Bounding Interviewer Effects: Data Quality when Objective Measures are Missing

Andrew Dillon
Northwestern University, Evanston, USA (andrew.dillon@kellogg.northwestern.edu)
Dean Karlan
Northwestern University, Evanston, USA (karlan@northwestern.edu)
Christopher Udry
Northwestern University, Evanston, USA (christopher.udry@ northwestern.edu)

February 5, 2026

Abstract

For most survey-based research, interviewers identify respondents, pose survey questions, record answers, and often navigate complex social interactions to collect data. Using a "backcheck" quality-control process from nine large household surveys, we find strong evidence of interviewer variance in face-to-face and telephone interviews. We reject the null hypothesis that interviewer effects are consistent with classical measurement error. We then extend the Abowd and Stinson (2013) econometric strategy to estimate reliability ratios using prior beliefs on mode accuracy to bound interviewer measurement error. Reliability ratios differ by variable type and confidence in the accuracy of resurvey data. Pooling across all outcomes, we find reliability ratios are 18 percent lower in telephone surveys than face-to-face surveys. We discuss the implications of these reliability ratios as data quality measures.

Acknowledgements: The Global Poverty Research Lab, an academic hub at Northwestern University, generously funded this work. Data collection was conducted by Innovations for Poverty Action (IPA). We are grateful for research assistance from Isabella Contreras, Emma Davies, Navishti Das, and Deepika Nagesh, Yuting Ren, and comments from the participants at the Ninth International Conference on Agricultural Statistics. All errors are ours. This manuscript has been edited for submission to the ICAS x 2026 conference. Full paper with tables is available from the authors upon request.

1. Introduction

Survey-based research is commonplace in economics yet research on the survey process is not. A large empirical literature in economics explores the causes and consequences of measurement error. However, the main characters in the 'data-generating process' that produce survey data, the interviewers, are rarely high-lighted. Interviewers pose survey questions, record responses, and often navigate complex social interactions. Interviewers also skip survey responses, move too quickly through a survey or vary effort in administering questionnaires. Interviewers are people whose strenuous job to collect data in difficult circumstances brings out their best and worst characteristics; all of which affects the data they collect. Data quality is predicated on an interviewer's skills and experience; external conditions like travel fatigue and work load, and inter-viewer supervision by the survey firm. However, defining interviewer data quality is challenging within and between surveys because we rarely observe the true value of any variable.

We address the 'missingness of truth data' by constructing data with repeated observations of the same household with different interviewers. Our re-interview data[1] is a second survey conducted by other interviewers on a random subset of the original households and on a small set of static (i.e., not expected to change over the course of a few weeks) responses. Though we do not have objective data, the repeated observations from re-interview data hold questionnaire design fixed, an alternative source of survey response bias. With repeated observations on the same variables, we construct reliability ratios, a signal to noise ratio, which we interpret as a data quality measure. We draw on insights from the labor economics literature where the variance components of multi-level random effects models are used to improve estimates of reliability ratios, to compare earnings reported from self-reported and administrative data Abowd and Stinson (2013), Kapteyn and Ypma (2007). Our reliability ratios are calculated using the ratio of 'true' variance to total variance. Reliability ratios for the survey data can be calculated for any assumed "truth" model; the accuracy of the re-interview data might be very high, or as low as that of the survey data. The reliability ratios allow us to compare interviewer data quality across countries, across survey modes, between interviewers, and within surveys depending on our priors about which data source is closest to the "truth", rather than assuming a data source is the truth. Due to the standardization of the data collection process at Innovations for Poverty Action, we aggregate nine household surveys to estimate data quality measures with repeated observations with initial interview survey responses and re-interview survey responses.

Before we estimate our measures of data quality, we first establish that the distribution of interviewer quality is meaningfully disparate. First, we estimate interviewer fixed effects by variable and country to illustrate that training programs alone with similar supervision structures does not dampen observable variation in interviewer fixed effects. Second, we find important variation across interviewers, variation that is not consistent with classical measurement error. Treating re-interview responses as "truth", we reject the null hypotheses that the difference between interviewer and re-interview survey responses has mean zero for all interviewers, and that these differences are uncorrelated with the true value of the variable.

