# QualMix: Using Mixture Models to Assess Survey Quality

Simon Hoellerbauer[*]

September 10, 2021

## Abstract

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

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

# 1 Introduction

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

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

I first briefly summarize the literature on survey quality, focusing in particular on data reliability – data issues that will induce measurement error. I also discuss issues with present approaches to dealing with data quality concerns. I then describe the general QualMix model, which relies on the logic underpinning probabilistic record linkage (Enamorado, Fifield and Imai, 2018; Fellegi and Sunter, 1969). I go on to underline the models flexibility by explaining possible extensions.

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

Using QualMix to assess survey data quality is not meant to replace other approaches to estimating survey response quality.[1] Yet, it streamlines and makes less arbitrary a step that is already part of researchers' and survey firms' quality assessment workflow. In addition, it provides respondent-level (and potentially enumerator-level) summary assessments that can potentially be incorporated into analysis.

## 2 Reliability and Survey Quality

The goal of the QualMix model is to assess survey data quality. There are many sources of data quality issues with surveys, all of which can contribute to survey error. Respondent satisficing, mode effects, implementer policies, poorly thought out questions, survey data fabrication, low quality enumerators — all of these and more can create a mismatch between a respondent's "true" response and the response that is actually collected. In other words, they can induce measurement error. Unfortunately, "[w]hile there has been considerable conceptual work regarding the measurement of [survey data] validity, translating the concepts into measurable standards has been challenging" (Madans et al., 2011, 2) In response to this, Alwin (2016) has proposed that "*the reliability of measurement should be used as a*

[1] For example, this method is conceived to assess data reliability and is not optimal for assessing issues of data quality due to lack of concept or construct validity, which usually requires "multiple statements on the focal domain," and subsequently try to assess similarity of these statements (Blasius and Thiessen, 2012, 12), often using a cohesive analytic framework such as multitrait-multimethod (MTMM) designs (Alwin, 2011, 2007). The method described here also is not intended to be used to analyze the quality of responses to particular survey items.

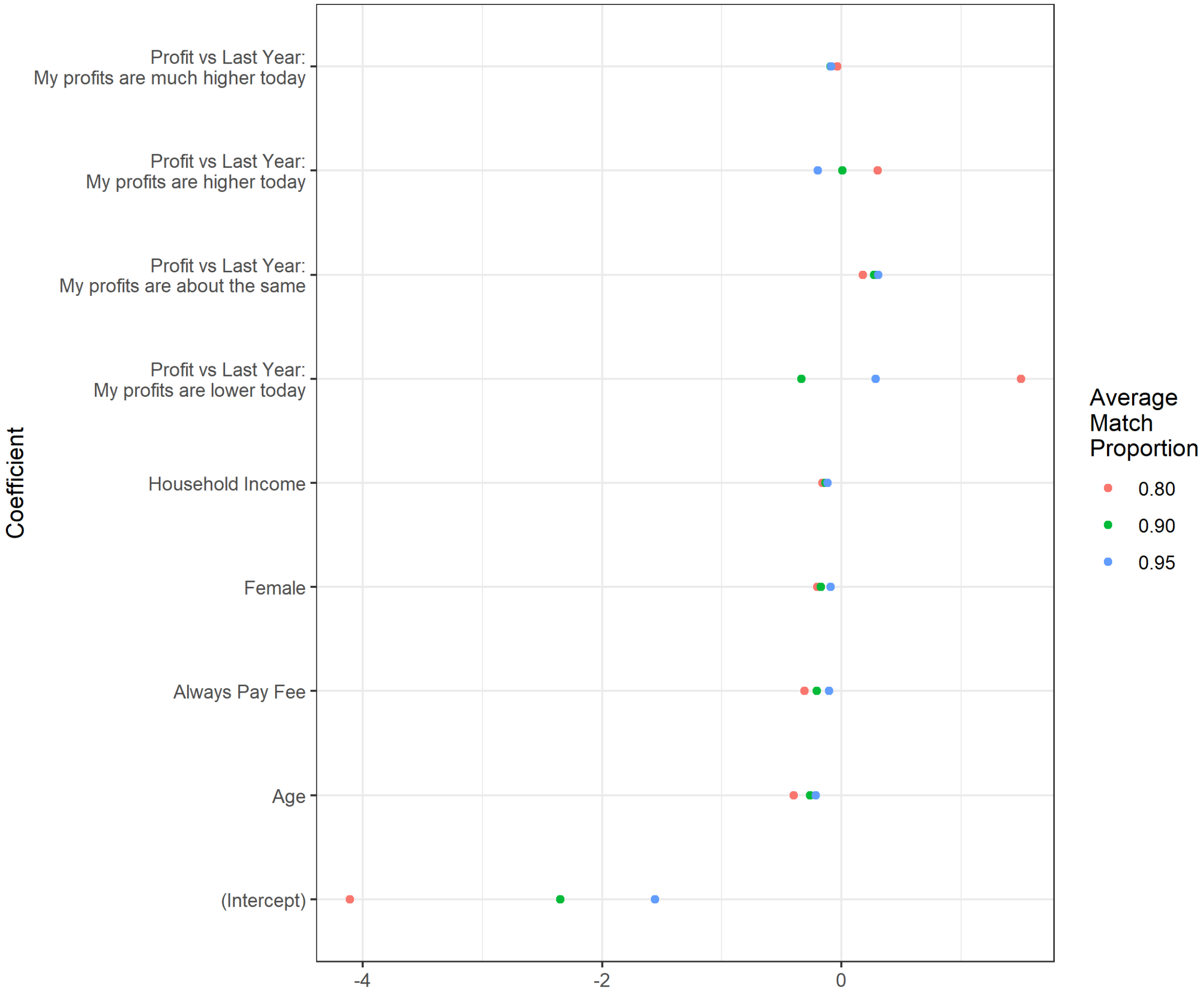
*major criterion for assessing differences in measurement quality"* (3, italics in original). Reliability refers to "agreement between two efforts to asses the underlying value using *maximally similar*, or replicate, measures" (Alwin, 2016, 7). Alwin points out that without data reliability, there cannot be data validity. In other words, without accurately recording data, it is difficult to make judgments about whether data reflect concepts.

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

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

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

Figure 1: Bias as Proportion of Original Effect Size



finding fabricated data, rather than assessing quality outright.

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

## 2.1 Backchecks

Backchecks—also called re-interviews, recontacts, callbacks or field audits—form a core part of the data quality assessment strategy at most major survey firms that do interviewer-administered surveys (Murphy et al., 2016). For example Innovations for Poverty Action includes re-interviews (backchecks) in their "Minimum Must Dos" for "every research project at IPA" (IPA, 2018). The World Bank states that "[b]ackchecks are an important tool to detect fraud" and "help research assess accuracy and quality of the data collected" (DIME, N.d.). The U.S. Census uses re-interviews extensively as part of its quality assessment procedures (Schreiner, Pennie and Newbrough, 1988; Forsman and Schreiner, 1991; Krejsa, Davis and Hill, 1999).

Forsman and Schreiner (1991) explain that re-interviews can serve multiple purposes:

> There are two major purposes for conducting reinterviews: (1) to evaluate field work; and (2) to estimate error components in a survey model. In the first area, reinterview is used to identify interviewers who (la) are falsifying data, and (lb) misunderstand procedures and require remedial training. In the second area reinterview is used to (2a) estimate simple response variance, and (2b) estimate response bias. (280-281)

Conceptually, therefore, re-interviews can be used to assess survey quality *and/or* find falsified data. This is reflected by the protocols developed by IPA and the World Bank. Nevertheless, the survey literature primarily discusses re-interviews in the context of finding falsified data, starting with Crespi (1945).

Most of the literature points out that random re-interviews as a way to identify data falsification by enumerators is somewhat inefficient (Schreiner, Pennie and Newbrough, 1988; Krejsa, Davis and Hill, 1999; Bredl, Storfinger and Menold, 2013). In such an approach, it is always possible that an observation chosen at random for a "cheating" enumerator will not have been fabricated, unless the enumerator fabricated *all* of their responses. Random sampling might lead to too many "good" enumerators being chosen for backchecking; in other words, random backchecking may be inefficient and costly. As such, survey analysts and statisticians have proposed a series of methods for detecting interviewer falsification without using backchecks, relying instead on paradata and characteristics of the response data, such as applying Benford's Law to numeric data entries. Researchers have suggested using these features in logistic regression (Li et al., 2011), unsupervised clustering algorithms (De Haas and Winker, 2014, 2016; Rosmansyah et al., 2019), and random forests (Birnbaum et al., 2013). Each of these methods shows promise for identifying observations that may be fraudulent.

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

The QualMix model can be used to assess overall quality *and* to detect falsified data when applied to backchecks. It can also generate uncertainty estimates about the quality of individual observations. In the next section, I turn to a formal presentation of the model.

# 3 QualMix: Survey Quality and Mixture Models

This section describes the general approach and the probabilistic model behind QualMix. The method is inspired broadly by the probabilistic record linkage model proposed by Fellegi and Sunter (1969; see also Enamorado, Fifield and Imai (2018)).[3]

## 3.1 General Approach

Suppose that for $n$ survey respondents we have two sets of responses to the same $K$ questions, $\boldsymbol{R_a}$ and $\boldsymbol{R_b}$, both with dimensions $n \times K$, where $\boldsymbol{r_1}_i$ 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

[3] It is important to note that it is not, however, quite the same. Whereas in probabilistic record linkage, the goal is to identify potential matches in the absence of identifiers, the aim here is to identify potential non-matches where identifiers for two sets of responses already exist. In some ways, the approach described here is the reverse of probabilistic record linkage, as the assumption is that “correct” unique identifiers exist. By *correct*, I mean that while we are uncertain that the two sets of information correspond to the same respondent, we are certain that they *should*. For example, in the context of re-interviews, discussed below, if respondent $i$ is chosen to be re-interviewed, the contact information provided by respondent $i$ in their interview is used to create re-interview $i$. The assumption is that there are not systematic administrative mix-ups in the creating of respondent identifiers and the collection of information. At the same time, this method can be used to identify such issues as well, if they exist, although then the specific quality measures described below would be confounded by administrative mistakes.

| | **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

between $\boldsymbol{r_{ak,i}}$ and $\boldsymbol{r_{bk,i}}$ as $\gamma_{ik}$, 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.[4] Table 1 presents a concrete hypothetical example of the general approach, with examples of four different variable types 1) strings, 2) ordered categorical, 3) unordered categorical, and 4) continuous numeric. I first set $L = 3$, for "Complete Agreement," "Similar," and "Complete Disagreement."

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

## 3.2 QualMix Model

If data quality issues exist, we can think of two kinds—or *clusters*—of agreement summary vectors: one with more agreements — like $\boldsymbol{\nu}_1$ in Table 1 — and one with more disagreements — like $\boldsymbol{\nu}_2$ in Table 1. Crucially, we can expect that not all agreement summary vectors for sets of high quality responses — that is, response vectors that match, overall — will consist of *only* complete agreements (purely due to random chance and sporadic data entry mistakes), nor that agreement summary vectors for sets of low-quality responses will consist of *only* complete disagreements.[5] This complicates what we do with the agreement summary vectors. We could use them deterministically, by establishing another decision rule. For example, if $K = 4$ and $L = 3$, we could say that agreement summary vectors with at least three complete agreements represent matches between $\boldsymbol{r_{a}}_i$ and $\boldsymbol{r_{b}}_i$. However, even besides the fact that such a decision rule is highly arbitrary and becomes harder to make as the number of questions and agreement categories grow, there are advantages to a probabilistic approach. 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? Furthermore, even once we have made the decisions, it is hard to conceptualize our uncertainty about their validity. The only solution we are left with is to drop observations that we unsure about. This potentially results in a waste of observations, a reduction of power, and potential selection problems.

