Abstract Title Page
Not included in page count.

Title: A Statistical Model for Misreported Binary Outcomes in Clustered RCTs of Education Interventions

Authors and Affiliations: Peter Z. Schochet, Mathematica Policy Research, Inc.

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

In randomized control trials (RCTs) of educational interventions, there is a growing literature on impact estimation methods to adjust for missing student outcome data using such methods as multiple imputation, the construction of nonresponse weights, casewise deletion, and maximum likelihood methods (see, for example, Allison, 2002; Graham, 2009; Peugh & Enders, 2004; Puma, Olsen, Bell & Price, 2009; Schafer & Graham, 2002). Much less attention, however, has been devoted in education RCTs to developing statistical methods to adjust for the systematic *misreporting* of student outcome data for those with nonmissing data. Without appropriate adjustments, misreporting could lead to biased impact estimates, which could be exacerbated if the intervention leads to treatment-control differences in misreporting rates and the composition of students with misreported data. Misreporting could also affect the variance of the estimated impacts, and hence, significance levels from statistical hypothesis tests of intervention effects.

In some education RCTs, the extent of data misreporting can be assessed by conducting validation studies using "gold-standard" information from outside data sources and by conducting data reliability studies. In many education RCTs, however, data for such analyses may not be available that pertain to the specific outcomes and populations under investigation, and it may be prohibitively expensive to collect them.

Accordingly, this article develops a statistical model for education RCTs—that relies on study baseline data and distributional assumptions on model error terms—to obtain consistent estimates of average treatment effects (ATEs) and their standard errors in the presence of misreported outcome data. The focus is on two-level RCT designs where schools (or classrooms within schools) are randomly assigned to a treatment or control condition, although the methods developed in this article apply (collapse) to single-level designs where students are the unit of random assignment. We discuss both identification and consistent estimation of the ATE parameter with misreporting using the Neyman-Rubin model that underlies experiments.

The methods discussed in this article are germane to the SREE conference theme, because they can be used to examine the extent to which impact findings from standard HLM models might be sensitive to misreporting. Thus, analysts might consider adding the methods discussed in this article to the toolbox of exploratory analytic methods that can be used to assess the robustness of ATE findings to alternative model assumptions, specifications, and estimation methods.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

>

The focus of the article is on the systematic misreporting of a *binary* outcome, which is assumed to be coded so that a value of 1 pertains to an undesirable result (such as the student was not proficient in math or English, used illicit drugs, or dropped out of school) and a value of 0 pertains to a successful result. We consider the case where the binary outcome could be misreported as zero for those with a truly undesirable outcome but that it will always be reported

accurately for those with a truly successful outcome. Thus, the observed data will contain too many "zeroes," and estimates of the proportion of students with undesirable outcomes—labeled hereafter as "failure rates"—will be biased downwards for both the treatment and control groups, leading to impact estimates that could also be biased.

This article adapts the parametric "double hurdle" model proposed by Cragg (1971) for continuous outcomes and nonclustered settings to (1) the RCT context, (2) two-level clustered designs, and (3) binary outcomes. Cragg's double hurdle model for continuous outcomes has been used by many authors to model zero expenditures on food, alcohol, and tobacco from household surveys in various countries (see, for example, Deaton & Irish, 1984; Jones, 1989; Maki & Nishiyama, 1996; Su & Yen, 2000; Newman, Henchio, & Matthews, 2003; and Aristei & Pieroni, 2008), and was used by Blundell and Meghir (1987) to model the labor supply of married women. Hausman, Abrevava, and Scott-Morgan (1998) and Lewbel (2000) examine variants of the double hurdle model for binary outcomes using both parametric and semi-parametric estimation methods, but do not consider clustered designs or RCT settings.

