Measurement error in survey data: what is it and why does it matter

Technical Brief / June 2022

laterite
DATA | RESEARCH | ADVISORY

When did you last see an empirical study in the social sciences that reported on measurement error in the context of a quantitative survey and how these errors might affect the research findings?

Measurement errors are everywhere in survey data. They are unavoidable, however good the survey instrument is and however well trained and competent enumerator teams are. Errors can result from survey questions that invite different interpretations, from questions that are time-sensitive, from enumerator entry or judgment errors, or simply from respondent fatigue and approximations.

Measurement errors are so ubiquitous in human activity generally speaking that Kahneman et al. refer to it as a “flaw in human judgment” in their most recent book, *Noise*. The book is full of real-life examples of how well-trained physicians disagree on the diagnosis of patients, or how different judges can come to diverging sentence decisions for similar crimes. In the case of medical misdiagnosis or a prison sentence, the effects of error are hard to miss. But what happens in the realm of econometrics?

Measurement error: definition and implications

Measurement error is a well-researched problem in econometrics, but unfortunately it hasn’t made its way from the literature to applied research. We all analyze our data as if what we “observe” in a survey is indeed the ground truth, which it is not.

What is measurement error? An error in measurement can be defined as the difference between the value recorded during a measurement, or data collection, and what is the actual truth. Examples:

- the difference between the reading on a scale and the true weight of a child (**continuous variable**)
- an enumerator classifies a given farming practice as adopted, when in reality it was not (**categorical variable**)

Errors can be due to, for example, data coding errors, self-reporting, single measurements of variable longitudinal processes, or imprecise measurement instruments (Brakenhoff et al., 2018).

Measurement error vs misclassification

Measurement error is defined slightly differently for continuous variables and for categorical/binary variables, where we talk about misclassification.

To understand why the two are slightly different, imagine if the distribution of error were completely random. Then for a continuous variable we might be able to describe error as the true value plus a positive or negative error term. Under randomness, this error term would be completely independent of the true value of the variable.

That is impossible in the case of binary, categorical or discrete variables (or any variable that is bounded for that matter). If the error for each observation in a binary variable were defined as: Error = True value - Observed value. Then if the observed value is 0, the error term can only take the values 0 if there is no error or 1 if there is an error; if the observed value is 1, then the error term can only take the values 0 or -1. This implies that errors will always be correlated with the true value of the binary variable we are trying to measure.

For more on this, see this excellent blog post.

In very broad strokes, the econometrics literature distinguishes between:

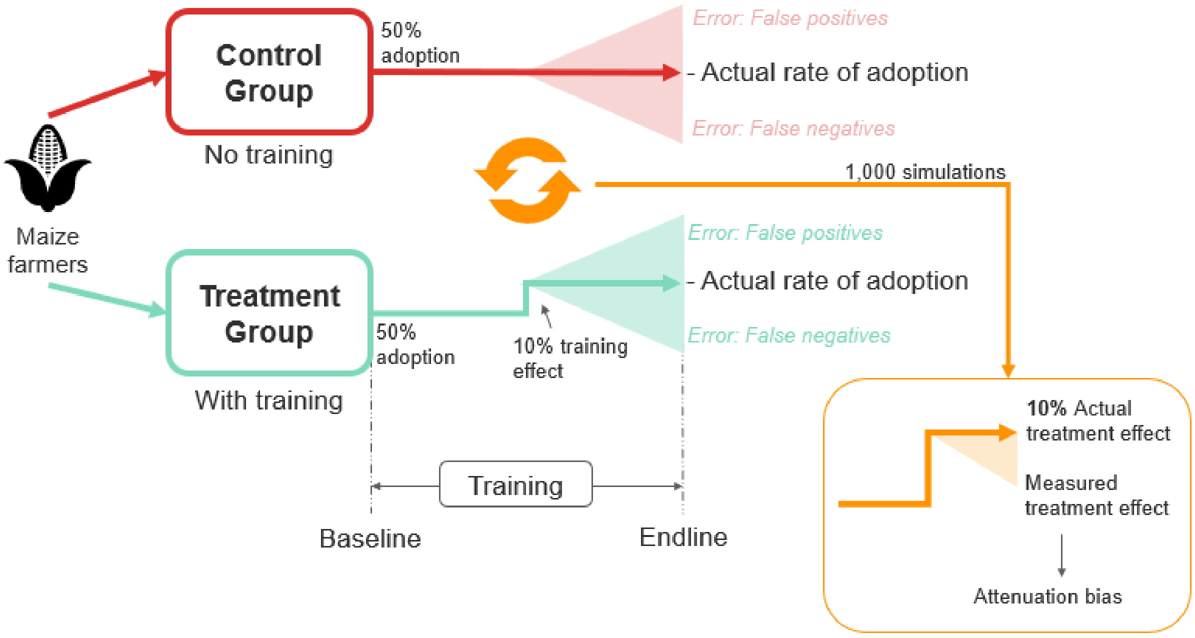
- **random measurement errors**, which are not correlated with any respondent characteristics and can be an under or over-estimation of a variable;
- **systematic measurement errors**, which are either under- or over-estimations of a variable; or
- **differential errors**, which correlate with one or more respondent characteristics.