We can instead use these agreement summary vectors as the data for a two-component finite mixture model (McLaughlan and Peel, 2000), resulting in the following model

$$\boldsymbol{\nu}_i | M_i = m \overset{\text{i.i.d}}{\sim} \text{Multinomial}(\boldsymbol{\pi}_m)$$

$$M_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\lambda)$$

[5] It is also important to note that a non-match agreement summary vector does not necessarily represent fabricated data. It is also possible that an enumerator did an extremely poor job of collecting true responses. In analytic terms there is no distinction between a falsified response and one full of errors — both induce measurement error — but the non-match cluster does not have to represent fabricated data.

where $m = 1$ when the two response vectors generally match (i.e. are of high quality) and $m = 0$ when they do not, $\lambda$ characterizes the overall probability that the agreement summary vectors from $\boldsymbol{R_a}$ and $\boldsymbol{R_a}$ are a match or not,[6] and $\boldsymbol{\pi}_m$ is an $L$ length vector of the agreement-level probabilities for distribution $m$. The probabilistic structure—and the distribution of the individual elements of the agreement summary vector—make it possible that pairs of matched observations can fail to coincide exactly on some of variables of interest, yet still count as high-quality.[7]

The observed-data likelihood for this model is

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

It is possible to estimate the model parameters using the Expectation-Maximization (EM) algorithm or, as I do in the empirical applications below, in a Bayesian framework.

$\lambda$ represents the *overall* probability that an observation comes from the high-quality distribution (that is to say, that $\boldsymbol{\nu_i}$ represents a matching response pair) $M = 1$. If all $\boldsymbol{\nu}_i$ seem to come from the same distribution, then the estimated $\lambda$ will be close to 1.

We can also estimate the observation-specific probability that observation $i$ represents a match via what are commonly called "responsibilities," or the posterior probability of coming from the high-quality component. Intuitively, this is just the amount that the observation $i$ contributes to the likelihood when $M = 1$ divided by observation $i$'s total contribution to the likelihood:

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

I discuss in Section 3.4 how these posterior probabilities can be used as a measure of the

[6] Also called the *mixing parameter*.

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

quality of observation $i$.

## 3.3 Possible Extensions to QualMix Model

### 3.3.1 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, $\gamma_{ik}|M_i = m \sim \text{Categorical}(\boldsymbol{\pi}_{mk})$, where $\boldsymbol{\pi}_{mk}$ is a vector of the probabilities of the $L_k$ categories for question $k$. This would be useful if survey implementers were interested in these probabilities for each question - for example if they wanted to see if different questions had different probabilities $\pi_{mkl}$, 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.

### 3.3.2 Incorporating Respondent-Level Characteristics or Survey Metadata

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

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

In the case of re-interviews, however, we have additional information that we can incorporate: the original data enumerators. In the original model $\lambda$ 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 $m$ as the outcome. This allows us to see how enumerators affect the probability of the survey-backcheck pair being a match. We will use a random intercept by enumerator in this regression, which also means that we must now index $\lambda$ by $e$, the enumerator. Thus, the extended model becomes

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

$\text{logit}^{-1}(\beta_0)$ in this context represents the overall probability of a match, and the intercepts by enumerator ($\beta_e$) represent the deviations from this probability. $\lambda_e$ 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 $\lambda$'s. The benefit of this approach is that it allows for match probability to vary by enumerator.

## 3.4 Quantities of Interest: Assessing Survey Quality

QualMix can be used to assess different aspects of survey data quality. Broadly speaking, due to the nature of the data, this approach can be characterized as a test-retest measure. As such, it is best situated to asses questions of reliability — how often do repeated measurements return the same response? The posterior probability of a match $\xi_i$ encapsulates how likely it is that $\boldsymbol{r_a}_i$ and $\boldsymbol{r_b}_i$ are actually the same — they vary from 0 to 1. I argue that we can then designate the *mean* of $\boldsymbol{\xi}$ as an indicator of overall survey quality

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

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

The estimated quantities $\hat{Q_S}$ and $\hat{\boldsymbol{\xi}}$ are then our estimates of survey quality, and our confidence in the quality of an individual observation. We use the average of the posterior probability of a match to estimate survey quality instead of $\hat{\lambda}$ because the former quantity incorporates the actual data as well.

It is important to note that this method simply allows us to express our uncertainty that two sets of responses match one another; it cannot tell us which response vector is more correct. As such, the uncertainty is about *both* $\boldsymbol{r}_{a_i}$ and $\boldsymbol{r}_{b_i}$.

However, the specific data quality issue assessed will depend on the $K$ questions chosen for comparison. Using the question typology drawn up by the Abdul Latif Jameel Poverty Action Lab (J-PAL), we can conceive of three main types of questions in this context, which lead to different interpretations of the parameter estimates (Gibson, N.d.). The first are questions that are factual in nature — for example, questions about age, gender, first name, last name, and occupation, among others. The responses to these kinds of questions should very rarely change, regardless of repetition, and so the parameter estimates drawn from a model fit with agreement summary vectors drawn from these questions will, at the survey and at the respondent level, indicate our uncertainty about whether the information has been accurately collected. In other words, it helps 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. Estimates $\hat{\lambda}$ and $\hat{\xi}_i$ would represent our confidence in how well the survey was administered.

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

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

This suggests fitting three different models. We can, however, also include questions of all three kinds in one model. The caveat is that once we do, our survey quality measures will represent overall survey quality, combining the various sources that would be identified via the separate questions. It is also possible to include several types of questions separately as well, for more flexibility.[8] In either, $\hat{\lambda}$ and $\hat{\boldsymbol{\xi}}$ would be estimates of the overall quality of the survey, combining the three different types of data quality issues.

### 3.4.1 Quantities of Interest Specific to Backchecks

As laid out above, we can use parameters to represent different aspects of survey quality, and for parameter estimates to serve as estimates of survey quality. The impact of the questions on the substantive meaning of these estimates remains. When we adjust the QualMix model to incorporate information on enumerators, we can define new quantities of interest.

First, unlike the general approach, where we are agnostic as to whether $\mathrm{R_a}$ or $\mathrm{R_b}$ represent the "truth," when using re-interviews, the aim is to assess the quality of the originally collected data, $\mathrm{R_a}$. This can be investigated by assuming the following:

1. The re-interview values are correct. This is a possibility that must be taken seriously — if there are problems with the initial survey, there may also be problems with the backcheck survey. If the backcheck process creates artificial non-matches — if, for example, backchecking relies on phone numbers, and the respondent provided an incorrect or temporary phone number — then the survey quality and enumerator quality estimates will be flawed.
2. Enumerators will not be aware *which* questions are selected for re-interviewing. If they are, it is possible that they could make sure to ask respondents those questions and then not be as careful with other questions.
3. Backchecking is performed randomly, either over all observations or stratified by enumerator.

[8] 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.

The model does *not* directly assume that differences will be due to the enumerator. The high-quality and low-quality distributions are both multinomials. If it is more likely that there will be differences due to response variability for a certain variable, the high-quality multinomial can take this into account.

With this in mind, we can use specific model parameters to derive survey evaluation measures:

**Posterior Probability of Match:**

$$\begin{aligned}\lambda_{e_i} &= \text{Logit}^{-1}(\beta_0 + \beta_{e_i})\\ \xi_{ie} &= \frac{\lambda_{e_i}\prod_{k=1}^{K}\pi_k^{\gamma_{ik}}}{\sum_{m=0}^{1}\lambda_{e_i}^{m}(1-\lambda_{e_i})^{m}\prod_{k=1}^{K}\pi_{km}^{\gamma_{ik}}}\end{aligned}$$

**Survey Quality:**

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

**Enumerator Quality:**

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

The posterior probability of a match indicates our estimate of how confident we can be that the individual interviewed originally and the individual interviewed during the re-interview are, in fact, the same. As such, this can be seen as a measure of how wrong an original observation was, compared to the re-interview. The advantage of using the probability that an observation belongs to a certain distribution in the mixture (in this case the high-quality distribution), is that it uses the actual observed agreement summary vector for respondent $i$, in addition to taking into account an observation's enumerator's overall probability of having an observation match its re-interview ($\lambda_{e_i}$). Some of the previously mentioned efforts to identify "cheating" enumerators are not useful for assessing whether individual observations are falsified because they rely on enumerator level characteristics. This approach allows us to assess the probability that each observation chosen for the re-interview process has been falsified or was originally recorded incorrectly.

Using the average respondent-level posterior probability of a match instead of $\hat{\lambda}_e$ ensures that the enumerator quality estimates incorporate the data directly. This means that enumerators who have similar $\hat{\lambda}_e$ but different associated $\nu_i$ will have different quality scores. Because the posterior probability of a match varies between 0 and 1, this measure is also bounded by 0 and 1. An enumerator with a quality score closer to 0 will be of lower quality than an enumerator with a quality score closer to 1. From the perspective of the model, a lower quality enumerator is one who either fabricates data outright or who is not able to solicit correct or consistent responses from a respondent. In other words, a poor quality enumerator will produce poorer quality data and lead to more measurement error, while a higher quality enumerator will produce higher quality data and lead to less measurement error.

The survey item type is still important. If a researcher is interested in assessing enumerator quality, they should not solely choose survey items whose values can be expected to change between the original enumeration and the backcheck.[9]

# 4 Application: Backchecks

Backchecks describe a data quality checking process where a subset of respondents is interviewed again. It can be useful to identify data quality issues, including data falsifica-

[9] The goal of the method I advance in this paper is not exclusively to find “cheating” enumerators. Instead, I propose a way to use survey backchecks to get a sense of the quality of the enumerators carrying out a survey and of the survey itself. This is what Forsman and Schreiner (1991) term as the “evaluate field work” motive for re-interviews (280). This will invariably involve assessing to a certain extent whether some original observations seem to have been fabricated.

In order to identify fabrications (not just problematic enumerators) wholesale, however, it will be more effective to oversample enumerators based on various factors, including not only the quality assessments the mixture model procedure provides, but also incorporating metadata and data collected from respondents in the initial survey. The latter could potentially be done without the need for backchecks of any kind, for example. The benefit of the approach I advance here is that a survey company could use some method to oversample suspicious enumerators, but still use the scope of the backcheck data to assess general survey and enumerator quality. A balanced approach would involved weighted observations from oversampled enumerators so that the total of every enumerator’s observations count equally for assessing their quality.

tion, with interviewer-administered surveys (Crespi, 1945; Schreiner, Pennie and Newbrough, 1988; Forsman and Schreiner, 1991; Murphy et al., 2016). However, up to now there has not been a clear way to *analyze* backcheck data, nor how to use them to express confidence in a survey or the interviewers in a systematic, widely applicable way.

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

## 4.1 Simulation Study

I conduct a simulation study to assess QualMix's sensitivity to real world conditions. Specifically, I vary the percent of respondents backchecked (stratified by enumerator in the simulation) and average match proportion. Under different conditions defined by these parameters, I check 1) if the model can identify low-quality observations, 2) if the model assesses overall survey quality accurately, and 3) whether the model does a "good" job identifying problematic enumerators. Survey administrators are often wary of doing more backchecks due to their costs — less than the original survey, but potentially still substantial. Varying the backcheck rate allows me to see how the model performs with varying number of observations per enumerator — can survey administrators cut cost from the budget here yet still be confident in how well the model assesses quality?[10] Also, not all survey processes will go equally smoothly. Varying the average match proportion allows me to assess how the model performs when there are generally more or less mismatches (akin to falsified or poorly

[10] In other words, does the number of observations matter significantly for parameter estimation.

collected data).[11]

### 4.1.1 Set-Up

For the simulation tests, I varied the following parameters (fixed parameters are given in App. C):

- Percent of Respondents Backchecked (%B) $\in \{0.05, 0.10, 0.15, 0.2\}$
- Average Match Proportion ($\text{logit}^{-1}(\beta_0)$) $\in \{0.8, 0.9, 0.95\}$[12]

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

To increase the external validity of the simulation exercise, I use existing survey data for the simulation, creating artificial dissimilarities. The survey used for the basis of the simulation was carried out from October to December 2018 in 128 Malawian markets. I use the following variables for the simulation:

- whether respondent is female or not (binary)
- the respondent's age (numeric, 18-86)
- the 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)
- whether the respondent's stall offers primarily goods or primarily services (categorical, fifty-two levels)

[11] In other words, does relatively fewer observations in the non-match cluster matter for parameter estimation.

