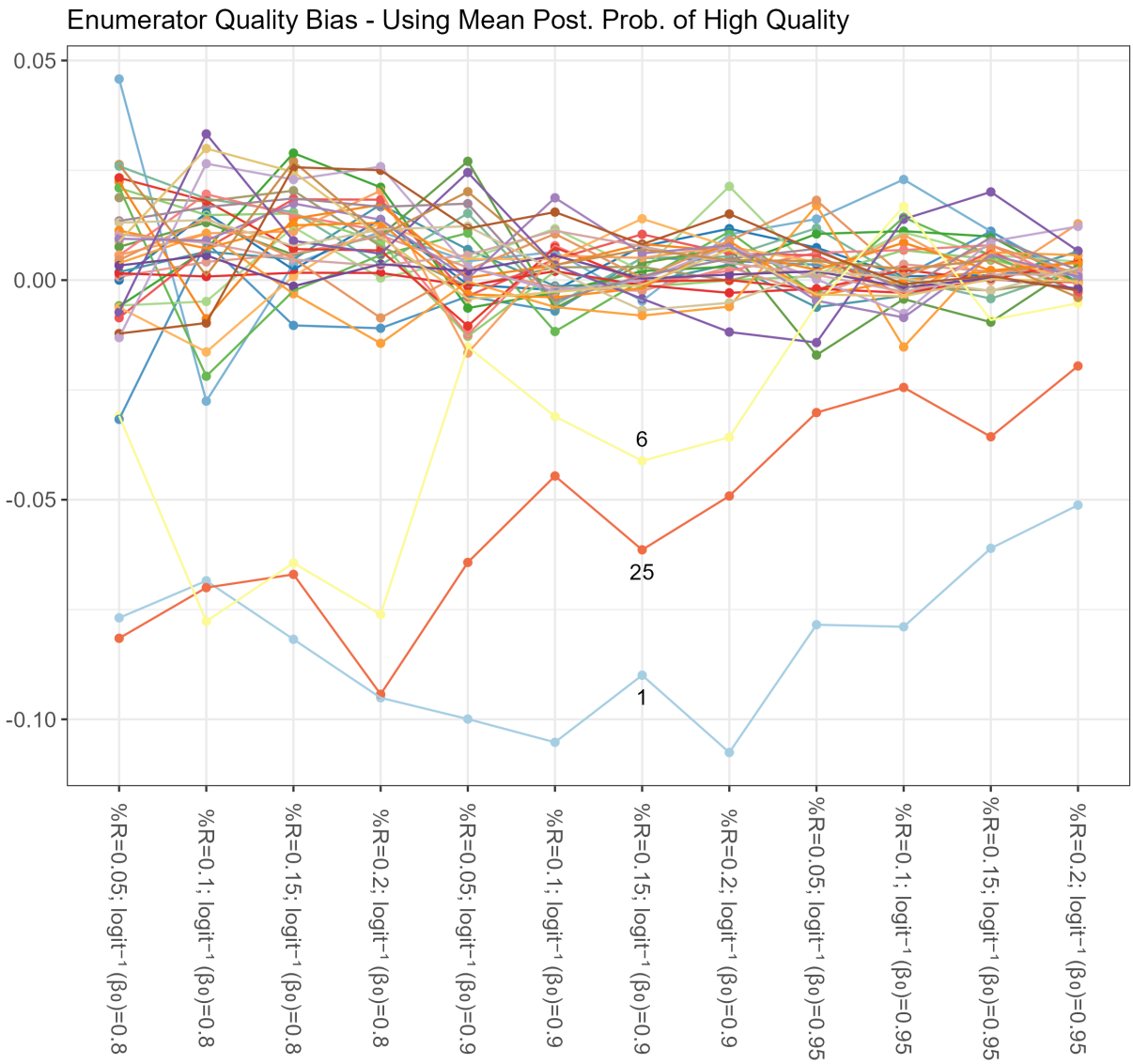


Figure H1: Bias of enumerator data quality for different parameter combinations, where enumerator quality is estimated as the mean posterior probability of high quality of all respondents originally interviewed by that enumerator and selected for the reinterview. Text labels enumerator ID for three enumerators with most biased estimates of quality.