In the context of a regression, measurement errors or misclassification can affect the dependent variable and/or one or more covariates.

The literature has established that errors or misclassification can result in a ‘triple whammy’ of bias, imprecision, and an attenuation, that can make it difficult to establish relationships between variables (Carroll et al., 2006). In the example below we show that it can also lead to a fourth issue, that, to the best of our knowledge, does not seem to be addressed in the literature: the effect of misclassification on the validity of experimental results.

How misclassification undermines the assumption of parallel trends

An example always helps. Let’s create an imaginary experiment, where a sample of maize farmers are randomly allocated to two groups: a treatment group, where farmers receive some training on good maize farming practices, and a control group. We assume that at baseline 50% of maize farmers in the treatment and control groups adopt a specific farming practice, let’s say mulching. We also assume that the treatment leads to a 10 percentage point increase in the adoption of mulching between baseline and endline. Since we are the creators of this synthetic dataset, we can create the true values of adoption and we can also determine what we observe at baseline and endline.



We assume that during field observations enumerators struggle to determine whether a farmer has mulched their farm or not (which is, from our own field experience, a realistic scenario). Sometimes the enumerators misclassify mulching as non-mulching, a false-negative, and sometimes they might think a farmer has mulched, when in fact they haven't, in which case we talk about a false-positive. We assume that the probability of both types of errors is different.

We can now take this hypothetical scenario and run 1,000 simulations, each time with a different combination of false-positive and false-negative rates, assuming these rates are consistent at baseline and endline. This will give us 1,000 baseline/endline datasets which we can use to measure treatment effect.

For this thought experiment, we will measure the treatment effect using two separate approaches:

- **a difference-in-difference approach**; and
- **a difference-with-a-lag approach**, by regressing the change in adoption over time against adoption at baseline and the treatment status.

We find that in the presence of misclassification (irrespective of whether false positive errors or false negative errors dominate):

- We obtain an erroneous measure of adoption at baseline and endline (which means the constants in the regressions are wrong)
- We almost always underestimate the treatment effect (this is called attenuation bias). Using the actual values of the data, we estimate a treatment effect of 9.8 percentage

points, versus 6.6 percentage points using the difference-in-difference method based on observed data;

- We find that the attenuation bias increases along with the average error rate, estimated as the proportion of measurements that were misclassified (see figure 2);
- But interestingly, we obtain different estimates of the treatment effect using the difference-in-difference and difference-with-a-lag approach. The difference with a lag approach almost always yields estimates that are closer to the real value of the treatment effect (see figure 1).

Figure 1 - Distribution of treatment effects using the two models, 1,000 simulations