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

- how respondent's profits this year compare to profits last year (ordered, 5 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 whether the respondent is female, the respondent's age, the respondent's household income, how the respondent's profits compare to last year's profits, and how many vendors paid the market tax according to the respondent. Created during each iteration.[13]

These are all variables that could plausibly be used for backchecking, with responses that should not change during the time between first interview and the backcheck re-interview. age, education, and sell_freq were three (but not all) of the variables used during the actual backchecking done for this survey. Variables like name and phone number would normally be used for backchecking, but for privacy reasons I did not include those in this simulations. These variables all represent values that should not change much between the original survey and the backcheck, making them type one variables according to the typology referenced above in Section 3.4. As such, in this simulation, we are assessing data falsification and factual accuracy.

### 4.1.2 Fitting the QualMix Model

I fit the QualMix model to the resulting set of agreement summary vectors using `Stan`'s `R` interface `rstan` (Team, 2020).[14] I ran each model for 1500 iterations each on four chains (for a total of 3000 post-warm-up samples from the posterior).[15]

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 $\boldsymbol{\pi}_1$ 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

[13] See Appendix C.2 for more information how this variable was simulated.

[14] See Appendix C.1 for the full model specification.

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

high-quality records tend to have a higher probability of similar entries in the agreement summary vector, and vice-versa for low-quality records.

### 4.1.3 Results

I first show results for out-of-sample model fit, then for how the well the model assesses survey and enumerator quality.[16]

**Model Fit**

Figure 2 shows the Area Under the Receiver Operating Characteristic Curve for all parameter combinations, calculated out of sample (on the observations not selected for the backcheck in each simulation). The figure demonstrates that the model does exceptionally well at identifying observations that are not the same between $\boldsymbol{R_a}$ and $\boldsymbol{R_b}$, with AUCs very close to 1. There are some minor differences in performance – the change between the lowest median AUC and the highest is only 0.005965. In general, performance improves the higher the average match proportion. Model performance also somewhat improves as the backcheck portion increases, although this is not consistent across overall match proportion.

In this application, we are worried about both false negatives and false positives – determining that two sets of data do not match when they do, and determining that two sets of data match when they in fact do not. Figure 3 shows the false negative and false discovery rate (also known as the false positive rate), considering an observation a match if its responsibility is greater than or equal to .5. As the figure shows, both decrease strongly as the average match proportion increases. In other words, if there are more observations that match, the model makes fewer mistakes with respect to both false positives and false negatives. The figure also shows that, keeping the overall match probability constant, there is little change when the backcheck proportion increases — the false negative rate goes slightly up (which

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

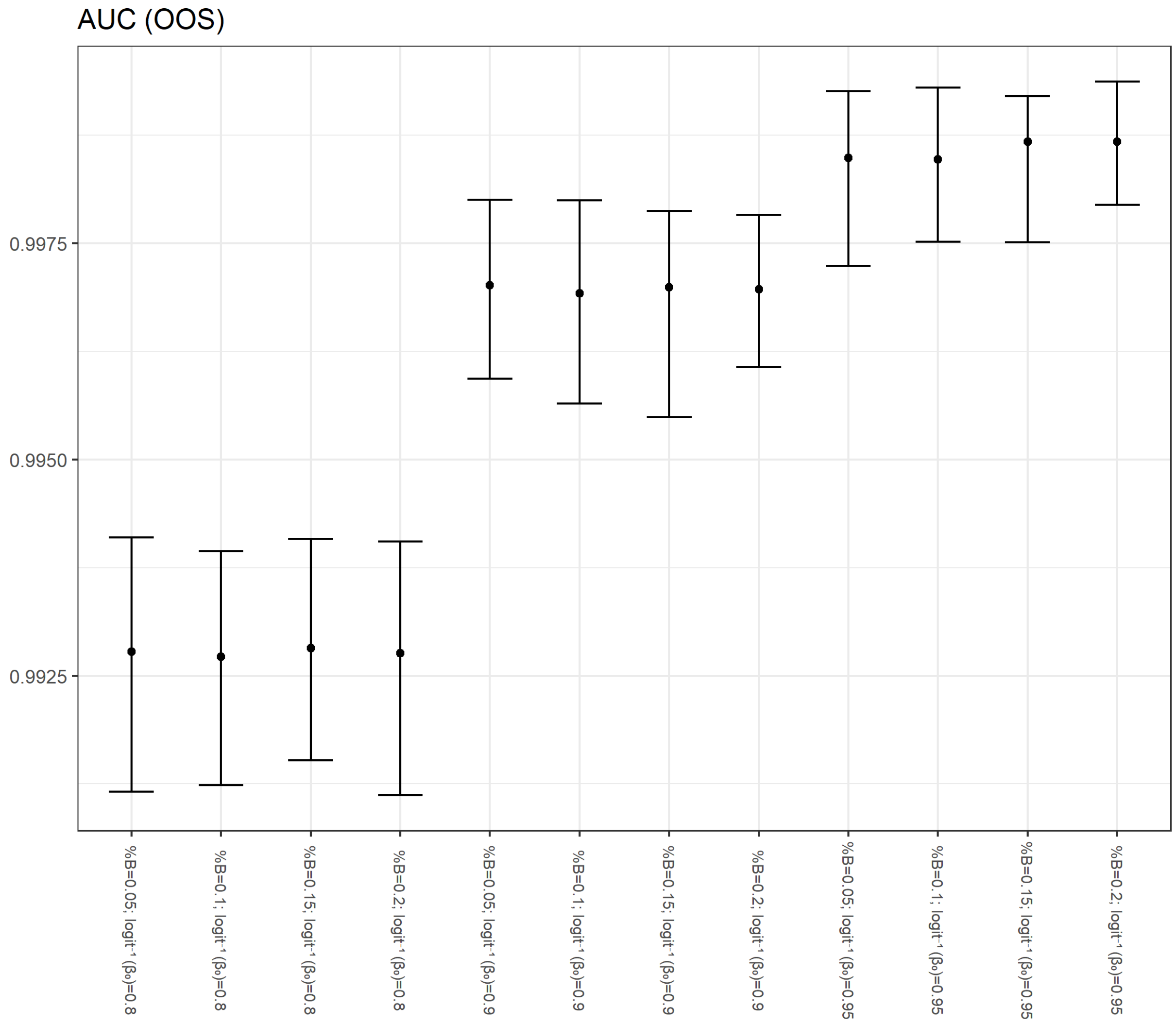


Figure 2: Median of the posterior of the Area Under the Receiver Operating Characteristic Curve for different parameter combinations with 95% credible intervals.

makes sense, because there are more chances to make mistakes), while the false discovery rate generally goes down (which again makes sense, because the model has seen more data). However, these differences are very small in substantive terms. This should be reassuring to researchers and survey implementers, as more backchecks require more resources.

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

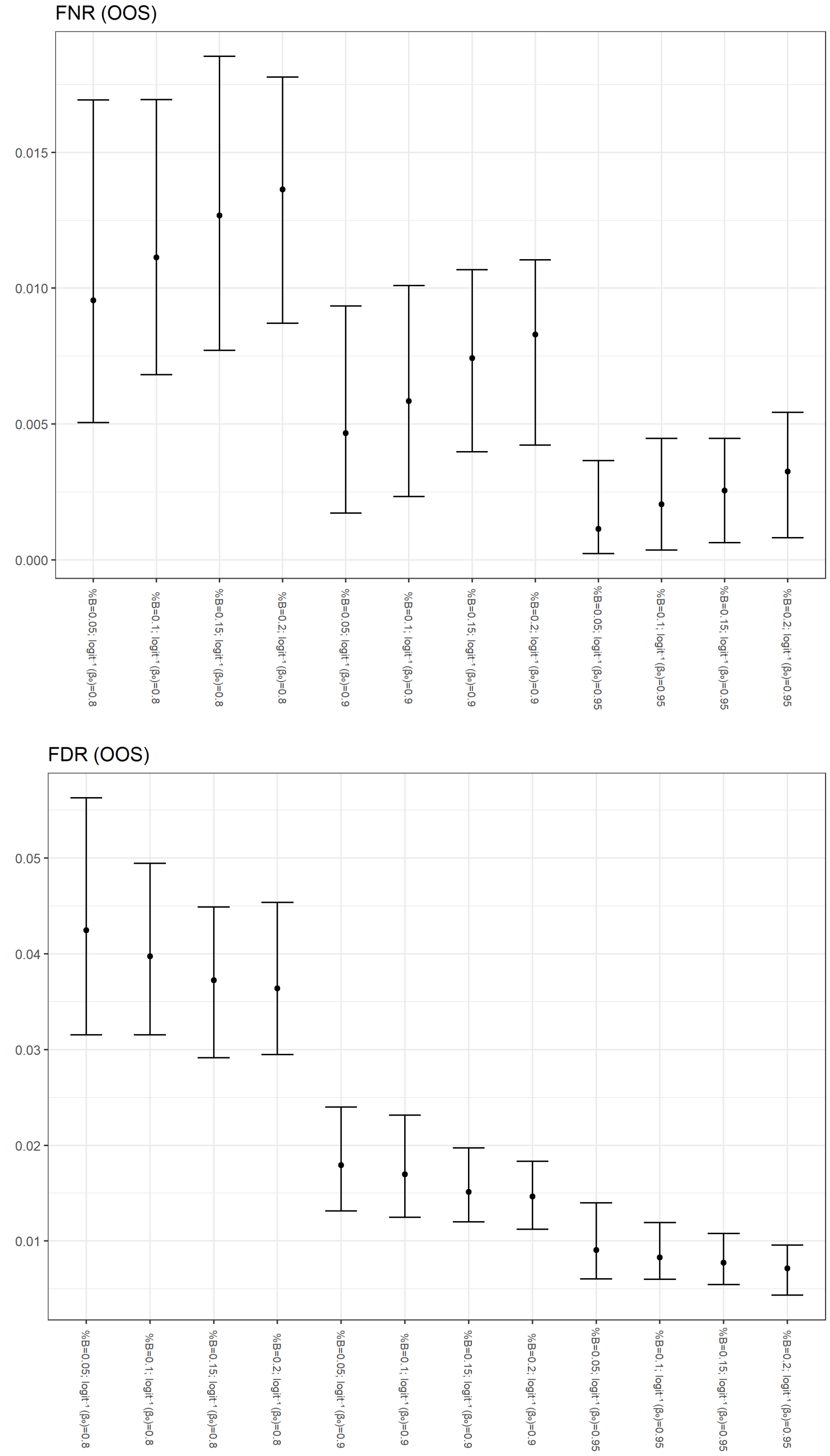


Figure 3: Median of the posterior of the False Negative Rate and the False Discovery Rate for different parameter combinations with 95% credible intervals.

**General Quality Evaluation**

Next, we turn to seeing how well the model helps assess survey and enumerator quality. Because of the questions chosen for this simulation (see Section 4.1.1), the data quality in question here reflects data collection accuracy. In order to calculate the error, I use the true proportion of matches for the survey and for each enumerator, respectively, as true values of $Q_S$ and $Q_e$. I then use the estimators for these quantities proposed in Section 3.4.1 and calculate the error. Figure 4 shows the error under different parameter combinations, using $\frac{\sum_i^N \hat{\xi}_i}{N}$ (bottom panel). We can see the errors are around 0 for all parameter combinations, with the mean error decreasing slightly as the average match proportion goes up. Keeping the overall match probability constant, the mean error is perhaps somewhat lower when only 5% of observations are backchecked, although the credible intervals are much larger. After that, the error does not change much as the backcheck percentage increases, with credible intervals becoming smaller due to the larger number of observations exposed to the model.