I Results Dropping Failed Reinterviews

There were some idiosyncrasies with the reinterviewing process for this survey. Reinterviews 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 reinterviewing occured. Interviewers were directed to ask no further reinterview 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 reinterview was completed. When we removed "failed" reinterviews, we can see in Figure I1 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 I2 shows. The 95% credible interval for the Jensen-Shannon Distance for this application is [0.355, 480]. The large interval comes from the uncertainty around the non-match distribution.

Figure I3 shows updated enumerator data quality estimates. Unfortunately there is considerable uncertainty about enumerator data 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 [0.910, 0.957]. It is important to note that the incomplete reinterviews 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.

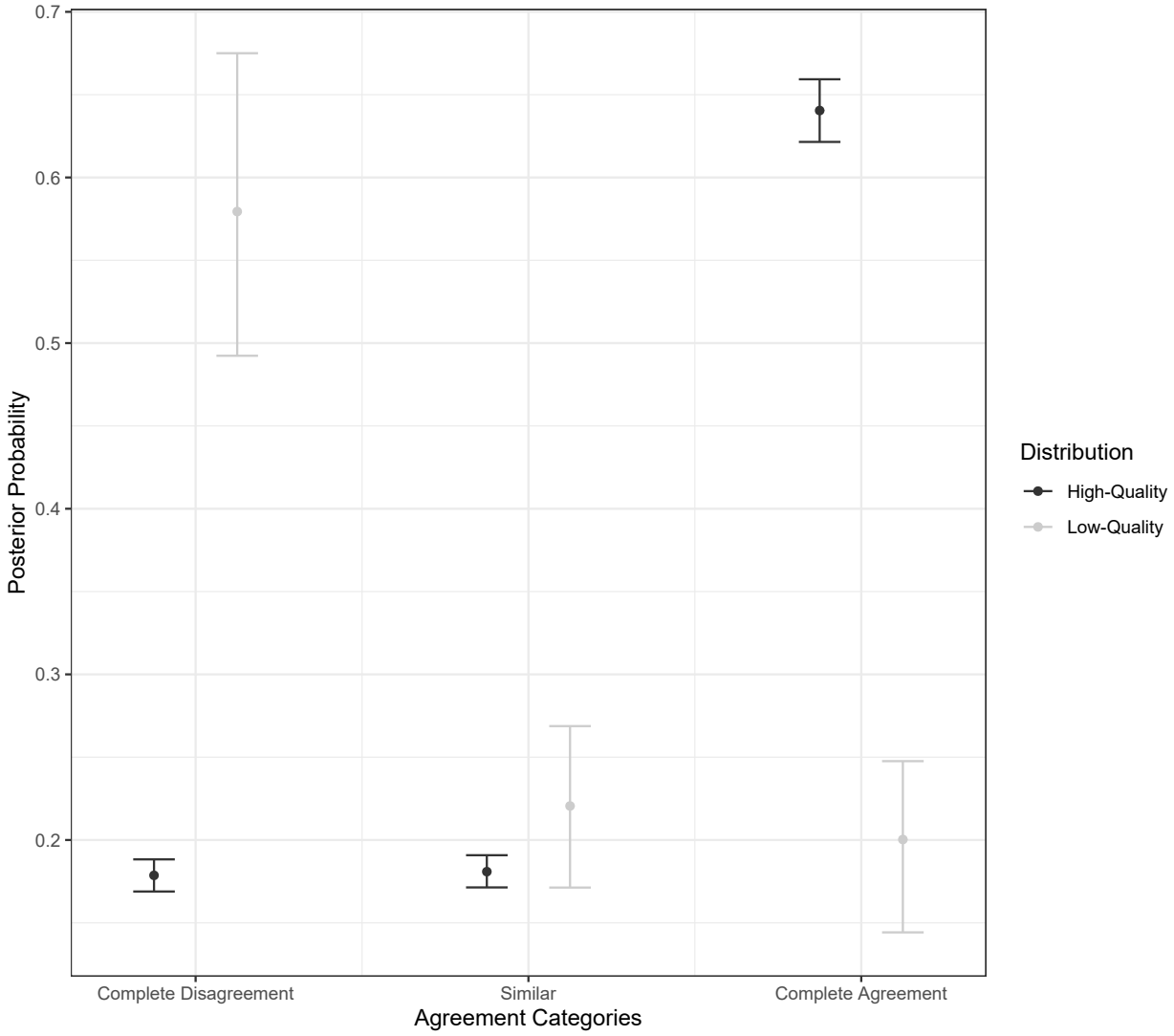


Figure I1: $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_0$ when reinterview pairs where $\nu_0 = 6$ are omitted.