In our context, the double hurdle model specifies that a value of 1 for the binary outcome will be observed only if two hurdles are overcome: (1) the student has a true binary value of 1 and (2) the student's outcome is recorded correctly in the data. Using a latent index approach, a random effects probit model is specified for each hurdle—separately for treatments and controls—and a quasi-Newton maximum likelihood (ML) approach is discussed for estimating the model parameters and their standard errors. In this framework, we do not observe which particular students have misreported outcomes, but we can estimate overall misreporting rates for the treatment and control groups. These estimated misreporting rates can then be used to obtain consistent ATE parameter estimates that are not contaminated by misreporting.

This article demonstrates the statistical approach using survey data from a large scale RCT of Job Corps, the nation's largest vocationally focused education and training program for disadvantaged youths between the ages of 16 and 24 (Schochet, Burghardt, & McConnell, 2008). The binary outcome for this analysis is the student-reported arrest rate during the four year follow-up period after random assignment. The Job Corps evaluation is a good case study for this article, because the literature suggests that adolescents tend to underreport their criminal activities in surveys. Furthermore, students in the study treatment sample may have had greater incentives to underreport their arrests than control students, because Job Corps students who violate Job Corp's zero tolerance policy are expelled from the program.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

> See above.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

> The double hurdle model used in the article is based on the following latent index variable framework where binary decisions are made depending on whether or not latent indices cross a threshold value of zero:

(9) $Y^*_{qij} = \mathbf{Q_{qij}}\boldsymbol{\beta}_\mathbf{q} + (\theta_{qi} + u_{qij})$

$Y_{qij} = 1$ if $Y^*_{qij} > 0$

$Y_{qij} = 0$ if $Y^*_{qij} \leq 0$

(10) $R^*_{qij} = \mathbf{X_{qij}}\boldsymbol{\gamma}_\mathbf{q} + \varepsilon_{qij}$

$R_{qij} = 1$ if $R^*_{qij} > 0$

$R_{qij} = 0$ if $R^*_{qij} \leq 0$.

In these equations, Y^*_{qij} is a continuous latent variable underlying the true potential binary outcome value for student j in school i and research condition q, and R^*_{qij} is a continuous latent index variable underlying the reporting accuracy of the student's data, which in our context is germane only for those with $Y^*_{qij} > 0$. The row-vectors $\mathbf{Q_{qij}}$ and $\mathbf{X_{qij}}$ are observed baseline covariates that contain student- and school-level variables (including the intercept) as well as random assignment blocking (stratification) variables such as school district indicators. It is assumed that conditional on the covariates, θ_{qi} are $iid\ N(0, \sigma^2_{\theta_q})$ school-specific random error terms that capture the correlation between latent index values for students in the same school. It is further assumed that conditional on the covariates and the school random effects, u_{qij} and ε_{qij} are $iid\ N(0,1)$ student random errors. The random errors within and across equations are assumed to be distributed independently of each other. The coefficient vectors $\boldsymbol{\beta}_\mathbf{q}$ and $\boldsymbol{\gamma}_\mathbf{q}$ and the variance $\sigma^2_{\theta_q} > 0$ are parameters to be estimated.

Equation 9 defines a random effects probit model for clustered RCT designs (see, for example, Gibbons, Hedeker, Charles & Frish 1994), where separate regression models are specified for the treatment and control groups. Equation 10 defines the misreporting process where the effects of covariates on reporting decisions and error variances are allowed to differ across the treatment and control groups. Equations 9 and 10 formalize a sequential decision-making process, where decisions are first made regarding binary outcome values, followed by decisions regarding reporting accuracy (for those with $Y^*_{qij} > 0$).

Using the parameter identifying assumptions laid out in the paper, the data generating process for the *observed* data is as follows:

(11) $y_{qij} = 1$ if $Y^*_{qij} > 0$ and $R^*_{qij} > 0$

$y_{qij} = 0$ if $Y^*_{qij} \leq 0$ or [$Y^*_{qij} > 0$ and $R^*_{qij} \leq 0$].