Figure 5 shows the bias in enumerator quality (measured by the proportion of matches per enumerator) using $Q_e$ as the estimator for enumerator quality. 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 figure demonstrates that enumerator quality bias is mostly clustered tightly around 0. A few enumerators display somewhat large negative bias - the largest absolute bias in any parameter combination is 0.084. These three enumerators are the three "worst" enumerators with respect to $\beta_E$ (aka they deviate the most from the average match proportion). I suspect that because the probability of a match is generally low for these enumerators, the probability of a match making into the backcheck is lower. Also, these enumerators will make more errors in match observations because of their lack of quality (see Appendix C for why this is the case) — the model struggles somewhat to pick this up because they already are very "bad." In other words, the match and non-match observations of these enumerators may be in fact very similar. Thus, the model assesses these enumerators as worse than their actual match proportion. The variance and

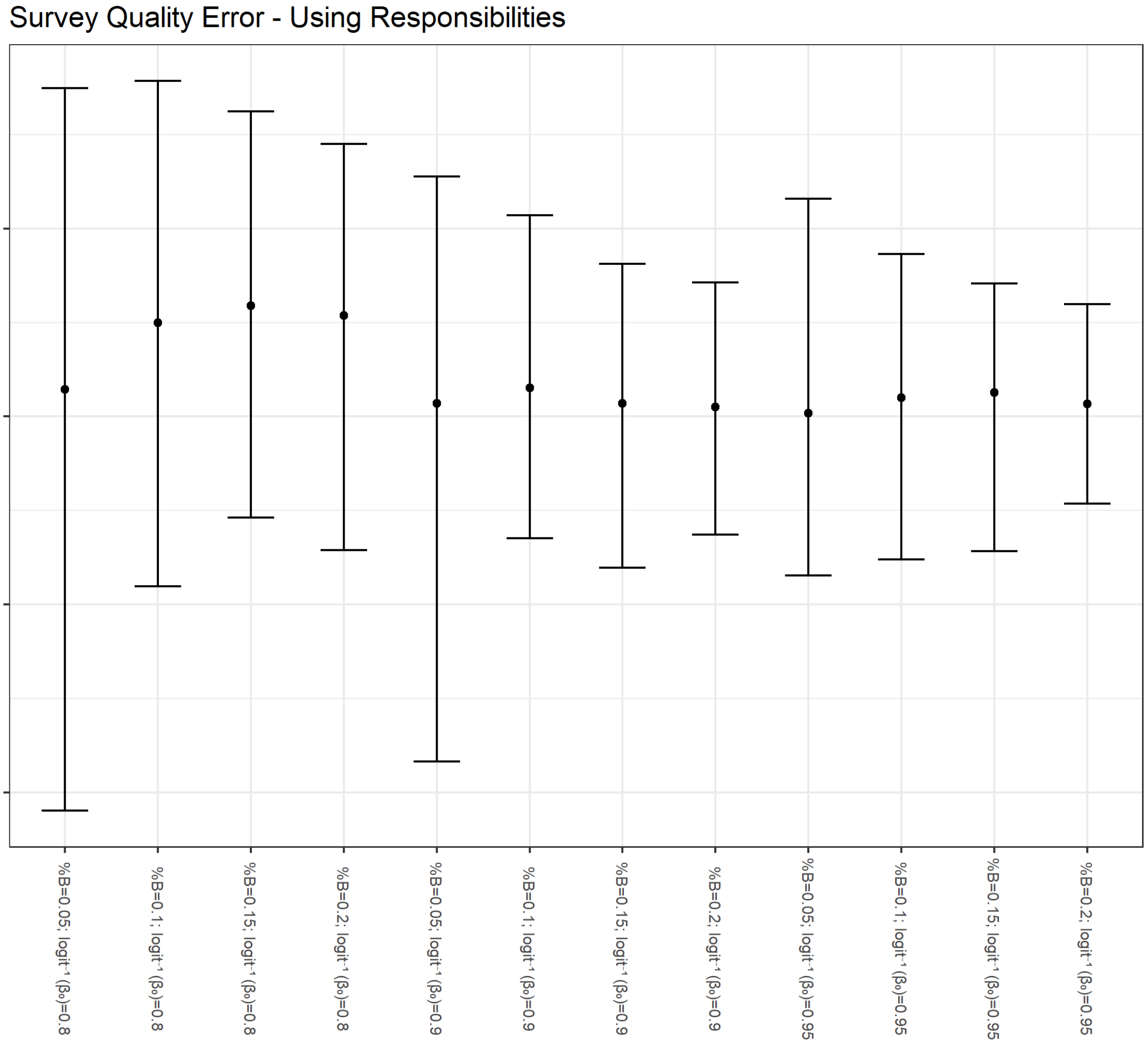


Figure 4: Mean of the posterior of the error (Empirical Bias) of overall survey quality for different parameter combinations with 95% credible intervals.

mean of the bias across enumerators go down as the average match proportion goes up (as overall quality increases, it becomes easier to identify poorly performing enumerators).

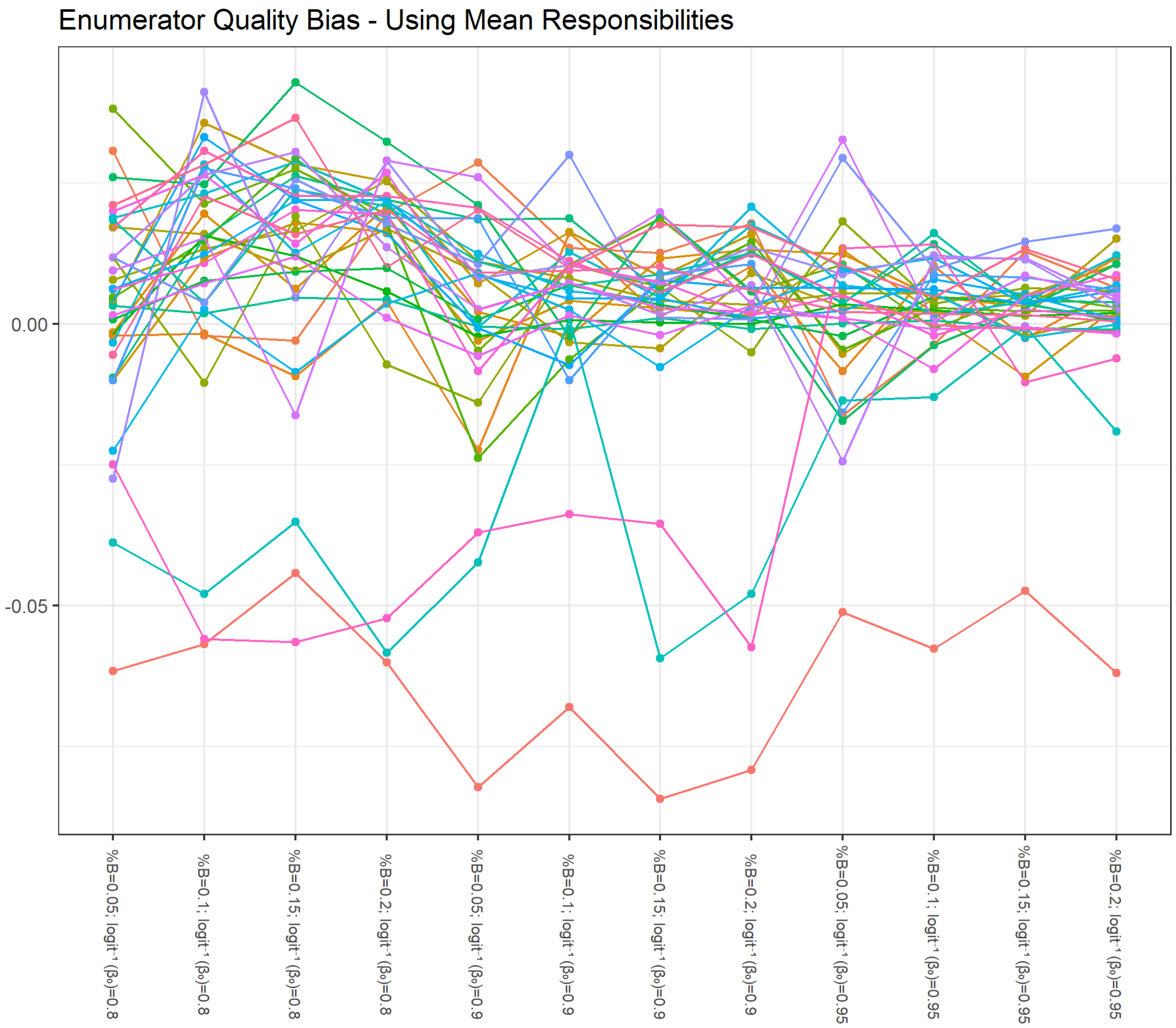


Figure 5: Bias of enumerator quality for different parameter combinations.

## 4.2 Real World Application

I next apply the QualMix model to a real world case: the survey used as the starting point for the simulation, carried out in Malawi between October and December 2018. This time, I use the actual backcheck data. Of the 12,370 respondents, 657 (5.3% of the sample) were re-contacted by telephone in November and December 2018. Backchecks were not stratified by enumerator. Fifty-one enumerators were used for the study, but only forty-five had a respondent recontacted (the other six interviewed very few respondents; one of the forty-five interviewed only one respondent, who was then randomly chosen for a backcheck). Figure 6 shows the percentage of each enumerator's respondents who were randomly chosen for the backcheck process. The one enumerator with a 100% backcheck rate is omitted from the figure to make it easier to interpret.

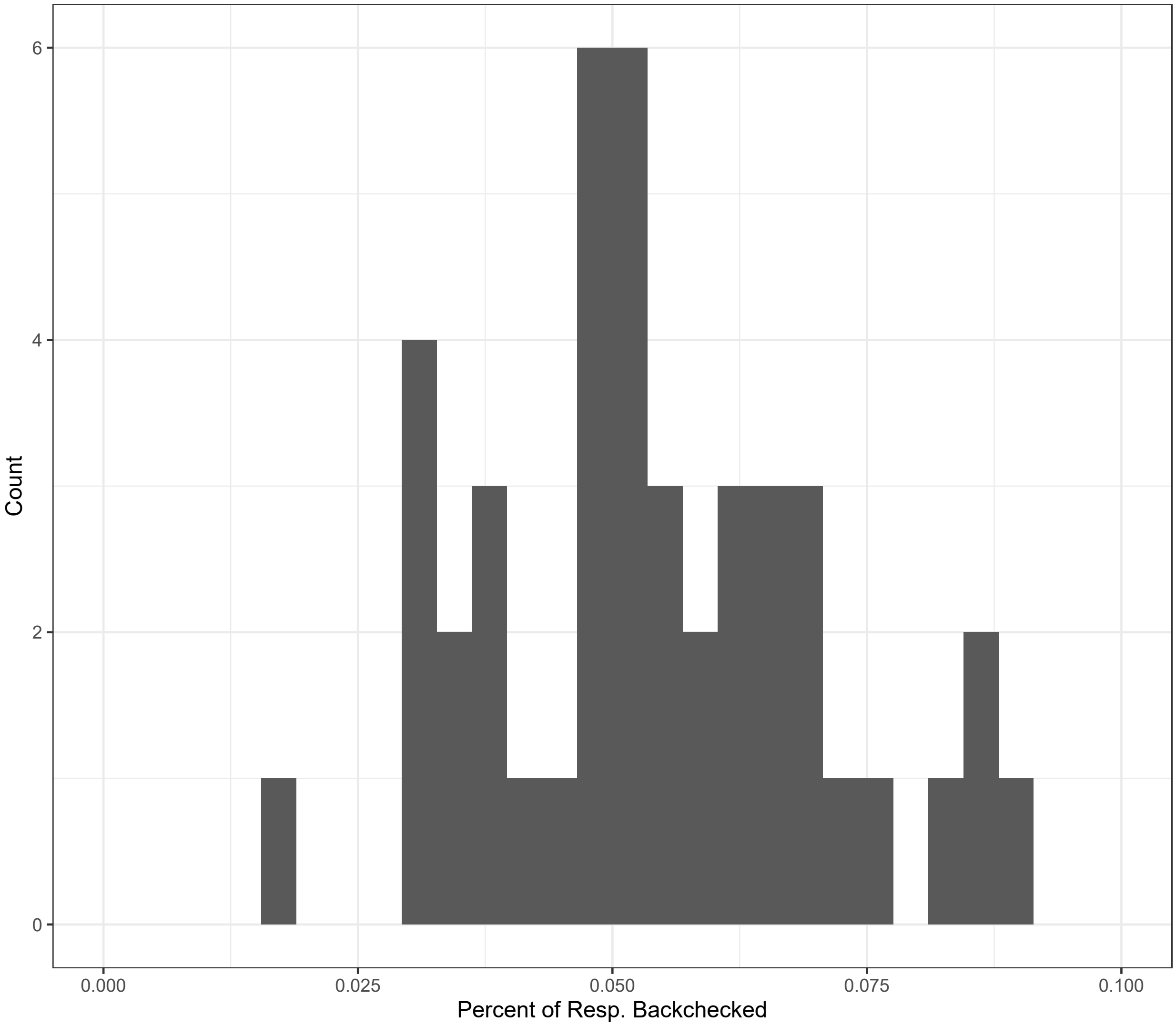


Figure 6: Histogram of Percentage of a Enumerator's Respondents Chosen for the Backcheck.