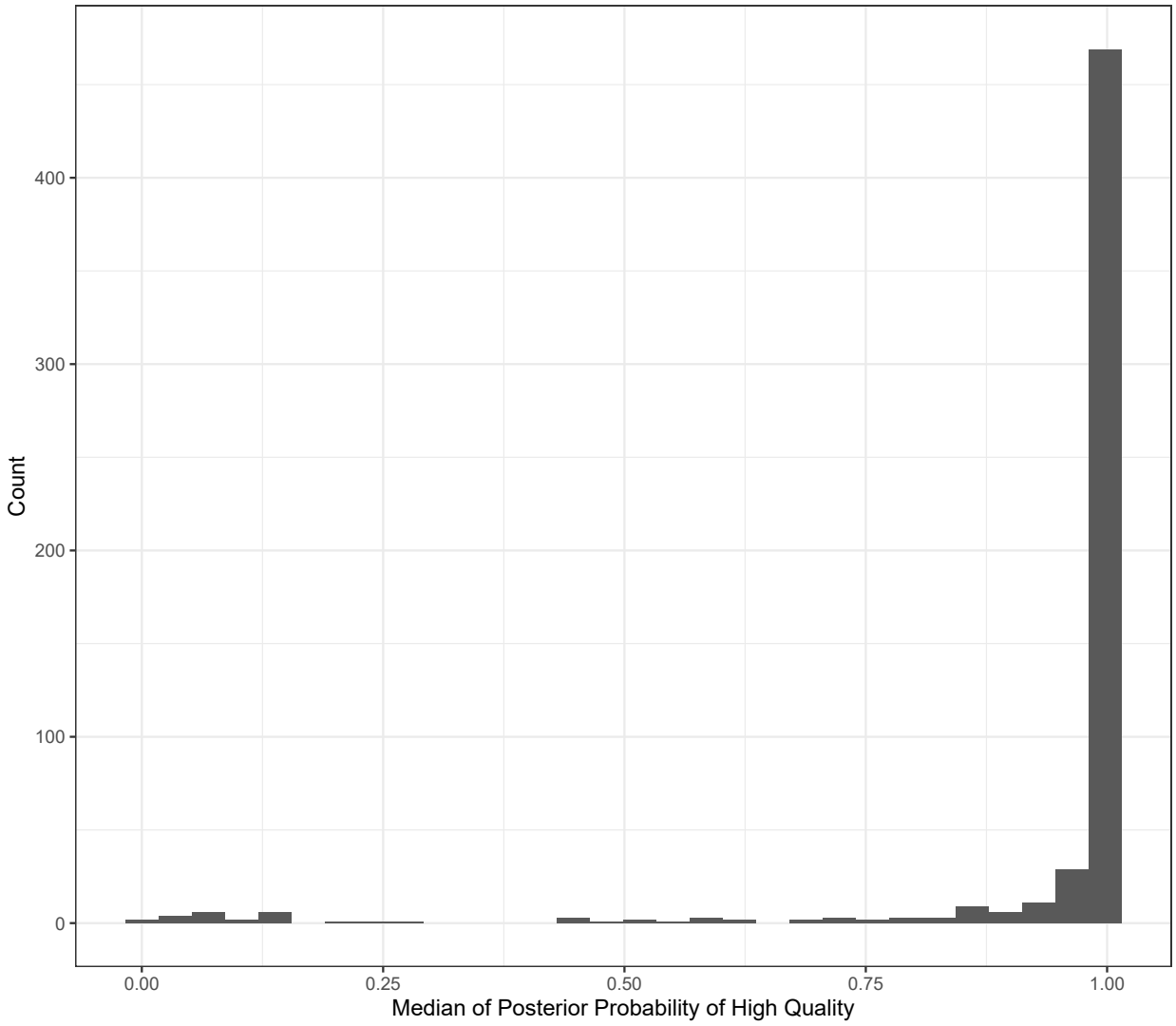


Figure I2: Histogram of the median of the posterior of the posterior probability of a match when reinterview pairs where $\nu_0 = 6$ are omitted.