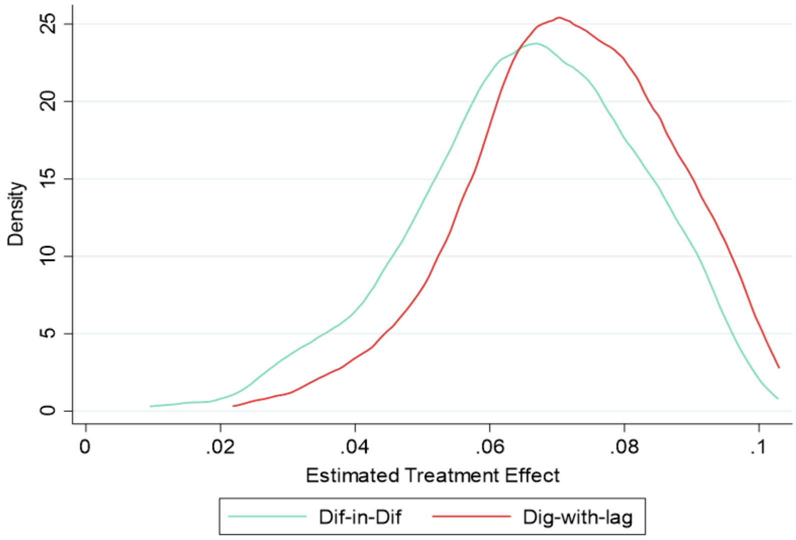
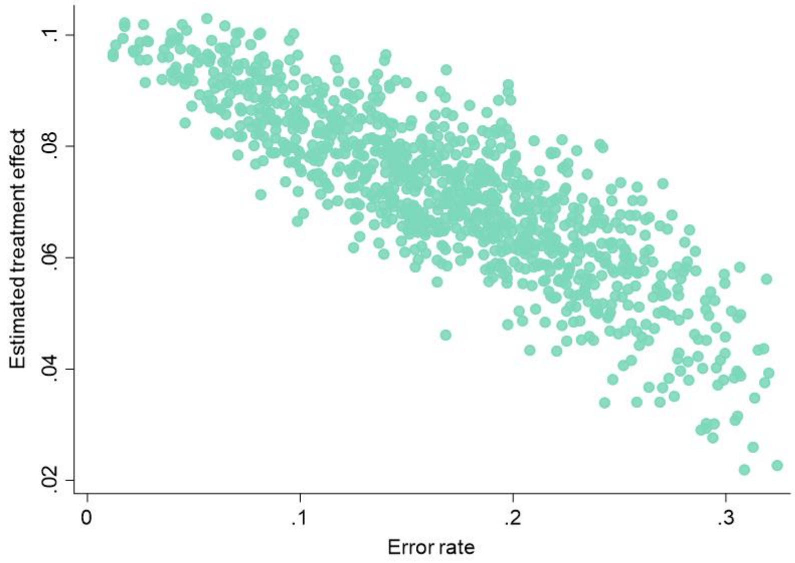


Figure 2 - Association between error rate and estimated treatment effect, using the difference with a lag, 1,000 simulations



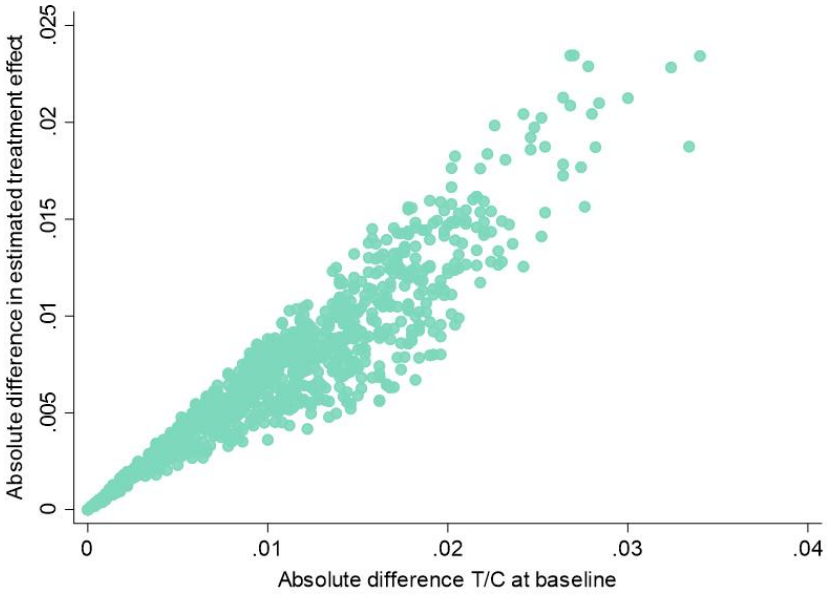
In this context, difference-in-difference estimates seem to be more susceptible to measurement errors. The most probable explanation is that the assumption of parallel trends does not hold anymore. This assumption, which is so central to difference-in-difference estimation, states that in the absence of any intervention, the treatment and control groups would have evolved in parallel. This assumption cannot hold in the presence of misclassification, if simultaneously there is:

- a difference in adoption rates between the treatment and control groups at baseline; and
- a difference in the probability of false positives and false negatives.