There were six variables chosen for all backchecks:[17]

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

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. how satisfied the respondent is with developments in the market.

All of these should not have changed between original enumeration and the backcheck, besides perhaps satisfaction with developments. As such, we can use these backcheck questions to assess the reliability of the survey. I create $\boldsymbol{\nu}$ for each backcheck pair. I categorize NA values as disagreements. I use the model described in Section 4.1.2 without alterations to derive estimates of enumerator and survey quality.

I fit the model using `stan` using the same priors as in the simulation.[18]

### 4.2.1 Results

Figure 7 shows the two estimated multinomials, demonstrating that the model was able to identify distinct distributions over agreement categories.[19]

It is important to note that the low-quality distribution puts almost all of the probability into the "Complete Disagreement" category. Of the 657 backcheck observations, 85 had no correct values - these were all cases where a different person answered the phone than the one interviewed for the survey (most often the individual who answered the phone did not know the person originally interviewed) or where no one answered the phone. The survey company considered these as failed backchecks, but it is important to take these cases into account - after all, it is possible that the original observations were fabricated.[20]. However, the sheer

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

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

[20] It is unfortunate that backchecks were done using telephones, although this saves on expense; it is possible that individuals would sell their phones or sim cards. It is also unfortunate that backcheck interviewers were directed to ask no further backcheck questions if the name of the person who picked up the phone did not match the one in the survey. Nevertheless, if this were a random process, then we would expect the distribution of such backcheck failures to be uniform among enumerators. Figure 8 clearly shows that this is not the case. See Appendix D for an analysis of the same data with all $i$ for which $\boldsymbol{\nu}_i = 6$. Once these

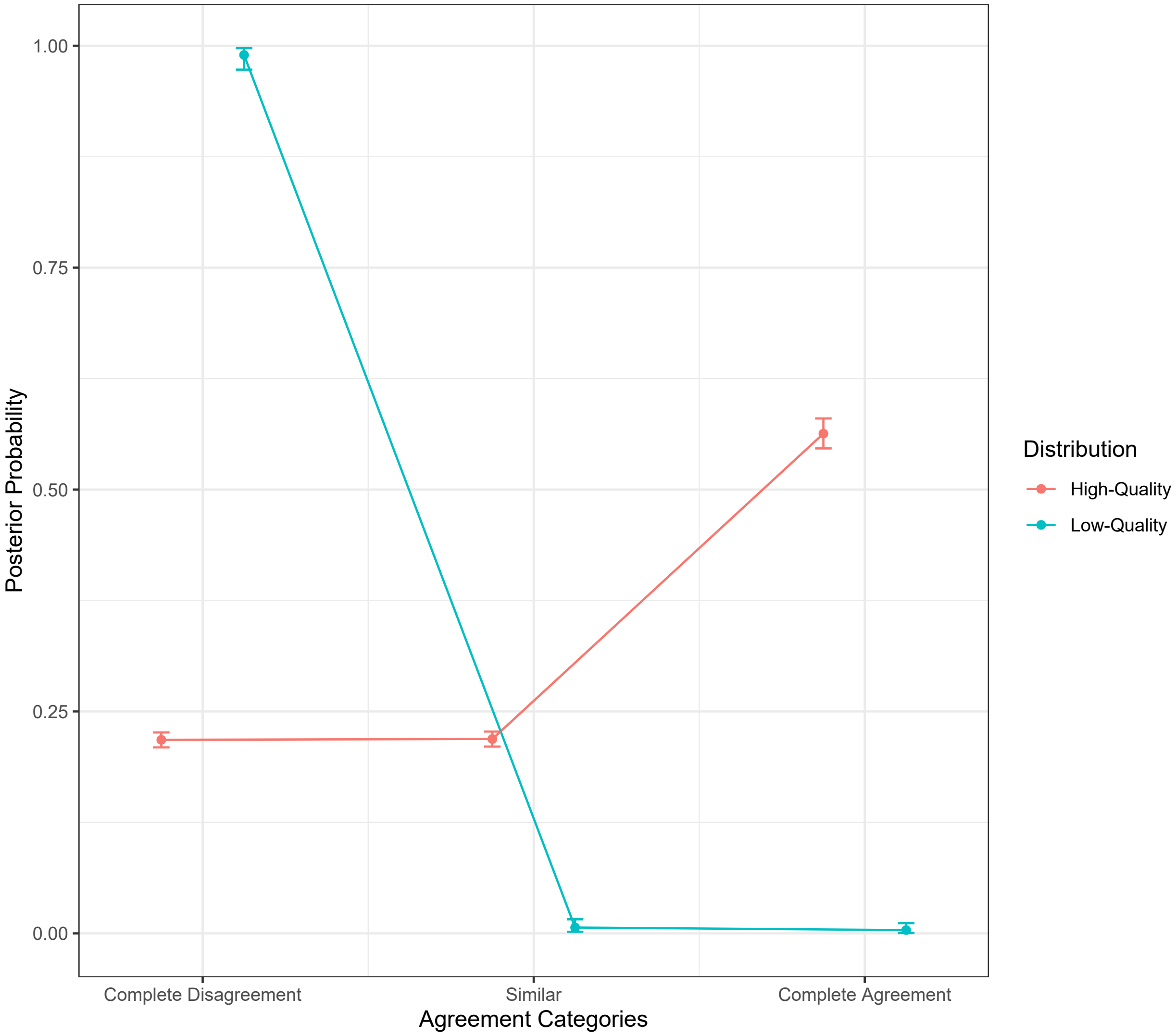


Figure 7: Median of posterior distributions of $\hat{\boldsymbol{\pi}}_0$ (non-match) and $\hat{\boldsymbol{\pi}}_1$ (match) along with 95% credible intervals.

number of these cases (12% of backchecked respondents) means that it was straightforward for the model to identify all of these observations as belonging to the same cluster. The match distribution clearly still contains "Complete Disagreement" values. The 95% credible intervals for the expected value of the categories are shown in Table 2.

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

Table 2: 95% Credible Intervals for Expected Values of the Match Distribution

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

The value of this analysis is that it identifies what our model considers a "match." Survey administrators and researchers must decide whether they are satisfied with the high-quality and low-quality distributions.[21] Even if they are *not* satisfied with a match distribution, however, the model and approached defined here still have utility, as they identify common patterns in the data.[22] If a match distribution is unsatisfactory, that is a sign in and of itself

observations are dropped, it becomes more difficult to detect two distinct distributions.

[21] If doing analysis in a Bayesian framework, they can also put stronger priors on the parameters of the two distributions, in accordance with what they consider acceptable or not.

[22] Note that this implies that will it will be possible and perhaps beneficial to compare high-quality and

that something might have gone wrong during data collection.

Using the model, we can derive enumerator quality estimates using Eq. 1, shown in Figure 8. There is considerable variation in enumerator quality. Some enumerators clearly seem to be of rather poor quality — eight have quality estimates of .75 or lower, with two below .5 (keeping in mind that quality ranges from 0 to 1, and is directly comparable within, but not between, surveys). Because of how the survey company performed the survey[23] and the backcheck, we cannot conclusively say that the enumerators were responsible for flawed data, but we can say that some are associated with many more backcheck failures: there are clear within-enumerator patterns that the model helps us see.

How can we be sure, however, that these enumerator quality estimates actually represent real world quality and not just variations in backcheck performance? As a validation exercise, I examine the relationship between the receipt variable and enumerator quality. As mentioned above, the receipt variable is one where quality can impact data collection quality starkly. More experienced enumerators would be more likely to get someone to show them a tax receipt because they may be better at getting a respondent's trust and less likely to rush through a survey.[24] Using a simple binomial logit regression model, I find that quality indeed impacts the probability that a respondent showed an enumerator a receipt. Increasing enumerator quality from 0.5 to .75 increases the probability that an enumerator reported being shown a receipt by .107; increasing enumerator quality from .75 to 1 increases it by a further .140.[25] As Figure 8 shows, this level of variation in enumerator quality is observed in the data, underscoring the vast differences in how well enumerators were able to solicit receipts.

As such, these enumerator quality estimates help identify enumerators who may have

low-quality distributions between surveys, quality estimates will not be directly comparable.

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

[24] If an enumerator did not want to have to wait for a respondent to find a receipt or did not want to have to go through the procedure of taking a picture of the receipt, they could simply declare that the respondent had not shown them one.

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

produced problematic data. Survey administrators and researchers can then investigate potential causes of these issues, and can even see if certain enumerator characteristics (more experienced versus less experienced, for example) correlate with these quality estimates. Survey administrators can also see if estimated enumerator quality is associated with certain survey measures. In addition, they can use this information in an actionable way, by down-weighting observations or enumerators about whose quality they are uncertain. It is also possible to run this model in real time, consistently updating it with data from the field. Researchers do not need to wait until the end of enumeration to apply this model to their data. In such ways, the model more systematically facilitates the identification, investigation, and solving of data quality issues than deterministic backcheck comparisons.