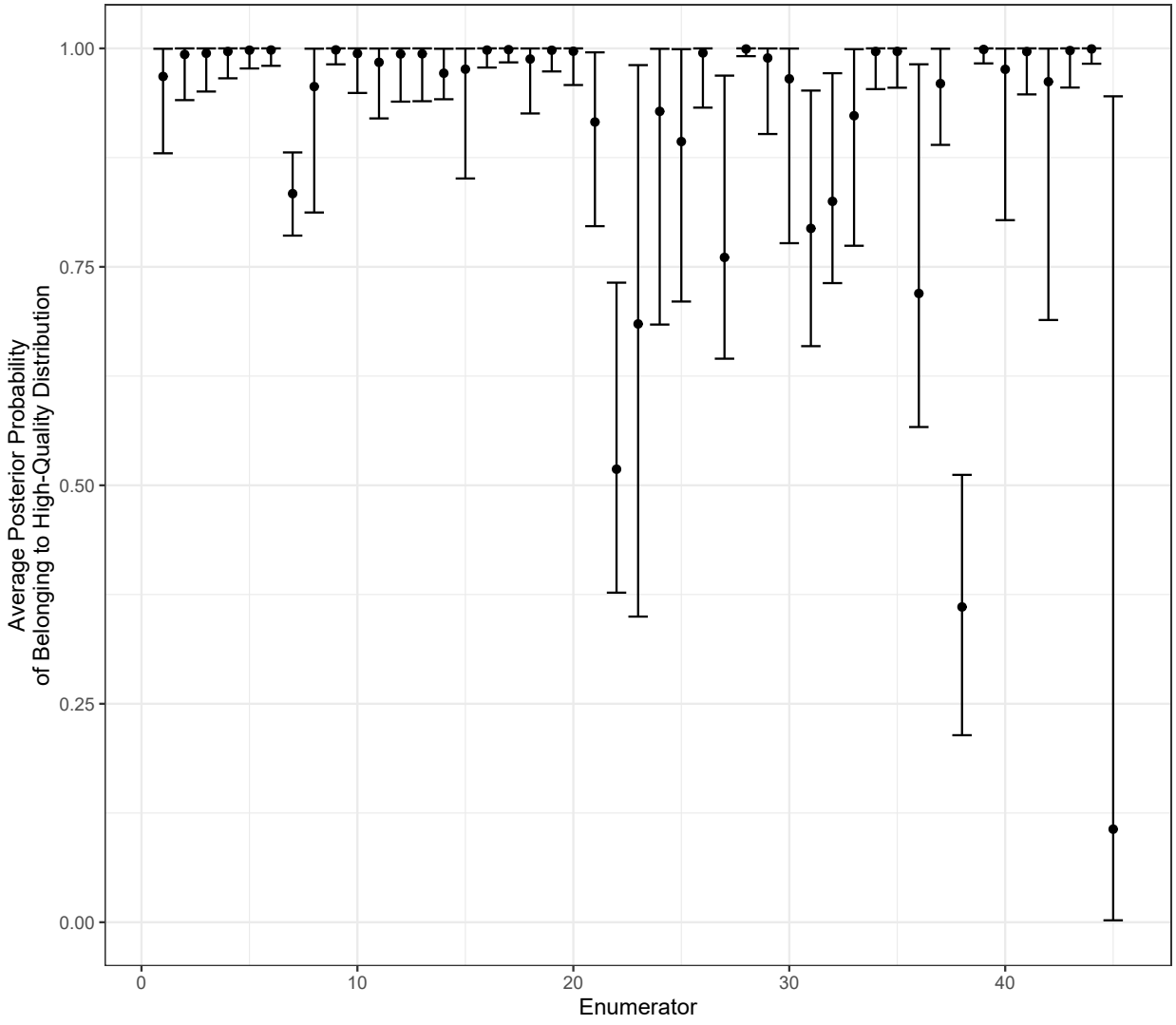


Figure I3: Median of the Posterior of the Average Posterior Probability of a Match for all 45 Enumerators when reinterview pairs where $\nu_0 = 6$ are omitted. Error bars show 95% credible intervals.

J Receipts and Enumerator Data 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.

J.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 data quality Q_e. For each enumerator e:

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

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.

J.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:

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

I ran each model for 1000 post-warm up iterations on two chains.[11] This results in 1000 model fits, each with 2000 draws from the posteriors of the parameters.[12] 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.

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

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

J.3 Results

Figure J1 shows the change in probability of a receipt being reported shown by an enumerator for different changes in enumerator data quality. Figure J2 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 J1: Median of the posterior predictive distribution of the difference between the probability at different values of enumerator data quality. Error bars show 95% credible intervals.