Thus, we observe $y_{qij} = 1$ (an undesirable outcome) if the true binary outcome value is 1 and the data are reported accurately. Conversely, we observe $y_{qij} = 0$ (a desirable outcome) if either the true binary outcome value is 0 or if the true binary outcome value is 1 and the data are misreported.

The log likelihood function for the vector of observed binary outcomes can be obtained in several steps using the approach of Butler and Moffit (1982) and ML methods can be used to

estimate the model parameters, separately for the treatment and control group samples under certain identifying assumptions discussed in the article.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

> The article applies the new methods to a large scale experimental evaluation of Job Corps, the largest federal training program for disadvantaged youths between the ages of 16 and 24. The empirical results suggest that accounting for survey misreporting for the arrest rate impact findings in the Job Corps study matters. Misreporting occurred for both treatments and controls, but was somewhat more common for the treatment students. Thus, accounting for misreporting increased the estimated arrest rates for both treatments and controls, and decreased the arrest rate impacts in absolute value.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

> This article develops a parametric statistical framework to test and adjust for the misreporting of binary outcomes in the estimation of ATEs for school-based RCTs. We consider a realistic scenario where it is assumed that binary outcomes on sensitive topics can be misreported for students with truly undesirable outcomes, but not for those with truly desirable outcomes. A latent index framework is employed where misreporting and binary outcome decision processes are modeled using available baseline data and normality assumptions about model error terms. This approach yields a “double hurdle” random effects probit model that can be estimated separately for treatments and controls. The article discusses quasi-Newton ML methods for obtaining consistent estimates of the unobserved misreporting rates, the ATEs on the considered binary outcomes, and standard errors of the estimates that are not contaminated by misreporting. The article also discussed how the approach can be applied to continuous outcomes and to nonclustered, student-level RCT designs.

Importantly, the success of the double hurdle model hinges critically on the predictive power of the baseline covariates used in the analysis. This is because the model uses the covariates to “adjust” the outcomes of students with reported successful outcomes who “look like” students with reported unsuccessful outcomes. Thus, the use of the double hurdle model will typically require the availability of detailed baseline data—including pre-intervention measures of the outcomes—that the literature suggests are correlated with the outcomes of interest for the study population. The article demonstrates, using simulations, the importance of predictive baseline data for the double hurdle model.

Finally, as shown in the case study using the Job Corps data, estimates using the double hurdle model might be sensitive to the choice of model covariates. Thus, researchers using this approach must carefully examine the robustness of study findings to alternative sets of covariates, and, in particular, to the inclusion of covariates that have significant predictive power in the models, but that could be endogenous or subject to measurement error, and thus, that might cause bias.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Allison, P.D. (2002) *Missing Data.* Thousand Oaks, CA: Sage.

Ansolabehere S. & E. Hersh (2011). Who really votes, In *Facing the Challenge of Democracy*, eds. P. Sniderman and B. Highton. Princeton: Princeton University Press.

David Aristei & Luca Pieroni, (2008). A double-hurdle approach to modelling tobacco consumption in Italy. *Applied Economics* 40(19), 2,463-2,476.

Barrera-Osorio, F., M. Bertrand, L. Linden & F. Perez-Calle (2011), Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia. *American Economic Journal: Applied Economics*, 3(2): 167–95.

Black D., S. Sanders & L. Taylor (*2003).* Measurement of higher education in the census and CPS, *Journal of the American Statistical Association*, 98(463):545–54.

Blundell R. & C. Meghir (1987*).* Bivariate alternatives to the Tobit model, *Journal of Econometrics*, 34(1), 179-200.

Bryk, A. and S. Raudenbush (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Butler, I. S. and R. Moffitt (1982): A computationally efficient quadrature procedure for the one-factor multinomial probit model", *Econometrica*, 50, 761-764.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with applications to the demand for durable goods, *Econometrica*, 39, 829-44.

Deaton, A. and M. Irish (1984), Statistical models for zero expenditures in household budgets, *Journal of Public Economics* 23, 59-80.

Dennis, J.E., Jr. & R. Schnabel (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, New York.

