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

Peter Z. Schochet

October 2012

Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08543-2393
Phone: (609) 936-2783
Fax: (609) 799-0005
pschochet@mathematica-mpr.com

Forthcoming in the Journal of Educational and Behavioral Statistics

Abstract

In education RCTs, the misreporting of student outcome data could lead to biased estimates of average treatment effects (ATEs) and their standard errors. This article discusses a statistical model that adjusts for misreported binary outcomes for two-level, school-based RCTs, where it is assumed that misreporting could occur for students with truly undesirable outcomes, but not for those with truly desirable outcomes. A latent variable index approach using study baseline data is employed to model both the misreporting and binary outcome decision processes, separately for treatments and controls, using random effects probit models to adjust for school-level clustering. Quasi-Newton maximum likelihood methods are developed to obtain consistent estimates of the ATE parameter and the unobserved misreporting rates. The estimation approach is demonstrated using self-reported arrest data from a large-scale RCT of Job Corps, the nation's largest residential training program for disadvantaged youths between the ages of 16 and 24.

Keywords: Randomized Control Trials, Misreported Outcomes, Clustered Designs, Average Treatment Effects, Causal Impact Parameters

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. School-based designs are common in education research, because education RCTs often test interventions that provide

enhanced services to teachers or that affect the entire school. Thus, for these types of interventions, it is infeasible to randomly assign the treatment directly to students. The methods developed in this article, however, apply (collapse) to single-level designs where students are the unit of random assignment.

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.

The remainder of this article is in five sections. Section 1 briefly discusses the literature on the misreporting of outcomes that is germane to school-based RCTs. Section 2 discusses the identification of the causal ATE parameter with misreporting, and Section 3 discusses the double hurdle model and the ML estimation approach. Section 4 presents case study findings using the Job Corps data and Section 5 presents a summary and conclusions.

1. The Misreporting Problem in Education RCTs

Data sources for collecting student outcome data in education RCTs will typically depend on the tested interventions and study research questions. The data sources, however, will typically include some combination of:

- *Student assessments* that provide data on study-administered achievement test and behavior scale scores that are collected using common instruments across study sites
- *Student surveys* that provide self-reported data on student activities, behaviors, and attitudes towards school
- *Parent surveys* that provide information on parents' perspectives on their children's school performance and activities, school satisfaction measures, and measures of parental involvement with their children's learning
- *Teacher surveys* that provide teacher-level data on students' classroom performance, behaviors, and attitudes towards learning
- *Administrative school records* that provide data on district-mandated achievement test scores, class grades, absences, suspensions, and grade promotions

There is a large literature that documents the systematic misreporting of *survey* data for outcomes that are germane to education RCTs (see, for example, Groves, Fowler, Couper, Lepkowski, Singer & Tourangeau, 2009). Survey misreporting could occur because respondents (students, parents, and educators in our context) do not understand certain questions, have trouble mapping responses into a response category, or do not recall events. Perhaps more problematic for education RCTs, evidence from validation studies suggest that survey respondents tend to systematically misreport responses to sensitive questions that are deemed to be embarrassing or intrusive. For instance, youth tend to underreport survey responses to sensitive questions about illicit drug use, the consumption of alcohol, criminal activities, abortion, and sexual behavior to avoid an unfavorable impression or to avoid perceived legal consequences (Tourangeau & Yan, 2007). For similar reasons, overweight adolescents often underreport their height and weight (Sherry, Jefferds, & Grummer-Strawn, 2007). Conversely,

survey respondents tend to overreport socially desirable outcomes and activities, such as their educational attainment (Black, Sanders, & Taylor, 2003; Kane, Rouse & Staiger, 1999; Mishel & Roy, 2006), school attendance (Barrera-Osorio, Bertrand, Linden, & Perez, 2011), and participation in such civic activities as voting (Ansolabehere & Hersh, 2011; McDonald, 2003) and going to church (Hadaway, Marler & Chaves, 1993).

The evidence suggests also that respondents are more willing to report sensitive information when the questions are self-administered than if they are administered by an interviewer either in-person or by telephone (Tourangeau & Yan, 2007). These mode effects could affect the quality of data collected from student, teacher, and parent surveys.

Administrative school records data could also be systematically misreported for a number of reasons. Misreporting could be due to data coding errors or to problems linking student, teacher, and school data over time. Furthermore, there have been recorded instances where educators have tampered with children's standardized test scores to inflate them in response to pressures from the No Child Left Behind Act and the increasing use of educator accountability systems to measure teacher and school performance (New York Times, Education Section, June, 11 2010).