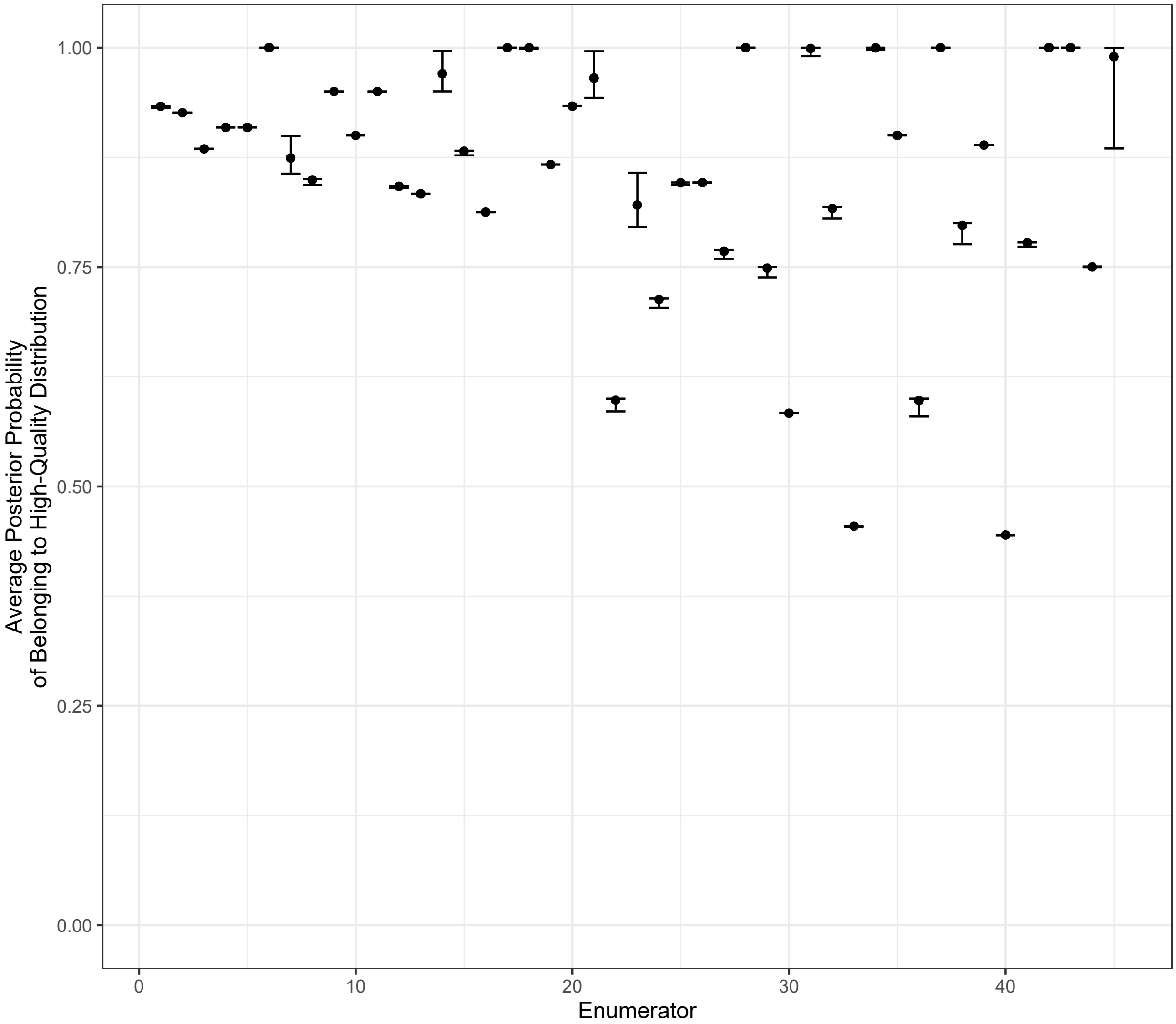


Figure 8: Median of the Posterior of the Average Posterior Probability of a Match for all 45 Enumerators. Error bars show 95% credible intervals.

# 5 Conclusion

In this paper, I lay out the QualMix model to help assess record quality — which directly affects measurement error — when two sets of responses exist for the same individual to the same questions. I suggest a mixture model approach that uses the number of inconsistencies and agreements between the two sets of records as data. The model allows us to assess our uncertainty regarding whether we have contacted the same individual, and therefore, how confident we can be that there is limited measurement error. With this model, survey administrators can estimate the overall quality of a survey, as well as the quality of enumerators implementing the survey.

The simulations demonstrate that the model does a good job of identifying problematic observations and does a fairly good job of assessing survey and enumerator quality. The model performs better when overall quality is not too low. It also shows that survey implementers can get better estimates of quality if they backcheck more than 5% of respondents, although there are not too many gains to be had by going higher than that number. The empirical application demonstrates how to apply the model to real world data. It also shows what can happen when the model struggles to separate out the two multinomials, or when there is greater uncertainty about one of the distributions. In order to make this situation less likely, I strongly suggest that survey implementers use more than six variables for backchecking. I also suggest that backcheckers ask all backcheck questions, regardless of whether a name matches — after all, it is possible that a *name* is incorrect, but that the other values are correct. Finally, in order to not confound variation in region of enumeration and enumerator identity, I suggest that, if possible, enumerators be sent to all regions of enumeration.[26]

While the approach I present here should not replace existing quality control measures (Cohen and Warner, 2021), it can be incorporated easily into existing quality control suites. The model is very flexible. It can be adapted to allow a more fine-grained analysis to

[26] This also ensures that enumerators do not confound region-specific treatment effects.

assess different kinds of survey quality. It can also easily incorporate other information, as the empirical application shows. Researchers are not limited to only evaluating survey backchecks with this model; other applicable scenarios include estimating the uncertainty that the correct respondents have been recontacted in a panel survey, for example.

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

# References

Ahmed, Bilal, Ali Ahmad, Akbar A Herekar, Umer L. Uqaili, Jahanzeb Effendi, S. Zia Alvi, Arif D Herekar and Timothy J. Steiner. 2014. “Fraud in a population-based study of headache: prevention, detection and correction.” *The Journal of Headache and Pain* 15:1–5.

Alwin, Duane F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement.* Hoboken, NJ: John Wiley & Sonds, Inc.

Alwin, Duane F. 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*, ed. Jennifer Madans, Kristen Miller, Aaron Maitland and Gordon Willis. Hoboken, NJ: John Wiley & Sons, Inc. pp. 265–293.

Alwin, Duane F. 2016. Survey Data Quality and Measurement Precision. In *The SAGE Handbook of Survey Methodology*, ed. Christof Wolf, Dominique Joye, Tom W. Smith and Yang chih Fu. Thousand Oaks, CA: SAGE pp. 527–557.

Birnbaum, Benjamin, Gaetano Borriello, Abraham D. Flaxman, Brian DeRenzi and Anna R. Karlin. 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 org and Victor Thiessen. 2012. *Assessing the Quality of Survey Data.* Thousand Oaks, CA: SAGE Publications.

Bohrnstedt, George W. 2010. Measurement Models for Survey Research. In *Handbook for Survey Research*, ed. Peter V. Marsden and James D. Wright. Bingley, UK: Emerald Group Publishing, Ltd. pp. 347–404.

Bredl, Sebastian, Nina Storfinger and Natalja Menold. 2013. A Literature Review of Methods to Detect Fabricated Survey Data. In *Interviewers’ Deviations in Surveys - Impact, Reasons, Detection and Prevention*, ed. Peter Winker, Natalja Menold and Rolf Porst. Frankfurt am Main: Peter Lang pp. 3–24.

Castorena, Oscar, Mollie J. Cohen, Noam Lupu and Elizabeth J. Zechmeister. 2021. “How Worried Should We Be? The Implications of Fabricated Survey Data for Political Science.” Working Paper. Version: July 26, 2021. `https://www.noamlupu.com/fabrication.pdf`.

Cohen, Mollie J. and Zach Warner. 2021. “How to Get Better Survey Data More Efficiently.” *Political Analysis* 29:121–138.

Cohen, William W., Pradeep Ravikumar and Stephen E. Fienberg. 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, Leo P. 1945. "The Cheater Problem in Polling." *The Public Opinion Quarterly* 9(4):431–445.

De Haas, Samuel and Peter Winker. 2014. "Identification of partial falsifications in survey data." *Statistical Journal of the IAOS* 30:271–281.

De Haas, Samuel and Peter Winker. 2016. "Detecting Fraudulent Interviewers by Improved Clustering Methods – The Case of Falsifications of Answers to Parts of a Questionnaire." *Journal of Official Statistics* 32(3):643–660.

DeDeo, Simon, Robert X. D. Hawkins, Sara Klingenstein and Tim Hitchcock. 2013. "Bootstrap Methods for the Empirical Study of Decision-Making and Information Flows in Social Systems." *entropy* 15:2246–2276.

DIME, World Bank. N.d. "Back Checks." DIME Wiki. https://dimewiki.worldbank.org/wiki/Back_Checks. Accessed: 2020-10-10.

Drost, Hajk-Georg. 2018. "Philentropy: Information Theory and Distance Quantification with R." *Journal of Open Source Software* 3(26).

Enamorado, Ted, Benjamin Fifield and Kosuke Imai. 2018. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." *American Political Science Review* pp. 1–19.

Endres, Dominik M. and Johannes E. Schindelin. 2003. "A New Metric for Probability Distributions." *IEEE Transactions on Information Theory* 49(7):1858–1860.

Fellegi, Ivan P. and Allan B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64(328):1183–1210.

Finn, Arden and Vimal Ranchhod. 2017. "Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South African Survey." *The World Bank Economic Review* 31(1):129–157.

Forsman, Gösta and Irwin Schreiner. 1991. The Design and Analysis of Reinterview: An Overview. In *Measurement Errors in Surveys*, ed. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz and Seymour Sudman. Chichester: Wiley pp. 279–301.

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

Imai, Kosuke and Dustin Tingley. 2012. "A Statistical Method for Empirical Testing of Competing Theories." *American Journal of Political Science* 56(1):218–236.

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

Jaro, Matthew. 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, Elizabeth A., Mary C. Davis and Joan M. Hill. 1999. Evaluation of the Quality Assurance Falsification Interview used in teh Census 2000 Dress Rehearsal. In *Proceedings of the Survey Research Method Section.* American Statistical Association pp. 635–640.

Kuriakose, Noble and Michael Robbins. 2016. "Don't get duped: Fraud through duplication in public opinion surveys." *Statistical Journal of the IAOS* 32:283–291.

Li, Jianzhu, J. Michael Brick, Bac Tran and Phyllis Singer. 2011. "Using Statistical Models for Sample Design of a Reinterview Program." *Journal of Official Statistics* 27(3):433–450.

Lin, Jianhua. 1991. "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37(1):145–151.

Madans, Jennifer, Kristen Miller, Aaron Maitland and Gordon Willis. 2011. *Question Evaluation Methods: Contributing to the Science of Data Quality.* Hoboken, NJ: John Wiley & Sons.

McLaughlan, Geoffrey and David Peel. 2000. *Finite Mixture Models.* New York: John Wiley & Sons.

Murphy, Joe, Paul Biemer, Chris Stringer, Rita Thissen, Orin Day and Y. Patrick Hsieh. 2016. "Interviewer falsification: Current and best practices for prevention, detection, and mitigation." *Statistical Journal of the IAOS* 32:313–326.

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

of Survey Quality, The SAGE Handbook. 2016. Another Look at Survey Data Quality. In *The SAGE Handbook of Survey Methodology*, ed. Christof Wolf, Dominique Joye, Tom W. Smith and Yang chih Fu. Thousand Oaks, CA: SAGE pp. 613–629.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.
**URL:** *https://www.R-project.org/*

Rosmansyah, Yusep, Ibnu Santoso, Ariq Bani Hardi, Atina Putri and Sarwono Sutikno. 2019. "Detection of Interviewer Falsification in Statistics Indonesia's Mobile Survey." *International Journal on Electrical Engineering and Informatics* 11(3):474–484.

Sarracino, Francesco and Malgorzata Mikucka. 2017. "Bias and efficiency loss in regression estimates due to duplicated observations: a Monte Carlo Simulation." *Survey Research Methods* 11(1):17–44.

Schnell, Rainer. 1991. “Der Einfluß gefälschter Interviews auf Survey-Ergebnisse.” *Zeitschrift für Soziologie* 20(1):25–35.

Schräpler, Jörg-Peter and Gert G. Wagner. 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, Irwin, Karen Pennie and Jennifer Newbrough. 1988. Interviewer falsification in census bureau surveys. In *Proceedings of the Survey Research Method Section*. American Statistical Association pp. 491–496.

Stan Developers and their Assignees. 2021. *CmdStanR*. R Package Version 0.4.0 (Not published on CRAN).
**URL:** *https://mc-stan.org/cmdstanr/*

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

Team, Stan Development. 2020. *RStan: the R interface to Stan*. R package version 2.21.2.
**URL:** *http://mc-stan.org*

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

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

**Appendix**

# A Forming Agreement Summary Vectors

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

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

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

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

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

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

# B 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.[27]. This would result in estimated responsibilities "heaping" around .5 in a histogram. I suggest three strategies for detecting such issues. First, looking at $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_1$. Second, plotting a histogram of the estimated responsibilities. Third, seeing how similar the two estimated multinomial distributions are using the Jensen-Shannon Distance, the square root of the Jensen-Shannon Divergence. The Jensen-Shannon Distance is bounded by 0 below and 1 above, with 0 indicating that two distributions are the exact same (Lin, 1991; Endres and Schindelin, 2003; Nielsen, 2011; DeDeo et al., 2013).