Depending on how strongly we weight either the accuracy of the survey or re-interview data, we find

[1]Survey firms like Innovations for Poverty Action often call re-interview data, backcheck or survey monitoring data. We describe our repeated observations as re-interview data, in part, because these observations themselves are not objective truth and are collected to revalidate survey responses rather than only ensure field compliance

considerable variation in reliability ratios for different commonly measured variables (household size, land size, a crop count index and asset index). In our preferred truth model, reliability ratios vary between 0.60 and 0.94. These estimates provide a practical measure of data quality for different variable types and underscore that data quality can vary considerably within the same survey, across surveys and by survey mode. Twelve out of the forty-two comparisons between reliability ratios from different surveys find that data quality is statistically different. With respect to differences by variable, one somewhat surprising example is that of asset index variables. Though relatively standardized in economic surveys, we find that asset indexes collected in face-to-face interviews are twice as reliable as those collected with phone surveys. Disaggregating reliability ratios by interviewer characteristics, we find older interviewers have the highest reliability ratios with interviewers between 31-40 years old and 41-46 years old scoring the highest data quality. Female interviewers also had higher reliability ratios relative to male interviewers pooling across all phone survey observations.

We make two main contributions. First, discussions of measurement error within economics often focus on survey design De Weerdt et al. (2020), respondent biases due to self-reporting Meyer and Mittag (2019) or non-response Heffetz and Reeves (2019). West and Blom (2017) extensively reviews the interviewer effects literature focused on variation in survey response due to interviewer characteristics such as gender or co-ethnicity. They find that much of this literature is based on descriptive studies where causal identification of the interviewer effect could be confounded by non-random respondent assignment. A few studies randomly assign interviewers to respondents and largely find little variation for non-sensitive questions and higher variation for sensitive questions Himelein (2016). For example, Di Maio and Fiala (2020) finds interviewer effects explain 30 percent of response variation for a four category political Likkert scale question. With random assignment of interviewers to respondents, the relative effect of interviewers on survey response can be estimated which has been useful to understand for which types of variables additional interviewer training or ex-post econometric corrections might be used to minimize interviewer effects. However, the relative intent to treat effect does not provide an estimate of interviewer measurement error which requires the 'true' measure of the variable. The survey methodology literature has also focused on the relationship between interviewer characteristics and data quality West and Blom (2017), data quality measures themselves often capture only limited dimensions of measurement error like nonresponse or outlier frequency which themselves are extreme outcomes from the data generating process.

Second, for many economic questions, quantifying measurement error, rather than signing the bias, is integral to assessing bias in parameter estimates and estimating statistical power in experimental designs. Classical measurement error is mean-reverting, but studies including Kapteyn and Ypma (2007), Abowd and Stinson (2013) and Hyslop and Townsend (2020) reject the hypothesis that measurement error is randomly distributed. Concerns about non-random measurement error in self-reported income are often addressed in the empirical labor literature using repeated observations. Abowd and Stinson (2013) estimate measurement error in earnings data collected in self-reported and administrative data sources by arguing that absent "truth" data, the unobservable true value of a variable can be estimated as a weighted average of available measures. When both data sources include measurement error, a reliability ratio can be estimated which provides a measure of data quality and a range of these estimates based on data accuracy by data source. Differences in the literature on estimating reliability ratios depend on which variance components to include in the reliability ratio. We provide a detailed derivation below of differences in estimation, but highlight significant differences do exist between leading methods like Kapteyn and Ypma (2007) and Abowd and Stinson (2013) which are not simple level effects. Our preferred method relies on variance components

directly estimated from a random effects model rather than assumptions about the true variance structure of the measured variable of interest.

2. Data Structure

The data structure for our analysis depends on three data sources: household survey data, re-interview survey data, and interviewer data. We use household survey data collected by Innovations for Poverty Action (IPA) that is 'raw' in that it is the initial download of data from the CAPI or paper form. This is the noisiest form of data that does not capture any corrections made as part of data quality checks. We use household level 'raw' data because it captures the interaction between the interviewer and the respondent before data quality checks are implemented to assess the measurement error in that stage of the data generating process. While there are many low-cost ways to reduce measurement error with data quality checks or corrections in the field, data quality protocols and supervision quality vary across surveys. To avoid potential biases from differences in data quality protocols and to measure unadjusted interviewer measurement error illustrating the full distribution of measurement error, we use the raw form of household survey data.

Household survey data is merged and linked with back-check data which according to IPA protocols is normally a 10% sample randomly selected to be resurveyed. Households selected into the 'back-check' sample receive a short form of the survey to assess the concordance between the repeated observations. Back-check data is collected in the same form as the household survey data and exported in its raw format. Merging the household and back-check data results in repeated observations for multiple household surveys where interviewers vary by survey and data source (household survey and back-check). Finally, we merge interviewer characteristics to household and backcheck data. Each IPA country office collects different interviewer characteristics as part of their hiring process. We take a standardized set of interviewer characteristics that include age, gender, education and experience.