The prevalence of misreported data in education RCTs could also differ across the treatment and control groups, although whether these effects are larger for treatments or controls will depend on the context. For instance, students, parents, and teachers in the treatment schools may be invested in demonstrating that the intervention is effective, and thus, might have incentives to overreport beneficial student outcomes. On the other hand, respondents in the treatment schools may be more willing to cooperate with study data collection procedures than those in control schools, and thus, might provide more truthful responses. Misreporting rates could also differ across the treatment and control groups if reported student outcomes are directly related to the

receipt of intervention services. This might be the case, for example, for an after-school intervention where participation is contingent on good behavior and strong academic performance. Another example is the Job Corps evaluation that is used for the case study below.

2. The Causal ATE Parameter with Misreported Data

Consider an experimental design with n total schools where n_T schools are randomly assigned to a single treatment group and n_C schools are randomly assigned to a control group. It is assumed that the sample contains m_{qi} students in school i and research condition q ($q = T$ for treatments and $q = C$ for controls) and that there are $m = \sum_{q \in (T,C)} \sum_{i=1}^{n_q} m_{qi}$ total students.

To discuss the causal ATE parameter for school-based RCTs in the presence of misreported binary outcome data for students, we adapt the potential outcomes framework developed in Rubin (1974, 1977) and Holland (1986). Let T_i be an indicator that equals 1 if school i is randomly assigned to the treatment group and 0 if the school is randomly assigned to the control group, and let $\mathbf{T}$ be the $1 \times n$ vector of treatment assignments for all n study schools.

Let $Y_{ij}(\mathbf{T})$ be the *true* potential binary outcome for student j in school i at a follow-up data collection point, given the random vector of school treatment assignments $\mathbf{T}$. Importantly, we assume without loss of generality that the binary outcome is coded so that a value of 1 pertains to an undesirable result (such as, the student is not proficient in a math or English, used illicit drugs, was arrested, or dropped out of high school) and a value of 0 pertains to a positive result. Finally, let $R_{ij}(\mathbf{T})$ denote a potential binary indicator that equals 1 if the student's binary outcome is recorded correctly and 0 if it is misclassified, given $\mathbf{T}$.

For the remainder of this article, it is assumed that a student's potential outcomes, $Y_{ij}(\mathbf{T})$ and $R_{ij}(\mathbf{T})$, are unrelated to the treatment statuses of other schools (Rubin, 1980):

A1. ***Stable Unit Treatment Value Assumption (SUTVA):*** If $T_i = T_i'$, then $Y_{ij}(\mathbf{T}) = Y_{ij}(\mathbf{T}')$ and $R_{ij}(\mathbf{T}) = R_{ij}(\mathbf{T}')$.

SUTVA allows us to express $Y_{ij}(\mathbf{T})$ as $Y_{ij}(T_i)$ and $R_{ij}(\mathbf{T})$ as $R_{ij}(T_i)$. It implies that potential outcomes will not depend on the treatment or control assignments of other study schools. These conditions are likely to be plausible unless there is substantial interaction between students and staff across study schools.

It is assumed that $Y_{ij}(T_i)$ are random draws from Bernoulli distributions in the study superpopulation with probability parameters $p_T = P(Y_{ij}(1) = 1)$ for treatments and $p_C = P(Y_{ij}(0) = 1)$ for controls. We assume similarly that $R_{ij}(T_i)$ are Bernoulli random variables. Note that randomization ensures that T_i is independent of $Y_{ij}(1)$, $Y_{ij}(0)$, $R_{ij}(1)$, and $R_{ij}(0)$.

We invoke two reporting assumptions. The first assumption pertains to students whose true binary outcome values are 0 (successes):

A2. $P(R_{ij}(T_i) = 1 \mid Y_{ij}(T_i) = 0) = 1$**.**

This assumption states that data will always be reported accurately for students with truly successful outcomes. This condition implies, for example, that students who do not use illicit drugs will always report this non-usage in student surveys, and that educators do not have incentives to tamper with the test scores of students who are proficient in math or English. This simplifying assumption is consistent with the results of the data quality validation studies discussed above, and thus, is likely to be realistic in practice for many binary outcomes that are used in education research. Hausman et al. (1998) relax this assumption by considering

parameter identification and estimation for a more general misreporting model under nonclustered designs where both outcome values of 0 and 1 are allowed to be misclassified.

The second reporting assumption that we invoke pertains to students with undesirable outcomes:

A3. $0 < P[R_{ij}(T_i) = 1 \mid Y_{ij}(T_i) = 1] \leq 1$**.**

This assumption states that there will be at least one student with an undesirable outcome who will correctly report that outcome. As discussed further below, this assumption is required for parameter identification in the estimation models.