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

Figure 9 shows the Jensen-Shannon Distance (JSD). We can see that for all parameter combinations it is around .5. Given that the JSD is bounded by 0 and 1, where 0 means identical distributions, this is key evidence that the component distributions of the mixture are sufficiently different.[28]

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

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

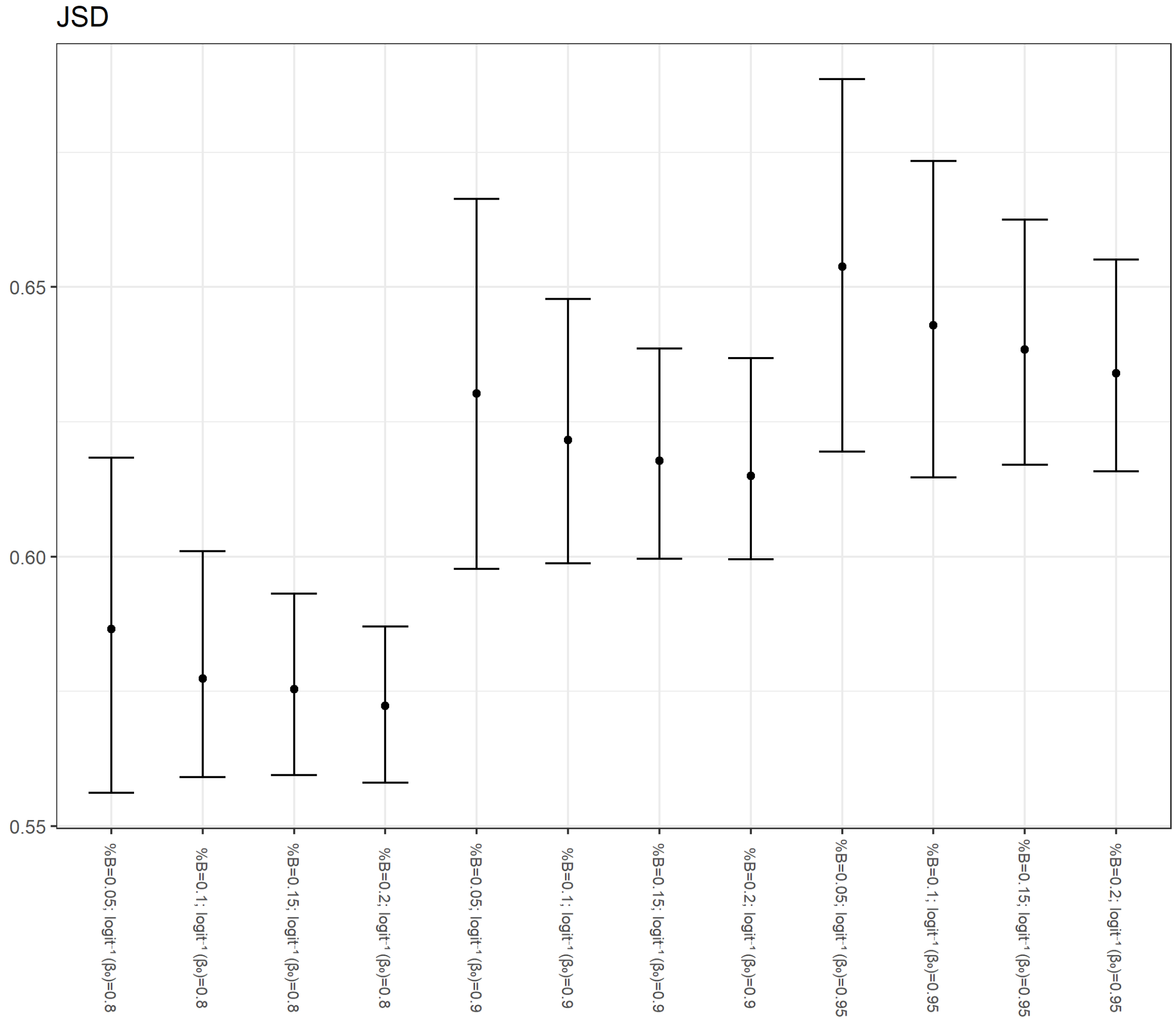


Figure 9: Median of the posterior of the Jensen-Shannon Distance for different parameter combinations with 95% credible intervals.

# C General Simulation Process

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

1. Start with an existing survey data set ($\boldsymbol{R_b}$). This will be the "correct" set.
2. Choose a baseline *proportion of mistake*: $\lambda$ (actually the inverse logit function applied to $\beta_0$).
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 $\beta_e$ from $\mathcal{N}(0, \sigma_{beta_e})$, with some user-determined *variance* (**1**). Calculate each enumerator's match proportion $\lambda_e$ by combining $\beta_0$ and $\beta_e$.
5. Decide which observations will be matches and which will be non-matches. This is random with respect to the observations picked, but the proportion of matches is fixed.
6. Create a copy of the original data set, $\boldsymbol{R_a}$. Replace non-match observations in $\boldsymbol{R_a}$ with observations chosen at random from all match observations (there will then be duplicates).
7. Next, induce small mistakes in match observations $\boldsymbol{R_a}$ via the following steps:

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

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

- [enumerator] a binomial distribution where the number of trials is the maximum decided in the previous step and where $\pi$ 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).[29]
- A discrete distribution where the categories are the numbers 0 to the maximum number of variables possible, with $\pi_i = \frac{(\text{Max. \# of Vars.}+1)-i}{\sum_{i=0}^{\text{Max \# of Vars.}} i+1}$, $i = 0, ..., \text{Max \# of Vars.}$

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

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

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

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

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

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

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

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

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

**Data Preparation**

To prepare the data for the simulation, I drop all observations with NA values in these variables. I also randomly choose thirty-five enumerators from all enumerators with more than 150 observations.[30] This results in 9,973 total observations. For each set of parameters, I then simulate fifty original data–backcheck data pairs. For each simulated original data–backcheck data pair, I then calculate agreement summary vectors for all original data–backcheck data observation pairs in the manner described in Section 3.4 and Appendix A.

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

## C.1 Simulation Model Specification

I then used fit the QualMix model, with the following hyperparameters to the agreement summary vectors:

$$\begin{aligned}
\gamma_i &\sim \text{Multi}(\boldsymbol{\pi}_1)\\
\gamma_i &\sim \text{Multi}(\boldsymbol{\pi}_0)\\
\beta_0 &\sim \mathcal{N}(\mu_{\beta_0}, \sigma_{beta_0})\\
\beta_e &\sim \mathcal{N}(0, \sigma_e)\\
\sigma_e &\sim \text{Gamma}(1, 1)\\
\boldsymbol{\pi}_1 &\sim \text{Dir}(1, 2, 3)\\
\boldsymbol{\pi}_0 &\sim \text{Dir}(3, 2, 1)\\
\mu_{\beta_0} &\sim \mathcal{N}(0, .1)\\
\sigma_{beta_0} &\sim \text{Gamma}(1, 1)
\end{aligned}$$

## C.2 Simulating Effects of Measurement Error

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

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

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

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.[31] 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.

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

# D Results Dropping Failed Backchecks

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

Figure 12 shows updated enumerator quality estimates. Unfortunately there is considerable uncertainty about enumerator quality — this is because there is similar uncertainty about the posterior probability of a match for each respondent, which comes from uncertainty around the two distributions. We see a similar issue with the estimate of overall survey quality, with a 95% credible interval of [0.880, 0.999].

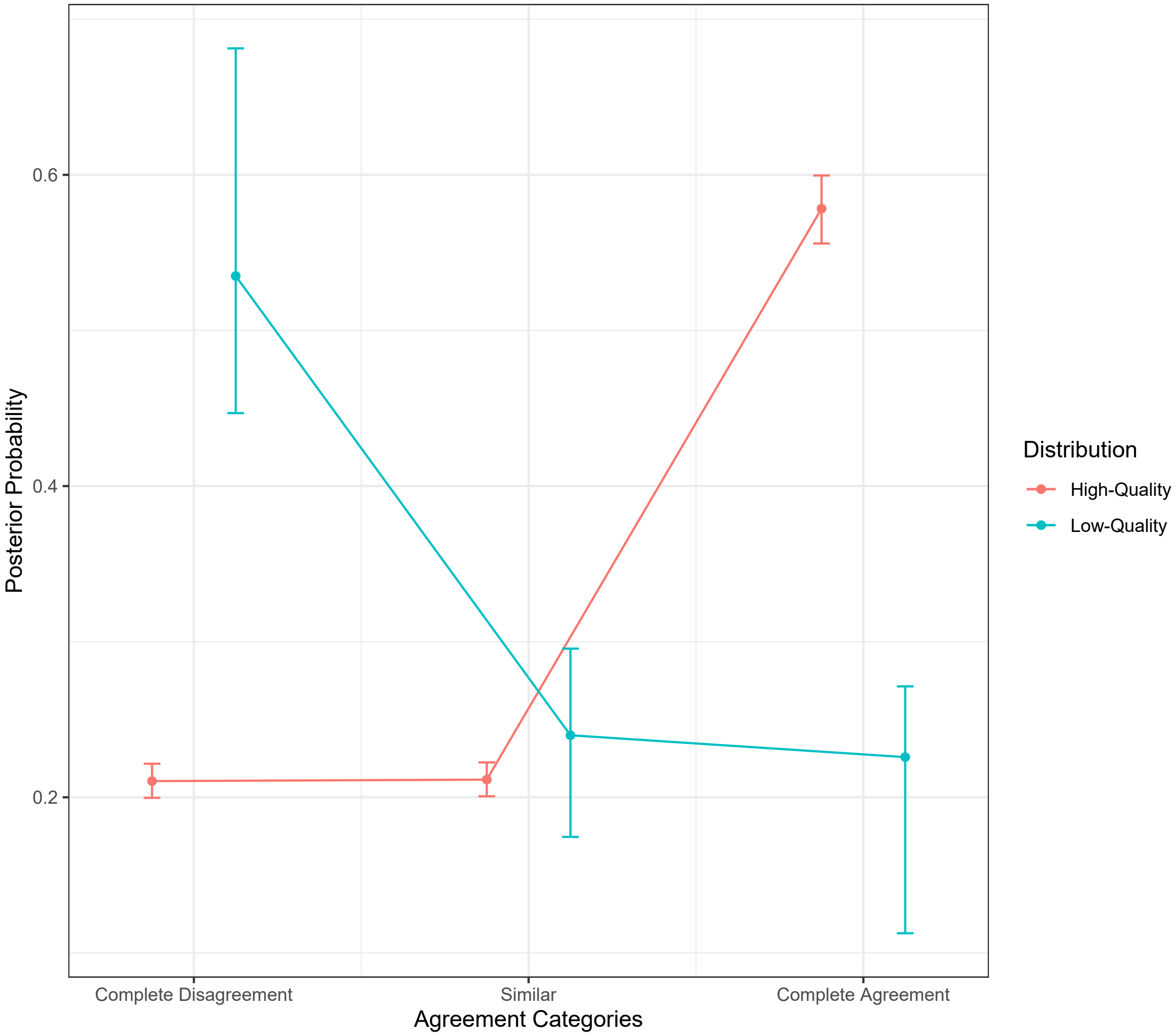


Figure 10: $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_1$ when backcheck pairs where $\nu_0 = 6$ are omitted.