Household survey data was collected using either face-to-face interviews or as phone surveys. Mode differences influence the interaction of interviewers and data quality. A large international literature compares differences between face-to-face surveys and phone surveys (de Leeuw 1992, add others here). For the purposes of our analysis, we estimate data quality by survey mode to further illustrate the importance of a standardized data quality measure.

To determine eligibility, metadata on the project was collected. A household survey wave is eligible to be included if re-interview survey data existed and households selected for re-interview were randomly assigned; and information on interviewer characteristics was available. An initial set of 16 projects were included.

Of these, projects were declared ineligible due to questions relating to the randomization of interviewers to respondents or the completeness or re-interview and household data. Our final data set includes 9 household surveys with matched household, re-interview responses and interviewer data. Face-to-face surveys include recent surveys from Burkina Faso, Mali and the Philippines. Phone surveys are included from Burkina Faso, Cote d'Ivoire, Mali, Philippines and Rwanda

3. Estimation

Estimating interviewer measurement error is difficult without true observations of the data. We approach this challenge in three steps. First, we observe across surveys and variable types a pattern of interviewer fixed effects which suggests that an interviewer's effect creates a bias which differs between interviewers and variable types. We present interviewer fixed effects estimates by regressing the z-score of the difference between the survey variable and the backcheck variable on interviewer fixed effects and controlling for the main respondent characteristics. We plot the interviewer fixed effects in Figures 1-8 and overlay the normal distribution for comparison. We expect interviewer fixed effects to be normally distributed. When they are not normally distributed, this is consistent with non-classical measurement errors.

Second, we test whether interviewer biases are consistent with classical measurement error. The results of the test lend itself to a practical recommendation for study design. If tests are consistent with classical measurement error, increasing sample size will address the 'noise' caused by interviewer measurement error. If tests are not consistent with classical measurement error, then interviewer measurement error biases parameter estimates. In this case, study design choices related to questionnaire design to reduce measurement error or recruitment and training to improve interviewer quality ex-ante could minimize parameter estimate biases. After establishing variability across interviewers and differences by survey outcomes, we test whether interviewer variance is consistent with classical measurement error. Hyslop and Townsend (2020) uses different surveys to estimate whether variation in earnings data is classical. Here for each outcome we adapt this test under the strictest assumption that our backcheck data accurately measures the true outcome ($Y_B = Y^*$) and the survey data are reported with classical measurement error ($Y_S = Y^* + \epsilon$, $\epsilon \sim iid$ (μ_ϵ, σ^2_ϵ)). The implication of these assumptions and data structure is that a regression of Y_B on Y_S provides a testable hypothesis that the regression coefficient will be one, while the regression of Y_S on Y_B will be attenuated towards zero. Rejection of this null hypothesis provides evidence that the variation in the data are not consistent with classical measurement error.

Lastly, we estimate variance components from a random effects model to construct reliability ratios which quantify data quality given different beliefs about survey and resurvey mode accuracy. Our data structure provides multiple observations of household, h, using data sources (survey and back-check) which are administered by different interviewers. Due to random assignment within data source, interviewers and data sources are independent at each household-level observation. We estimate measurement error from inter-viewers for a set of dependent variables (z-scores of household size, land size, an asset index, and a crop count variable) which vary by household, data source s, and interviewer i. Observable respondent characteristics, x, are included as independent variables with a set of random effects for interviewers, households, and data sources. Equation 1 is the regression specification:

$$Y_{his} = \beta_0 + \beta_1 X_{his} + u_i + v_h + w_s + \epsilon_{his} \quad (1)$$

Here, $u_i = \alpha_i + \eta_I$ where $i = 1, \ldots, I$ interviewers, $v_h = \kappa_h + \eta_h$ where $h = 1, \ldots, H$ households, and $w_s = \gamma_s + \eta_s$ where $s = 1, \ldots, S$ sources, represent the random effects which are iid. $\eta_i \sim N(0, \sigma_I)$, $\eta_h \sim N(0, \sigma_H)$, $\eta_s \sim N(0, \sigma_s)$, and $\eta_{his} \sim N(0, \sigma_\epsilon)$.