In order to simplify the notation, we hereafter write the potential outcomes as $Y_{Tij} = Y_{ij}(1)$, $Y_{Cij} = Y_{ij}(0)$, $R_{Tij} = R_{ij}(1)$, and $R_{Cij} = R_{ij}(0)$. In addition, we define y_{ij} to be the *observed* binary outcome for student j in school i. Because our focus is on misreporting, to keep the presentation manageable, we do not simultaneously model the missing data and misreporting processes, but assume that data on y_{ij} are missing at random conditional on the available model covariates.

Using this RCT framework, the causal ATE impact parameter for the population failure rate is defined as follows:

(1) $ATE_Y = E(Y_{Tij} - Y_{Cij}) = P(Y_{Tij} = 1) - P(Y_{Cij} = 1)$.

This parameter is the difference between the population failure rates in the treatment and control conditions.

If the binary outcome is recorded correctly for all students, we have the relation

(2) $y_{ij} = T_i Y_{Tij} + (1 - T_i) Y_{Cij}$.

In this case, we can link the observed data with the ATE_Y parameter as follows:

(3) $E(y_{ij} \mid T_i = 1) - E(y_{ij} \mid T_i = 0) = E(Y_{Tij} \mid T_i = 1) - E(Y_{Cij} \mid T_i = 0) = ATE_Y$,

where the second equality holds because $E(Y_{Tij} \mid T_i = 1) = E(Y_{Tij})$ and $E(Y_{Cij} \mid T_i = 0) = E(Y_{Cij})$ due to school-level randomization. Thus, in the absence of misreporting, ATE_Y can be consistently estimated as the treatment-control difference in observed failure rates using standard estimation approaches such as hierarchical linear modeling (HLM) (Raudenbush & Bryk, 1992), where the models could include student- and school-level baseline covariates to improve precision.

With misreporting, Equations 2 and 3 no longer hold. In this case, standard estimation approaches based on the observed data will not typically provide consistent estimates of ATE_Y, as shown in the following proposition:

Proposition 1. Under assumptions A1 and A2, standard impact estimators (such as HLM) will consistently estimate the following impact parameter:

(3) $ATE_Y^M = E(y_{ij} \mid T_i = 1) - E(y_{ij} \mid T_i = 0) = r_T P(Y_{Tij} = 1) - r_C P(Y_{Cij} = 1)$,

where $r_T = P(R_{Tij} = 1 \mid Y_{Tij} = 1)$ and $r_C = P(R_{Cij} = 1 \mid Y_{Cij} = 1)$ are correct reporting rates for those with unsuccessful outcomes in the treatment and control conditions, respectively.

Proof. Conditioning on Y_{Tij}, we can express $E(y_{ij} \mid T_i = 1)$ as follows:

(4*a*) $E(y_{ij} \mid T_i = 1) = E(y_{ij} \mid Y_{Tij} = 1, T_i = 1) P(Y_{Tij} = 1 \mid T_i = 1)$
$+ E(y_{ij} \mid Y_{Tij} = 0, T_i = 1) P(Y_{Tij} = 0 \mid T_i = 1).$

Conditioning further on R_{Tij} and noting from **A2** that $E(y_{ij} \mid Y_{Tij} = 0, T_i = 1) = 0$, we find that:

(4*b*) $E(y_{ij} \mid T_i = 1) = E(y_{ij} \mid Y_{Tij} = 1, R_{Tij} = 1) P(R_{Tij} = 1 \mid Y_{Tij} = 1) P(Y_{Tij} = 1)$,

where the redundant conditioning on T_i is dropped on the right-hand side of Equation 4b because of randomization. Because $E(y_{ij} \mid Y_{Tij} = 1, R_{Tij} = 1) = 1$ and $r_T = P(R_{Tij} = 1 \mid Y_{Tij} = 1)$, it follows that

$E(y_{ij} \mid T_i = 1) = r_T P(Y_{Tij} = 1)$. A parallel argument shows that $E(y_{ij} \mid T_i = 0) = r_C P(Y_{Cij} = 1)$, and Equation 3 follows.