Fletcher, R. (1987). *Practical Methods of Optimization*. New York: Wiley.

Gibbons, R. D., D. Hedeker, S. Charles & P. Frisch (1994). A random-effects probit model for predicting medical malpractice claims. *Journal of the American Statistical Association*, 89, 760-767.

Gil, A., J. Segura & N. Temme (2007), Gauss quadrature, *Numerical Methods for Special Functions*, SIAM Publication.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60, 549*-576.

Greene, W. (2000). *Econometric Analysis.* 4th Edition. Upper Saddle River, NJ: Prentice Hall.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer & R. Tourangeau (2009). *Survey Methodology*. Hoboken, NJ: John Wiley and Sons.

Hadaway, C., P. Marler & M. Chaves (1993). What the polls don't show: A closer look at church attendance. *American Sociological Review*,58(6), 741-752.

Henchion, M. & A. Matthews (2003). A double hurdle model of the Irish household expenditure on prepared meals, *Applied Economics*, 35, 1053-1061.

Hausman, J.A., J. Abrevaya & F.M. Scott-Morton (1998), "Misclassification of the Dependent Variable in a Discrete-Response Setting," *Journal of Econometrics*, 87, 239-269.

Hindelang, M, T. Hirshi & J. Weis (1981), *Measuring delinquency.* Sage Beverly Hills. CA.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945-960.

Huizinga, D. & D. Elliott (1986), Reassessing the reliability and validity of self-report delinquency measures. *Journal of Quantitative Criminology* 2, 293-327.

Jones, A. (1989). A double hurdle model of cigarette consumption. *Journal of Applied Econometrics*, 4, 23-39.

Jones, A. (1992). A note on computation of the double-hurdle model with dependence with an application to tobacco expenditure. *Bulletin of Economic Research*, 44, 67-74.

Kane, T., C. Rouse & D. Staiger (1999), Estimating returns to schooling when schooling is misreported, *National Bureau of Economic Research Working Paper #7235*.

Katz, J. & G. Katz. (2010).\Correcting for survey misreports using auxiliary information with an application to estimating turnout. *American Journal of Political Science*, 54(3), 815-835.

Lewbel, A. (2000). Identification of the binary choice model with misclassification, *Econometric Theory*, 16(4), 603-609.

Maki, A. & S. Nishiyama (1996). An analysis of under-reporting for micro-data sets: the misreporting or double hurdle model. *Economic Letters*, 52, 211-220.

McDonald, M. (2003). On the over-report bias of the National Election Study. *Political Analysis* 11(2): 180-186.

Mishel, L. & J. Roy (2006), *Rethinking high school graduation rates and trends*, Washington D.C: Economic Policy Institute.

New York Times (June, 2011). Under pressure, teachers tamper with tests. *New York Times Education Section*: New York, New York.

Peugh, J. L., & Enders, C. K. (2004). Missing data in education research: A review of reporting practices and suggestions for improvement. *Review of Educational Research , 74* (4), 525-556.

Puma, M., R. Olsen, S. Bell & C. Price (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Education Psychology*, 66, 688-701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate, *Journal of Education Statistics*, 2(1), 1-26.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177.

Schochet, P. Z., J. Burghardt & S. McConnell (2008). Does Job Corps work? Impact findings from the National Job Corps Study. *American Economic Review* 98 (5): 1864–1886.

Sherry, B., M. Jefferds & L. Grummer-Strawn (2007). Accuracy of adolescent self-report of height and weight in assessing overweight status: a literature review. *Arch Pediatr Adolesc Med*, 161 (12), 1154-1161.

Su, S. J. & S. T. Yen (2008). A censored system of cigarette and alcohol consumption, *Applied Economics* 32, 729-737.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24-36.

Tourangeau, R. & T. Yan. 2007. Sensitive questions in surveys. *Psychological Bulletin*, 133(5): 859-883.

Appendix B. Tables and Figures

Not included in page count.