With the variances from the random effects model, we can estimate the interviewer intraclass correlation coefficient and a reliability ratio as interviewer data quality measures. The intraclass correlation coefficient is expressed as the variance of the interviewer component of the intercept in

the random effects model and the variance of the residual. This provides a measure of the homogeneity of the responses obtained by the interviewer as the proportion of the explained variance.

$$\rho_{int} = \frac{\rho_{\mu}^2}{\rho_{\mu}^2 + \rho_{\bar{e}}^2} \quad (2)$$

The reliability ratio provides a 'signal to noise' ratio of the variance of errors relative to the overall variance in the data. A limitation of this derivation of the reliability ratio is that it is static for the population of interest and is independent of potential information about the accuracy of either data source. Kapteyn and Ypma (2007) and Abowd and Stinson (2013) use two equivalent derivations of the reliability ratio that rely on different variance components. Defining the variance of the observed variable as equivalent to the sum of the true variance and measurement error variance ($\sigma_X{}^2 = \sigma_T{}^2 + \sigma_E{}^2$), then Kapteyn and Ypma (2007) preferred reliability on the left is equivalent to Abowd and Stinson (2013)'s preferred reliability ratio on the right generally.

$$\frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \quad (3)$$

Because we estimate the interviewer variance component of the observed variable and the error variance directly with our data, the Abowd and Stinson (2013) reliability ratio does not potentially compound errors of variance component estimation by requiring the calculation of the true interviewer variance in the reliability ratio. In our empirical estimates, we estimate reliability ratios using both methods. An advantage of either the Kapteyn and Ypma (2007) or Abowd and Stinson (2013) reliability ratio estimate is that they can vary by the accuracy or 'truthfulness' we might assume about one of our data sources. We prefer estimating reliability ratios using the Abowd and Stinson (2013) approach as variance components of the error term and observable characteristics (X) are easily estimated using equation 1. Kapteyn and Ypma (2007)'s approach requires assumptions about the variance of the truth.

4. Results

4.1 Descriptive differences and correlations by data source

Table 1 and 2 provides descriptive statistics of household and backcheck variables for face-to-face and phone surveys respectively. We report the mean and standard deviation for each household survey and backcheck variable (columns 2 and 3), the difference between household and backcheck observations (column 4), a reliability ratio defined as $Y_b / (Y_b + Y_{s-bc})$ (column 5), and the correlation between the mean difference of the variable and the backcheck data (column 6).

Differences between survey and backcheck variables are not uniformly distributed in face-to-face surveys summarized in Table 1. In the Philippines sample, backcheck reports of the household size are 0.5 members higher, while in Mali, the household survey report of household size is 0.5 members higher. Correlations between survey and backcheck reports of household size vary between -0.42 and 0.15. Negative correlations indicate the association between survey and backcheck data when the resurvey information reports a higher value of the variable. The range of correlations for the asset index (-0.43 and 0.02), land size (-0.27 and -0.11), and crop count (-0.61 and -0.42) provide evidence that the

backcheck mode reports higher values relative to the survey mode when the value of the variable is higher.

Phone survey differences in survey and backcheck data are not uniformally biased as reported in Table 2. We find higher reports of household size in surveys relative to backchecks in Mali and Philippines surveys. Pooling all phone survey observations, we also find higher survey responses for asset indices than in backcheck observations. Correlations between the survey and backcheck data are negative, varying between -0.9 and -0.62.

In Table 1 (face-to-face surveys) and Table 2 (phone surveys), we calculate simple reliability ratios found in column 5 of both tables. Across countries, face-to-face survey reliability ratios range from 0.45 to 0.88, while phone survey reliability ratios range from 0.34 to 0.96. Pooled across all surveys, household size variables had high reliability ratios measured in both face-to-face and phone surveys, while there was a twofold difference in reliability ratios for asset indices measured with face-to-face surveys (0.68) relative to phone surveys (0.34). As asset indices are a common outcome of studies focused on poverty, measuring asset indices with interviewers over the phone resulted in considerably lower data quality.

Figure 1 presents interviewer fixed effects by country sample and mode for variables household size, asset index, land size and crop count variables. We estimate interviewer fixed effects on the mismatch between survey and backcheck responses. In Figure 1, we order these interviewer fixed effects with country color-coded lines to illustrate variation in interviewer fixed effects on response mismatch. Interviewer fixed effects can be negative or positive indicating under or over-reporting relative to the backcheck data. These figures illustrate that interviewer fixed effects are much larger for face-to-face surveys relative to phone surveys for example (Figure 1, Panel A), but smaller when assets are measured in face-to-face surveys relative to phone surveys (Figure 1, Panel B). Comparing between household size and asset index interviewer fixed effects, we see that generally interviewer fixed effects are larger magnitude for household size in comparison to asset index variables.