Clearly, the ATE_Y^M and ATE_Y parameters will equate if $r_T = r_C = 1$ (that is, if there is no misreporting). The two parameters will also equate if $ATE_Y = 0$ and $r_T = r_C$ (that is, if the null hypothesis of no treatment effects is true and misreporting rates do not differ for treatments and controls). In addition, if $0 < r_T = r_C < 1$ and $ATE_Y \neq 0$ then $|ATE_Y^M| \leq |ATE_Y|$. For other scenarios, ATE_Y^M could be bigger, smaller, or equal to ATE_Y depending on the specific misreporting and failure rates in the treatment and control conditions. However, if $ATE_Y \neq 0$, for most realistic scenarios, ATE_Y^M will be smaller than ATE_Y in absolute value. Thus, in our context, if the intervention improves outcomes, misreporting will likely lead to downwardly biased impact estimates.

The following proposition follows directly from Proposition 1.

Proposition 2. Under assumptions **A1** to **A3**, the ATE_Y parameter can be recovered from the data as follows:

$$(5) \quad ATE_Y = \frac{E(y_{ij} \mid T_i = 1)}{r_T} - \frac{E(y_{ij} \mid T_i = 0)}{r_C} = \frac{\alpha_T}{r_T} - \frac{\alpha_C}{r_C},$$

where $\alpha_T = E(y_{ij} \mid T_i = 1)$ and $\alpha_C = E(y_{ij} \mid T_i = 0)$.

Equation 5 suggests that a consistent estimator for ATE_Y with misreporting is as follows:

$$(6) \quad \hat{ATE}_Y = \frac{\hat{\alpha}_T}{\hat{r}_T} - \frac{\hat{\alpha}_C}{\hat{r}_C},$$

where $\hat{\alpha}_q$ is a consistent estimate of α_q and $\hat{r}_q$ is a consistent estimate of r_q ($q \in (T, C)$). The estimator $\hat{\alpha}_q$ can be obtained using observed treatment and control group failure rates, and, as

discussed below, $\hat{r}_q$ can be obtained from the double hurdle model. If available, $\hat{r}_q$ can also be obtained using pertinent information from validation studies on misreporting rates or known associations between covariates and misreporting probabilities that can be used to predict misreporting rates for the study sample (see Katz & Katz, 2010).

Equation 6 is a ratio estimator because the numerators and denominators are both measured with error. Thus, a variance estimator for $\hat{ATE}_Y$ can be obtained using an asymptotic Taylor series expansion of $\hat{ATE}_Y$ around the true value ATE_Y:

$$(7) \quad \left(\hat{ATE}_Y - ATE_Y\right) \approx [\frac{(\hat{\alpha}_T - \alpha_T)}{r_T} - \frac{\alpha_T(\hat{r}_T - r_T)}{r_T^2}] - [\frac{(\hat{\alpha}_C - \alpha_C)}{r_C} - \frac{\alpha_C(\hat{r}_C - r_C)}{r_C^2}].$$

Taking squared expectations on both sides of Equation 7, ignoring covariance terms, and inserting estimators for unknown parameters yields the following first-order asymptotic variance estimator for $\hat{ATE}_Y$:

$$(8) \quad \hat{AsyVar}(\hat{ATE}_Y) \approx [\frac{\hat{AsyVar}(\hat{\alpha}_T)}{\hat{r}_T^2} + \frac{(\hat{\alpha}_T / \hat{r}_T)^2 \hat{AsyVar}(\hat{r}_T)}{\hat{r}_T^2}] + [\frac{\hat{AsyVar}(\hat{\alpha}_C)}{\hat{r}_C^2} + \frac{(\hat{\alpha}_C / \hat{r}_C)^2 \hat{AsyVar}(\hat{r}_C)}{\hat{r}_C^2}].$$

Because $\hat{\alpha}_q$ and $\hat{r}_q$ are asymptotically normal, $\hat{ATE}_Y$ will also be asymptotically normal (see, for example, Greene, 2000).[1] The case study below presents results using this ratio estimator as well as an alternative approach where consistent parameter and standard error estimates for ATE_Y are obtained *directly* from the double hurdle model (as discussed below).

[1] The normal approximation may not be suitable for event rates that are very close to 0 or 1.

Finally, it is important to note that the analysis presented above applies not only to binary outcomes but also to continuous outcomes where (1) more positive variable values are associated with poorer outcomes and (2) those with true positive values have incentives to report zero values. In this case, the analysis and notation from above applies except that we can no longer express expected values of potential outcomes as event probabilities. Note that with continuous outcomes, the approach assumes that truly positive outcome values are either reported correctly or as zero, but not any value in between. This assumption may not always be realistic in education research, and is a primary reason why this article focuses on binary outcomes.