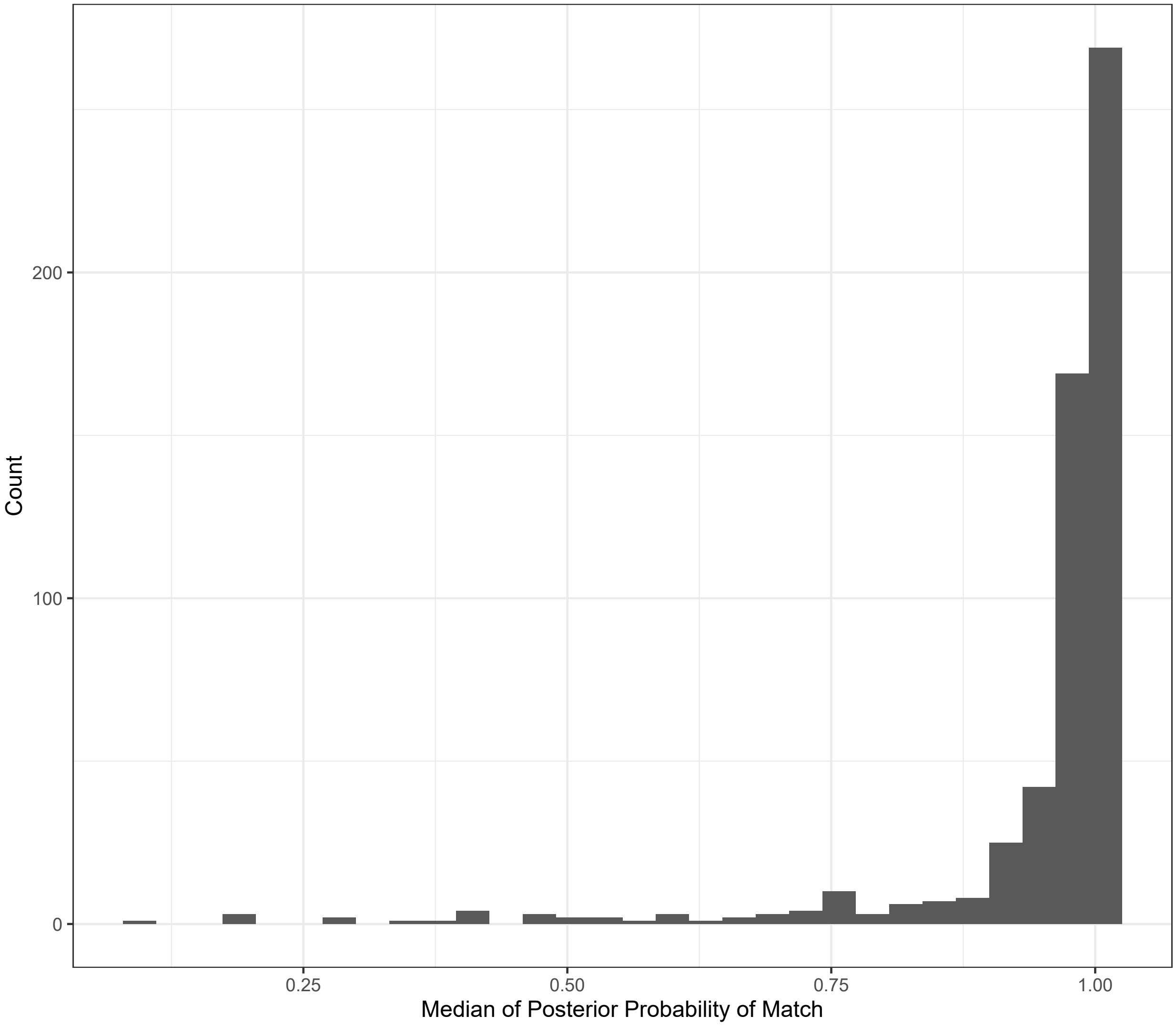


Figure 11: Histogram of the median of the posterior of the posterior probability of a match when backcheck pairs where $\nu_0 = 6$ are omitted.

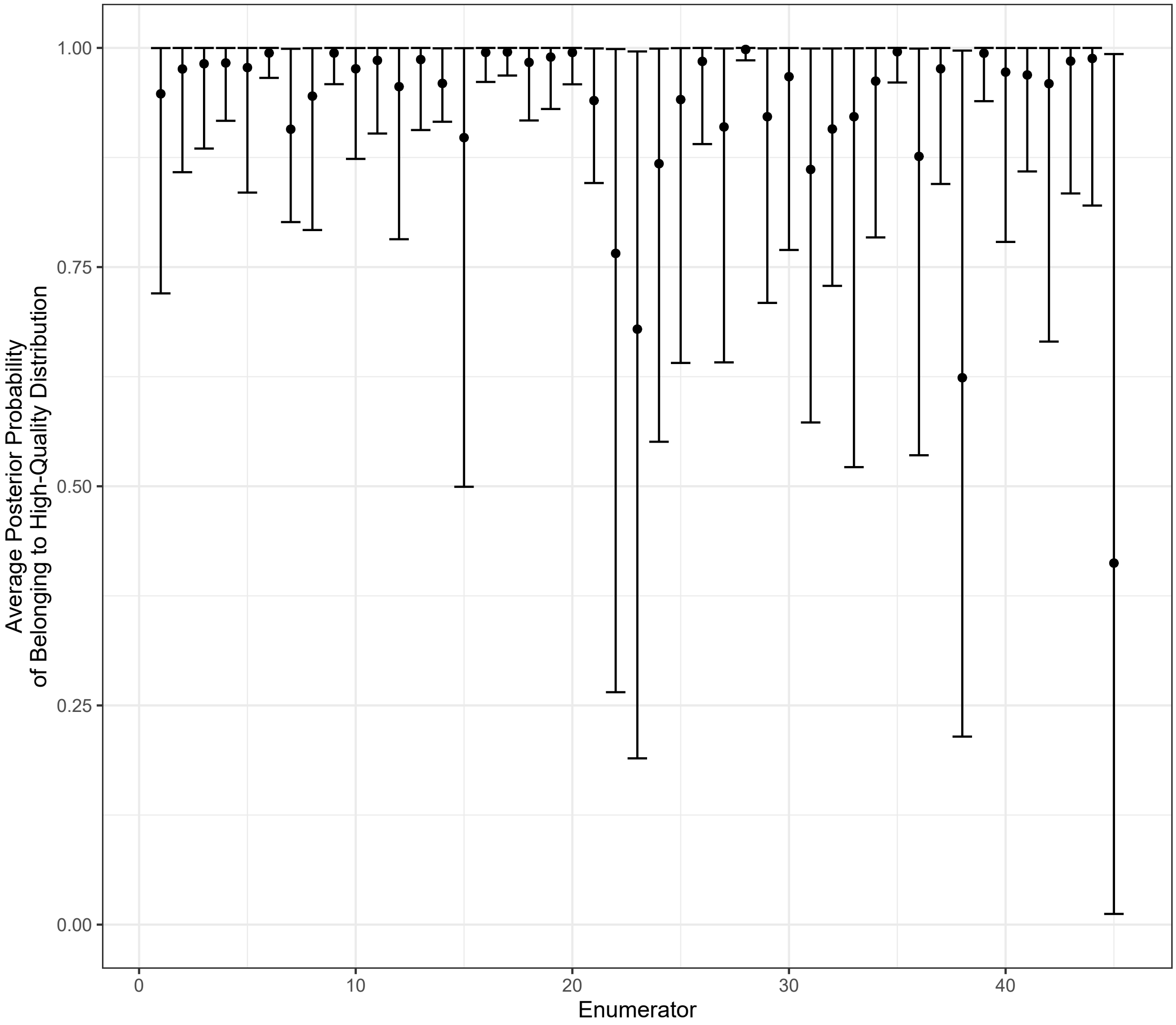


Figure 12: Median of the Posterior of the Average Posterior Probability of a Match for all 45 Enumerators when backcheck pairs where $\nu_0 = 6$ are omitted. Error bars show 95% credible intervals.

# E Receipts and Enumerator Quality

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

## E.1 Initial Modeling Attempt

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

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

I chose the Beta distribution for $Q_e$ because its support is [0, 1] — enumerator 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.

## E.2 Working Model

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

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

I ran each model for 1000 post-warm up iterations on two chains.[32] This results in 1000 model fits, each with 2000 draws from the posteriors of the parameters.[33] I pool the posteriors for $\beta_0$ and $\beta_1$, separately, across all 1000 model fits. This allows me to incorporate uncertainty about enumerator quality into this model.

I then use the pooled posteriors for $\beta_0$ and $\beta_1$ 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.

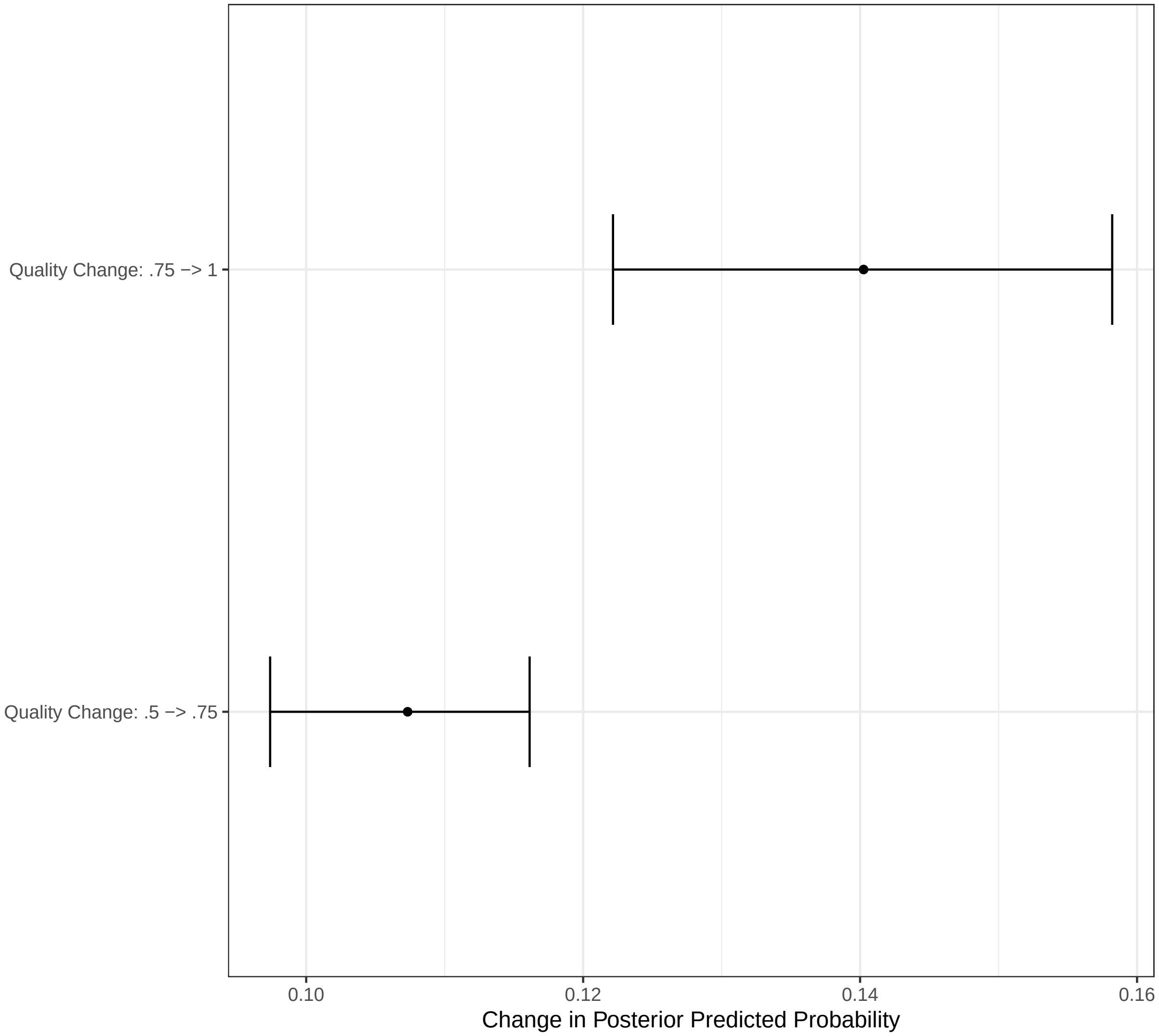
[32] Rhat values were all very close to 1, and ess_bulk and ess_tail were all above 200.

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

## E.3 Results

Figure 13 shows the change in probability of a receipt being reported shown by an enumerator for different changes in enumerator quality.

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


Quality Change: .75 −> 1
Quality Change: .5 −> .75
Change in Posterior Predicted Probability