In Figure 2, we present the kernel density of the interviewer fixed effects and underlying distribution plotted relative to the kernel density of the normal distribution in red. Both Figures 1 and 2 provide some context for the variation we observe of interviewer effects which differ by mode and variable. For example, we observe a flatter density for phone survey asset index variables relative to face-to face survey interviewer fixed effects. This comparison illustrates larger over and under-reporting of asset index variables in phone surveys due to interviewers.

4.2 Tests for Non-Classical Measurement Error

Variation between interviewers is not sufficient evidence that interviewer fixed effects represent interviewer-specific measurement error. First, interviewers are assigned different respondents, so variations in interviewer fixed effects may correspond to sampling variation. Second, interviewer associated measurement error biases point estimates only if non-random. If interviewer fixed effects represent classical measurement error, then increasing interviewers or sample size addresses attenuation bias from classical measurement error. We test for classical measurement error, presenting results in Table 3 for face-to-face surveys and Table 4 for phone surveys. In panel A in each table, we test the null hypothesis, following Hyslop and Imbens (2020), that the coefficient of the regression of Y_R on Y_S will be close to 1. In panel B in each table, we present the regression results when we

regress Y_S on Y_R. Coefficient estimates are predicted to be attenuated towards zero if consistent with the classical measurement error null hypothesis.

For face-to-face surveys, Table 3, panel A reports a few coefficients which are close to 1, but we reject the null hypothesis that the coefficient is one for each variable. In Table 3, panel B we test whether we can reject the null hypothesis that coefficients are equal to zero. We are unable to reject the null hypothesis for any variable across the four country surveys. We conclude from the null hypothesis test coefficients and the observation that coefficient ratios are generally below 1 that variation in the face-to-face survey data is not consistent with classical measurement error.

We repeat the same hypothesis tests for the phone survey data in Table 4. In Table 4, panel A, we test the null hypothesis that coefficients are one. This null hypothesis is rejected for all variables with the exception of the household size variable in Rwanda. In Table 4, panel B, we test the null hypothesis that coefficients are attenuated to zero. This null hypothesis is rejected in five out of the nine hypothesis tests conducted. Both hypothesis tests are rejected for the household size variable in Rwanda which provides evidence consistent with classical measurement error. For the other hypothesis tests where we were not able to reject both hypothesis tests, we cannot draw a definitive conclusion about classical measurement error.

4.3 Reliability Ratios

Interviewer effects may be contributing to non-classical measurement error in our data. We estimate reliability ratios based on Abowd and Stinson (2013) and Kapteyn and Ypma (2007), as a measure of interviewer effects and data quality. We do not have an unbiased measure of each variable. To address this issue, we estimate a range of interviewer effect bounds based on prior weights of confidence in the survey and re-interview data. From the survey data and random effects model described above, we estimate the variance of the outcome by data source (survey or re-interview), the variance of the signal, and the variance of the measurement error by data source. The variance of the signal and variance of the measurement error depend on the truth model weights.

Table 5 (face-to face surveys) and Table 6 (phone surveys) present the reliability ratios where the assumed 'truth' model bounds the confidence we have in either the survey or re-survey data. We focus on the range of reliability ratios for the most likely cases in the truth model where either the re-interview data is measured with small error relative to the survey data (0.1, 0.9) or the alternative where both data sources have the same likelihood of error (0.5, 0.5). The reliability ratios of most interest are those of survey data sources as measures of data quality. In Table 5, face-to-face survey reliability ratios calculated using Abowd and Stinson's formulation are highest for household size variables (0.819 and 0.943), land size (0.772 and 0.930), and asset index (0.701 and 0.907). The reliability ratio for the crop count variable is noticeably different (0.595 and 0.874) compared to our other variables, though we cannot statistically test these differences. When we pool normalized variable observations across the data, we find a reliability ratio of 0.739 when we have more confidence in resurvey data and 0.919 when truth model beliefs are equally weighted between the survey and resurvey observations. Comparing reliability ratios between the Abowd and Stinson method and Kapteyn and Ypma method when we believe that the re-interview data are equally or of higher quality, we generally find higher survey reliability ratios when we estimate the reliability ratio using Kapteyn and Ypma's method.

Figure 3 compares differences between face-to-face and phone survey reliability ratios. Pooling all coun-tries, we observe no differences in reliability ratios pool all outcomes or estimating reliability ratios for household size, a common variable collected across all surveys. Disaggregation across countries reveals much larger confidence intervals for the difference between face-to-face and phone surveys, but none of the reliability ratio differences are statistically different from zero.