3. The Double Hurdle Model and ML Parameter Estimation

The double hurdle model for continuous outcomes was introduced by Cragg (1971) to generalize the standard Tobit censored regression model (Tobin, 1958). This article adapts this model to (1) RCTs, (2) school-based designs, and (3) binary outcomes. The approach 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:

$$(9) \quad Y^*_{qij} = \mathbf{Q_{qij}\beta_q} + (\theta_{qi} + u_{qij})$$

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

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

$$(10) \quad R^*_{qij} = \mathbf{X_{qij}\gamma_q} + \varepsilon_{qij}$$

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

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

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 $Y^*_{qij} > 0$).[2]

[2] It is possible to allow for correlations of the errors across Equations 9 and 10 by specifying bivariate normal distributions (to allow for the possibility that R^*_{qij} and Y^*_{qij} are simultaneously determined). Jones (1992) used this

Note that it is theoretically possible to include school-level random effects in the reporting model in Equation 10. However, for sample sizes that are typically used in education RCTs, identification of the variance components for this specification will be problematic and statistical power will be low. Thus, we focus on a design that excludes random effects in Equation 10, but the approach presented below can be generalized to allow for this clustering.

Using assumptions **A1** to **A3**, 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. Note that to simplify the notation for expressing the log likelihood function below, we use the notation y_{qij} for the observed outcome rather than y_{ij} as above.

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). First, conditioning on the school random effects and the vector of observed covariates $\mathbf{Z_{qij}} = \left[T_i \;\; \mathbf{Q_{qij}} \; \mathbf{X_{qij}} \right]$, we have that

(12) $$P(y_{qij} = 1 \mid \mathbf{Z_{qij}}, \theta_{qi}) = P(Y^*_{qij} > 0 \mid \mathbf{Z_{qij}}, \theta_{qi}) P(R^*_{qij} > 0 \mid \mathbf{Z_{qij}}, \theta_{qi})$$
$$= \Phi(\mathbf{Q_{qij}}\boldsymbol{\beta_q} + \theta_{qi}) \Phi(\mathbf{X_{qij}}\boldsymbol{\gamma_q})$$

and

(continued)
approach for a nonclustered double hurdle design with continuous outcomes. Allowing for these correlations for clustered designs with binary outcomes, however, adds considerable computational complexity for parameter estimation, and is not performed in this article.

$$(13)\quad P(y_{qij}=0\mid \mathbf{Z_{qij}},\theta_{qi}) = P(Y^{*}_{qij}\le 0\mid \mathbf{Z_{qij}},\theta_{qi}) + P(Y^{*}_{qij}>0\mid \mathbf{Z_{qij}},\theta_{qi})P(R^{*}_{qij}\le 0\mid \mathbf{Z_{qij}},\theta_{qi})$$
$$= 1-\Phi(\mathbf{Q_{qij}\beta_q}+\theta_{qi})\Phi(\mathbf{X_{qij}\gamma_q}).$$

Second, because of the assumption that observed responses for students within school i are independent conditional on the random school effects, the joint probability of observing a vector pattern of binary responses $\mathbf{y_{qi}}$ for students in school i is equal to the product of their response probabilities:

$$(14)\quad l(\mathbf{y_{qi}}\mid \mathbf{Z_{qi}},\theta_{qi}) = \prod_{j=1}^{m_{qi}}[\Phi(\mathbf{Q_{qij}\beta_q}+\theta_{qi})\Phi(\mathbf{X_{qij}\gamma_q})]^{y_{qij}} \times$$
$$[1-\Phi(\mathbf{Q_{qij}\beta_q}+\theta_{qi})\Phi(\mathbf{X_{qij}\gamma_q})]^{1-y_{qij}},$$

where $\mathbf{Z_{qi}} = \left[\mathbf{Z_{qi1}}\ \mathbf{Z_{qi2}}\ \mathbf{...}\ \mathbf{Z_{qim_{qi}}}\right]$ contains covariates for all students in the school and $\Phi(.)$ is the standard normal cumulative distribution function.

Third, the marginal probability of observing $\mathbf{y_{qi}}$ that is *unconditional* on the random school effects can be obtained using Equation 14 by integrating out θ_{qi} over its assumed normal probability distribution:

$$(15)\quad h(\mathbf{y_{qi}}\mid \mathbf{Z_{qi}}) = \int_{-\infty}^{\infty} l(\mathbf{y_{qi}}\mid \mathbf{Z_{qi}},\theta_{qi})\frac{1}{\sigma_{\theta_q}}\phi(\frac{\theta_q}{\sigma_{\theta_q}})\partial\theta_q,$$

where $\phi(.)$ is the standard normal density function.[3] Finally, because student responses are independent across schools, the log likelihood function in research condition q for the full vector of student responses across all schools can be expressed as follows:

[3] If random effects are included in the reporting model in Equation 10, then these random effects would also need to be integrated out in Equation (15) in a symmetric way as for θ_{qi}.

$$(16) \quad \log L_q = \sum_{i=1}^{n_q} \log(h(\mathbf{y_{qi}} \mid \mathbf{Z_{qi}})).$$

If weights are used in the analysis to adjust for such factors as differential sampling probabilities and missing data, the log likelihood in Equation 16 can be generalized as follows:

$$(17) \quad \log L_q = \sum_{i=1}^{n_q} wgt_{qi} \log(h(\mathbf{y_{qi}} \mid \mathbf{Z_{qi}})),$$

where $wgt_{qi} = \sum_{j=1}^{m_{qi}} wgt_{qij}$ is the sum of the wgt_{qij} weights for students in school i (that are scaled to sum to n_q).

ML methods can be used to estimate the model parameters in Equation 17, separately for the treatment and control group samples. Parameter identification is driven by (1) the assumed nonlinear normal error distributions, (2) an *exclusion restriction* where at least one baseline covariate differs in $\mathbf{Q_{qij}}$ and $\mathbf{X_{qij}}$, and (3) assumption **A3**. The exclusion restriction is needed because the probabilities for each hurdle enter symmetrically in the likelihood function in Equation 14. Importantly, the predictive power of the $\mathbf{Q_{qij}}$ covariates plays a critical role in the model to help distinguish between $Y_{qij} = 0$ students (some of whom have misreported data) and $Y_{qij} = 1$ students (all of whom have accurate data). Intuitively, the $\mathbf{Q_{qij}}$ covariates are used to "adjust" the outcomes of students with reported values of 0 who "look like" students with reported outcome values of 1. Thus, the success of the model depends critically on the quality of the baseline covariates that are used to predict failure rates (see the simulation results presented in Appendix B).

In practice, there may not always be available scientific evidence to identify covariates that satisfy the exclusion restriction (that is, covariates that influence failure rates but not

misreporting rates, or vice versa). Furthermore, with sample sizes that are typically used in education RCTs, parameter identification will become problematic in practice if the $\mathbf{Q}_{\mathbf{qij}}$ and $\mathbf{X}_{\mathbf{qij}}$ covariates largely overlap. This parameter identification issue is a key reason why the previous literature using the double hurdle model has focused on models that *exclude* $\mathbf{X}_{\mathbf{qij}}$ covariates (see, for example, Deaton and Irish, 1984; and Hausman et al., 1998). Thus, in the empirical work below, we include a rich set of $\mathbf{Q}_{\mathbf{qij}}$ covariates in the model, but include either no $\mathbf{X}_{\mathbf{qij}}$ covariates beyond the intercept or a small number of $\mathbf{X}_{\mathbf{qij}}$ covariates (also included in $\mathbf{Q}_{\mathbf{qij}}$) that the literature indicates may be associated with misreporting rates in our context.

Simulation results shown in Appendix B suggest that the exclusion of relevant $\mathbf{X}_{\mathbf{qij}}$ covariates will not severely bias the estimated failure rates unless misreporting rates differ substantially across levels of the $\mathbf{X}_{\mathbf{qij}}$ covariates. Thus, if researchers have a priori evidence that some $\mathbf{X}_{\mathbf{qij}}$ covariates might matter, a small number of these covariates could be included in the model to examine the robustness of study findings. Alternatively, the model could be estimated separately within strata formed by the $\mathbf{X}_{\mathbf{qij}}$ covariates (where the misreporting equations would include an intercept only), and the estimates could then be aggregated across strata to obtain full sample estimates. It is theoretically possible, however, for studies to include a rich set of $\mathbf{X}_{\mathbf{qij}}$ covariates in the model if exclusion restrictions exist and sample sizes are large.

Commonly-used Gaussian-Hermite Quadrature (GHQ) procedures can be employed to approximate the integrals in Equation 15 (see Gil, Segura & Temme, 2007; Pennington, 1970; Stroud & Secrest, 1966). The general GHQ approximation is $\int_{-\infty}^{\infty} e^{-x^2} g(x) \partial x \approx \sum_{d=1}^{D} w_d g(a_d)$ for

some function $g(.)$, where the weights w_d and abscissas a_d can be obtained from published tables for D evaluation points.