Why? Because in the absence of any intervention, the observed trajectory of both groups would have been entirely determined by misclassification rates (assuming no other changes); if the starting point of those two groups is different, then their observed trajectories will be as well.

We can see in figure 3 that, indeed, the difference in adoption rates at baseline between the treatment and control groups, is strongly predictive of the difference in the estimated treatment effect using the difference-in-difference method compared to the difference-with-a-lag approach, which does not assume parallel trends.

Figure 3 - Difference in the estimated treatment effects vs difference in adoption at baseline between treatment and control, 1,000 simulations



In summary, misclassification, and measurement errors more broadly, can severely compromise well-crafted experiments and surveys. What's more, we're often unaware of it.

Should you worry about the effects of misclassification errors?

The answer is absolutely, especially if:

- You are collecting data on things that are difficult to measure (e.g., income, productivity)
- Your survey instruments ask questions that rely on recall or are time-sensitive
- You are working with data that depends on the judgment of enumerators;
- You are collecting data based on observations;
- You are asking the same question about a household, from different household members in different households (for example a woman in one household, a man in another, because they can perceive things differently)
- You think the assumption of parallel trends holds in the real world.

The good news: there are econometric ways to deal with measurement error and misclassification; the bad news: it's really not for the faint hearted and these methods also come with many limitations.

Our advice to limit measurement error in survey data

Here are a couple of best practices we learned from our experience:

- Invest in testing & improving your instruments ahead of time: a pilot survey will expose weaknesses to be corrected.
- Use measures of inter-rater reliability to understand where measurement errors might be coming from: it can be a specific question that needs to be improved, or from an enumerator that needs some extra training.
- Use back-checks on key variables of interest to estimate error rates: once you have an estimate of error rates, bias correction-methods become much more accessible (don't just use back-checks to check on enumerators).
- As tempting as they are, avoid asking questions about things that are so incredibly difficult to measure, like income and farm productivity. Try to use proxies that are more reliable.
- Reflect on the type of measurement errors you might be facing and acknowledge them in your paper/report.
- Run some simulations by injecting errors into your data and see what happens.
- Read up on the measurement-error literature and try out some of the solutions that are proposed.

Measurement error literature: a starter's guide

Barnett, A. G. (2004). "Regression to the mean: what it is and how to deal with it." *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Bollinger, C. R. (1996). "Bounding mean regressions when a binary regressor is mismeasured." *Journal of Econometrics*, *73*(2), 387–399. https://doi.org/10.1016/S0304-4076(95)01730-5

Bollinger, C. R. (2001). "Response Error and the Union Wage Differential." *Southern Economic Journal*, *68*(1), 60. https://doi.org/10.2307/1061511

Bollinger, C. R., & David, M. H. (1997). "Modeling Discrete Choice with Response Error: Food Stamp Participation." *Journal of the American Statistical Association*, *92*(439), 827–835. https://doi.org/10.1080/01621459.1997.10474038

Bollinger, C. R., & van Hasselt, M. (2017). "Bayesian moment-based inference in a regression model with misclassification error." *Journal of Econometrics*, *200*(2), 282–294. https://doi.org/10.1016/j.jeconom.2017.06.011

Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G. M., Groenwold, R. H. H., & van Smeden, M. (2018). "Measurement error is often neglected in medical literature: a systematic review." *Journal of Clinical Epidemiology*, *98*, 89–97. https://doi.org/10.1016/j.jclinepi.2018.02.023

Bruckmeier, K., Riphahn, R. T., & Wiemers, J. (2021). "Misreporting of program take-up in survey data and its consequences for measuring non-take-up: new evidence from linked administrative and survey data." *Empirical Economics*, *61*(3), 1567–1616. https://doi.org/10.1007/s00181-020-01921-4

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models* (0 ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420010138

DiTraglia, F. J., & García-Jimeno, C. (2019). "Identifying the effect of a mis-classified, binary, endogenous regressor." *Journal of Econometrics*, *209*(2), 376–390. https://doi.org/10.1016/j.jeconom.2019.01.007