Comparing across countries and within mode with a wider set of outcome variables, we do find reliability ratio differences (Figure 4) which suggests differences in data quality by survey. Twelve out of the forty-two comparisons between reliability ratios from different surveys find that data quality is statistically different. Comparisons in reliability ratios are potentially useful to monitor data quality, but also have applications to selecting samples for meta-analysis (more here or in conclusion).

Figures 5-7 estimate reliability ratios pooled across all phone surveys based on interviewer characteristics including age, gender, and survey experience. We find older interviewers have the highest reliability ratios with interviewers between 31-40 years old and 41-46 years old scoring the highest data quality. Female interviewers also had higher reliability ratios relative to male interviewers pooling across all phone survey observations. Finally, interviewers with less experience (fewer than 8 projects) had the highest reliability ratios relative to more experienced interviewers. This may be because interviewers with less experience might be more easily trainable or have furnish more effort on the job relative to their more experienced peers.

5. Conclusion

In empirical work, we often take data as given. When we collect data, researchers focus on how alternative study designs might increase statistical power of treatment effects or improve the measurement of key outcome variables. While an emerging literature on questionnaire design and survey experiments continues to attract the attention of empirical economists, the data-generating process itself and its main characters, the interviewers, are often absent. A compelling reason that interviewer effects might be overlooked is that data quality measures often face a missing objective data problem. We address objective missing objective data by relying on between and within variances of repeated observations on the same household outcomes with different interviewers. By using variance components to create reliability ratios and providing bounds of reliability ratios over a researcher's assumptions about the truth model (how accurate the survey versus re-survey data might be), we can interpret reliability ratios as a data quality measure.

We find reliability ratios vary within surveys amongst different variable types and between surveys for the same variables. This observation has important implications for how we interpret the external validity of survey methodology experiments. For example, we find large differences in reliability ratios for asset index variables, a variable that is relatively standardized in economic surveys. Despite this standardization, we find that reliability ratios of interviewer effects for asset index variables administered with phone surveys are twice as low as those administered with face-to-face interviews. By comparing between countries, we also note variation in reliability ratios for the same variables.

Using reliability ratios as a data quality measure has implications for research design, data collection, and policy. As a researcher, a common way to improve data precision is to either increase sample size or improve the measurement of key outcomes with better questionnaire design to reduce

bias. By bounding interviewer effects, we illustrate that field supervision and interviewer training are two alternative survey design choices where researchers might benefit from better estimates of bias-cost tradeoffs. As more data is generated, research firms use a variety of field supervision practices to ensure data quality, but often do not have comparable data quality measures. Data collection firms would benefit from conducting systematic training and monitoring interventions to assess bias-cost tradeoffs where reliability ratios are the outcomes. The benefit of such a research agenda is the comparability across studies. Lastly, this comparability across studies has important policy relevance for bounding data quality effects across data sources within countries over time. Which data to trust is often a dilemma for policymakers and national statistical offices that field multiple surveys. Reliability ratios can give some indication of which data sources or which variables within a survey are higher quality.

References

Abowd, J. M. and M. H. Stinson (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *The Review of Economics and Statistics 95* (5), 1451–67.

De Weerdt, J., J. Gibson, and K. Beegle (2020). What can we learn from experimenting with survey methods? *Annual Review of Resource Economics 12* (1), 431–47.

Di Maio, M. and N. Fiala (2020). Be wary of those who ask: A randomized experiment on the size and determinants of the enumerator effect. *World Bank Economic Review 34* (3), 654–69.

Heffetz, O. and D. B. Reeves (2019). Difficulty of reaching respondents and nonresponse bias: Evidence from large government surveys. *The Review of Economics and Statistics 101* (1), 176–91.

Himelein, K. (2016). Interviewer effects in subjective survey questions: Evidence from timor-leste. *International Journal of Public Opinion Research 28* (4), 511–33.

Hyslop, D. R. and W. Townsend (2020). Earnings dynamics and measurement error in matched survey and administrative data. *Journal of Business Economic Statistics 38* (2), 457–69.

Kapteyn, A. and J. Y. Ypma (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics 25* (3), 513–51.

Meyer, B. D. and N. Mittag (2019). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness, and holes in the safety net. *American Economic Journal: Applied Economics 11* (2), 176–204.

West, B. T. and A. G. Blom (2017). "explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology 5* (2), 175–211