To make the integrand in Equation 15 conform to the GHQ structure, we can make a change of variables in Equation 15 using the substitution $\psi_q = \theta_q / \sqrt{2}\sigma_{\theta_q}$. The GHQ method can then be applied, which after some algebra yields the following approximation to the log likelihood function:

$$(18) \quad \log L_q \approx \sum_{i=1}^{n_q} wgt_{qi}[\log(\frac{1}{\sqrt{\pi}}\sum_{d=1}^{D} w_d \prod_{j=1}^{m_{qi}} prob_{qij}^{y_{qij}} (1 - prob_{qij})^{(1-y_{qij})})], \text{ where}$$

$$(19) \quad prob_{qij} = \Phi(\mathbf{Q_{qij}\beta_q} + \sqrt{2}\sigma_{\theta_q} a_d)\Phi(\mathbf{X_{qij}\gamma_q}).$$

In the empirical work, quasi-Newton methods (see Fletcher 1987) were used to obtain ML estimates of $\boldsymbol{\beta}_\mathbf{q}$, $\boldsymbol{\gamma}_\mathbf{q}$, and σ_{θ_q}, and the inverse of the estimated Hessian matrix provided asymptotic variance estimates. Appendix A displays the gradient vectors of the log likelihood function that are required to apply this method as well as estimation details.

The ML estimates can then be used to obtain the following consistent estimates of $p_T = P(Y_{Tij} = 1)$ and $p_C = P(Y_{Cij} = 1)$:

$$(20) \quad \hat{p}_q = \sum_{f \in (T,C)} \sum_{i=1}^{n_f} \sum_{j=1}^{m_{fi}} wgt^*_{fij} \Phi(\frac{\mathbf{Q_{fij}}\hat{\boldsymbol{\beta}}_\mathbf{q}}{\sqrt{\hat{\sigma}^2_{\theta_q} + 1}}),$$

where $wgt^*_{fij} = wgt_{fij} / \sum_{f \in (T,C)} \sum_{i=1}^{n_f} \sum_{j=1}^{m_{fi}} wgt_{fij}$ are student-level weights. The estimator in Equation 20 is the weighted average of the predicted probabilities that a student in the sample has a value of $Y_{qij} = 1$, and is obtained using all treatment and control students. The corresponding ATE estimator—which we label the "direct ML estimator"—is $\hat{ATE}_Y = \hat{p}_T - \hat{p}_C$.

Similarly, the r_T and r_C reporting rate parameters can be estimated as follows:

$$(21) \quad \hat{r}_q = \sum_{i=1}^{n_q} \sum_{j:y_{qij}=1}^{m_{qi}} wgt_{qij}^{**} \Phi(\mathbf{X_{qij}} \hat{\boldsymbol{\gamma}}_\mathbf{q}),$$

where $wgt_{qij}^{**} = wgt_{qij} / \sum_{i=1}^{n_q} \sum_{j:y_{qij}=1}^{m_{qi}} wgt_{qij}$ and where the calculations are conducted using treatments (controls) with observed $y_{qij} = 1$ values. This estimator is the average predicted probability that a student with a reported undesirable outcome has his or her data classified correctly. The corresponding estimator for the misreporting rate is $(1 - \hat{r}_q)$. An alternative, more complex estimator for r_q that uses the full treatment and control samples is as follows:

$$(22) \quad \hat{r}_{q1} = \frac{1}{\hat{\alpha}_q} \sum_{f \in (T,C)} \sum_{i=1}^{n_f} \sum_{j=1}^{m_{fi}} wgt_{qij}^{*} \Phi(\mathbf{X_{fij}} \hat{\boldsymbol{\gamma}}_\mathbf{q}) \Phi\left(\frac{\mathbf{Q_{fij}} \hat{\boldsymbol{\beta}}_\mathbf{q}}{\sqrt{\hat{\sigma}_{\theta_q}^2 + 1}}\right).$$

To obtain asymptotic variance estimates for $\hat{p}_q$, $\hat{r}_q$, and $\hat{r}_{q1}$, we note first that these estimators are functions of subvectors of the full set of estimated model parameters $\hat{\boldsymbol{\lambda}}$. Let $\boldsymbol{\Omega}$ be the asymptotic variance-covariance (information) matrix for $\hat{\boldsymbol{\lambda}}$. The asymptotic variance of $\hat{p}_q$ can then be calculated using an asymptotic Taylor series expansion of $\hat{p}_q$ around the true value p_q (that is, using the Delta method) as follows:

$$(23) \quad \sqrt{n}(\hat{p}_q - p_q) \approx \sqrt{n} \mathbf{g_q} (\hat{\boldsymbol{\beta}}_\mathbf{q} - \boldsymbol{\beta}),$$