Feng, S., & Hu, Y. (2013). "Misclassification Errors and the Underestimation of the US Unemployment Rate." *American Economic Review*, *103*(2), 1054–1070. https://doi.org/10.1257/aer.103.2.1054

González Chapela, J. (2022). "A Binary Choice Model with Sample Selection and Covariate-Related Misclassification." *Econometrics*, *10*(2), 13. https://doi.org/10.3390/econometrics10020013

Griliches, Z., & Hausman, J. (1984). *Errors in Variables in Panel Data* (No. t0037; p. t0037). National Bureau of Economic Research. https://doi.org/10.3386/t0037

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). "Misclassification of the dependent variable in a discrete-response setting." *Journal of Econometrics*, *87*(2), 239–269. https://doi.org/10.1016/S0304-4076(98)00015-3

Hug, S. (2010). "The Effect of Misclassifications in Probit Models: Monte Carlo Simulations and Applications." *Political Analysis*, *18*(1), 78–102. https://doi.org/10.1093/pan/mpp033

Innes, G. K., Bhondoekhan, F., Lau, B., Gross, A. L., Ng, D. K., & Abraham, A. G. (2022). "The Measurement Error Elephant in the Room: Challenges and Solutions to Measurement Error in Epidemiology." *Epidemiologic Reviews*, *43*(1), 94–105. https://doi.org/10.1093/epirev/mxab011

Lewbel, A. (2000). "identification of the binary choice model with misclassification." *Econometric Theory*, *16*(4), 603–609. https://doi.org/10.1017/S0266466600164060

McHugh, M. L. (2012). "Interrater reliability: the kappa statistic." *Biochemia Medica*, 276–282. https://doi.org/10.11613/BM.2012.031

Meyer, B. D., & Mittag, N. (2017). "Misclassification in binary choice models." *Journal of Econometrics*, *200*(2), 295–311. https://doi.org/10.1016/j.jeconom.2017.06.012

Qiao, M., & Huang, K.-W. (2021). "Correcting Misclassification Bias in Regression Models with Variables Generated via Data Mining." *Information Systems Research*, *32*(2), 462–480. https://doi.org/10.1287/isre.2020.0977

Savoca, E. (2011). "Accounting for Misclassification Bias in Binary Outcome Measures of Illness: The Case of Post-Traumatic Stress Disorder in Male Veterans." *Sociological Methodology*, *41*(1), 49–76. https://doi.org/10.1111/j.1467-9531.2011.01239.x

Schochet, P. Z. (2013). "A Statistical Model for Misreported Binary Outcomes in Clustered RCTs of Education Interventions." *Journal of Educational and Behavioral Statistics*, *38*(5), 470–498. https://doi.org/10.3102/1076998613480393

Shiu, J.-L. (2016). "Identification and estimation of endogenous selection models in the presence of misclassification errors." *Economic Modelling*, *52*, 507–518. https://doi.org/10.1016/j.econmod.2015.09.031

Tang, L., Lyles, R. H., King, C. C., Celentano, D. D., & Lo, Y. (2015). “Binary regression with differentially misclassified response and exposure variables.” *Statistics in Medicine*, *34*(9), 1605–1620. https://doi.org/10.1002/sim.6440

Tennekoon, V., & Rosenman, R. (2016). “Systematically misclassified binary dependent variables.” *Communications in Statistics - Theory and Methods*, *45*(9), 2538–2555. https://doi.org/10.1080/03610926.2014.887105

Tennekoon, V. S. B. W., & Caudill, S. B. (2022). “Estimation of a selectivity model with misclassified selection.” *Communications in Statistics: Case Studies, Data Analysis and Applications*, *8*(1), 1–14. https://doi.org/10.1080/23737484.2020.1827465

van Hasselt, M., & Bollinger, C. R. (2012). “Binary misclassification and identification in regression models.” *Economics Letters*, *115*(1), 81–84. https://doi.org/10.1016/j.econlet.2011.11.031

This Technical Brief was written by Dimitri Stoelinga and José Rúbio-Valverde, with contributions by John DiGiacomo.

Suggested citation:

Stoelinga, D and Rúbio-Valverde, JR (2022). Measurement error in survey data: what is it and why does it matter. Technical Brief. Laterite.

laterite

DATA | RESEARCH | ADVISORY

From data to policy

Rwanda | Ethiopia | Kenya | Uganda | Tanzania | The Netherlands

www.laterite.com