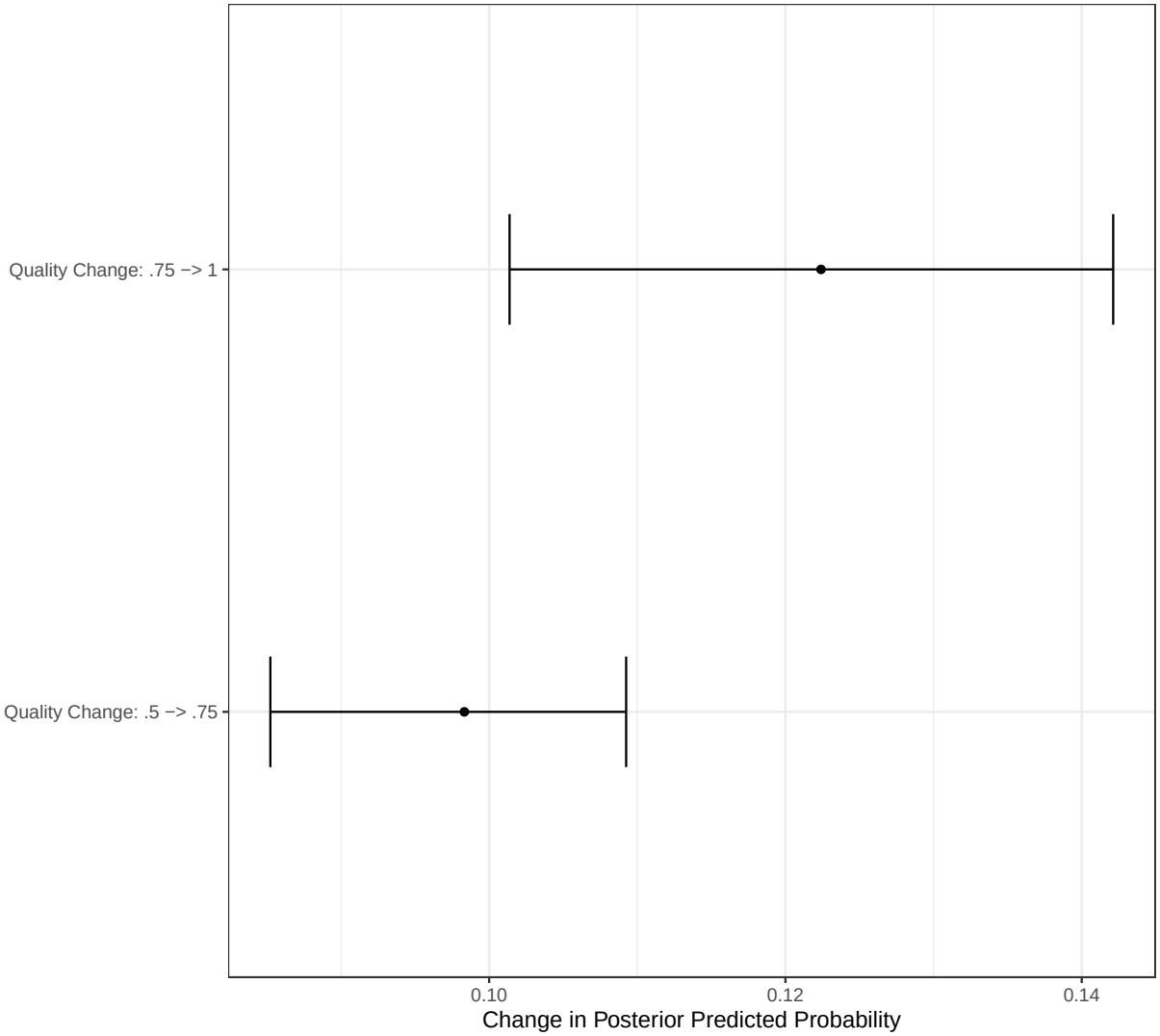
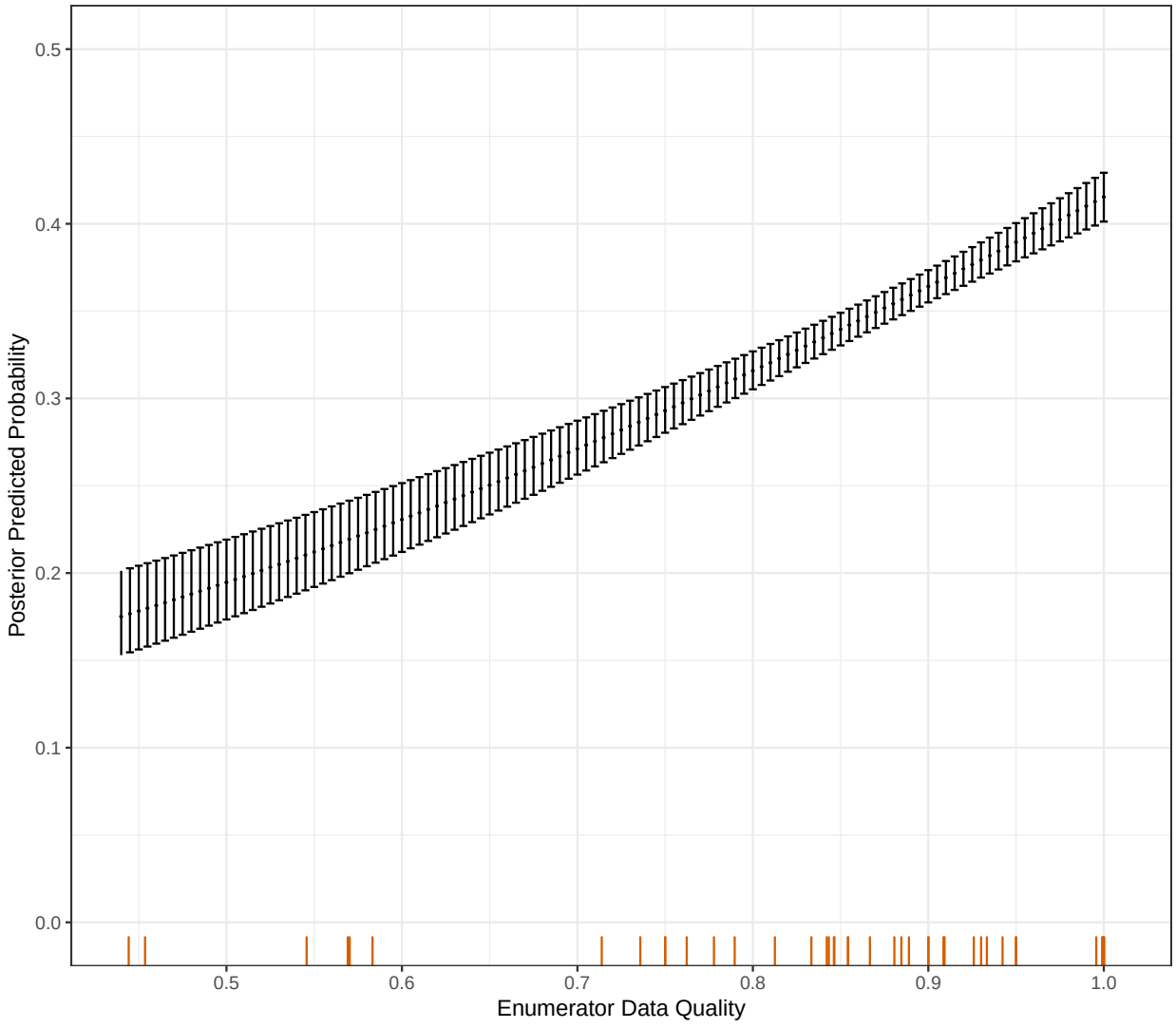


Figure J2: Median of the posterior predictive distribution of the probability of being shown a tax receipt at different values of enumerator data quality. Error bars show 95% credible intervals.



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

K.1 Data Collection Strategy

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

K.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 (IPA), under contract from NORC at the University of Chicago. IPA refers to the series of surveys and other monitoring work connected to the LGAP impact evaluation as the Tax Decentralisation Project (TAD). As part of its contract, IPA carried out three surveys at baseline and endline with different populations: 1) ward councilors 2) market tax (fee) collectors 3) market vendors. The survey used in this study and described in greater detail in this appendix is the endline survey of market vendors.

K.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.

K.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 **main** 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*)

K.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 reinterview and in the original survey except for the receipt question. The question used to check 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

K.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

K.4 Population Under Study

The population under study was market vendors in 128 markets[13] in eight districts[14] in Malawi from October 2018 to January 2019.

K.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

[13]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

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

enumerators then 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.

K.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.

K.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 reinterview 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.

K.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. IPA did not share a response rate for the vendor survey.

K.9 How the Data Were Weighted

The data were not weighted.

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

IPA performed high-frequency checks. Data were collected on tablets using SurveyCTO. Logic checks were built into the survey.

IPA also carried out a backcheck (reinterview; IPA uses the term backcheck in their materials and documentation) between November 28, 2018 and December 17, 2018 (inclusive). Four enumerators (not involved with in-the-field data collection) attempted to recontact randomly selected respondents by phone (respondents provided phone numbers during original in-the-field enumeration). IPA did not calculate a response rate for these backchecks. If the backchecking enumerators could not confirm identity, they considered a backcheck "failed." However, they did not consistently collect information on whether they could not confirm identity because backcheck respondents were not reachable, refused to participate in the backcheck, or claimed they were not the person selected for the backcheck.

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

K.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.