where $\mathbf{g_q} = (\partial p_q / \partial \boldsymbol{\beta}_\mathbf{q})$ is the gradient (row) vector. Equation 23 can then be used to obtain the following asymptotic variance for $\hat{p}_q$:

$$(24) \quad AsyVar(\hat{p}_q) = (1/n_q) \mathbf{g_q} \boldsymbol{\Omega}_{\boldsymbol{\beta}} \mathbf{g}_\mathbf{q}',$$

where $\boldsymbol{\Omega}_{\boldsymbol{\beta}}$ is the submatrix of $\boldsymbol{\Omega}$ that corresponds to $\boldsymbol{\beta}$. Because $\hat{\boldsymbol{\beta}}_{\mathbf{q}}$ is asymptotically normal, $\hat{p}_q$ will also be asymptotically normal (see, for example, Greene, 2000). Equation 24 can be evaluated using the estimators $\hat{\boldsymbol{\beta}}_{\mathbf{q}}$ and $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\beta}}$. The same method can be used to obtain asymptotic variance estimates for $\hat{r}_q$ and $\hat{r}_{q1}$.

Note that the statistical approach developed above simplifies considerably for *nonclustered* RCT designs where students are the unit of random assignment. In this case, the random school effect θ_{qi} does not enter the model (so that numerical integration is not required) and all student responses are independent conditional on the covariates. Thus, in nonclustered designs, the form of the likelihood function is based on a simplified version of Equation 14 where the random school effects are omitted and where the product of probabilities extends to the full treatment (control) sample.

The statistical approach can also be applied to continuous outcomes. The key changes are as follows: (1) the variance of u_q in Equation 9, $\sigma^2_{u_q}$, can now be specified and estimated; (2) the first condition in Equation 11 is that $y_{qij} = Y^*_{qij}$ is observed if $Y^*_{qij} > 0$ and $R^*_{qij} > 0$; and (3) the contribution to the likelihood function for those reporting nonzero outcomes becomes $\sigma^{-1}_{u_q}\phi([y_{qij} - \mathbf{Q_{qij}}\boldsymbol{\beta}_{\mathbf{q}} - \theta_{qi}]/\sigma_{u_q})\Phi(\mathbf{X_{qij}}\boldsymbol{\gamma}_{\mathbf{q}})$. The GHQ and quasi-Newton methods discussed above can then be used to obtain ML estimates.

4. Case Study Using National Job Corps Study Data

Job Corps is the nation's largest vocationally focused education and training program for disadvantaged youths. It serves youths between the ages of 16 and 24, primarily in a residential setting, in more than 100 Job Corps centers nationwide. It provides services to more than 60,000

new participants each year, at a cost of about $1.5 billion. While at centers, Job Corps students receive intensive vocational training, academic education, and a wide range of other services, including counseling, social skills training, and health education. Job Corps students enroll in centers for an average of about 8 months, but program duration is voluntary and varies: one quarter of students stay longer than one year and a similar fraction stay less than three months (Schochet et al., 2008).

The National Job Corps Study (NJCS) used an RCT design where from late 1994 to early 1996, nearly 81,000 eligible applicants nationwide were randomly assigned to a treatment group, who were allowed to enroll in Job Corps, or to a control group, whose 6,000 members were not (Schochet et al, 2008). Program impacts on key outcomes—education and training, employment and earnings, criminal activities, and drug use—were estimated using baseline and follow-up survey data collected during the four years after random assignment (and administrative earnings data covering the nine years after random assignment).

A key NJCS outcome was the arrest rate during the four-year follow-up period. Arrest data were obtained from follow-up surveys conducted 12, 30, and 48 months after random assignment. The surveys were conducted by telephone, and, if necessary, in-person. The NJCS estimated arrest rate impacts using a binary variable that was set to 1 for those who ever answered yes to the following question pertaining to the follow-up period: “Have you ever been arrested or charged with a delinquency or criminal complaint or for a probation or parole violation?” The sample for this analysis included those who completed the 48-month interview (81 percent of treatments and 78 percent of controls). The binary arrest variable was missing for less than 1 percent of survey respondents. All analyses were conducted using weights to adjust for the sample and survey designs and for interview nonresponse (Schochet et al., 2008).

The NJCS found that Job Corps participation significantly reduced self-reported arrest rates; 32.3 percent of controls reported that they were arrested during the 48-month period, compared to 29 percent of treatments, a statistically significant reduction of 3.3 percentage points (Schochet et al. 2008). Impacts were largest during the period when treatments were enrolled in Job Corps, but persisted during the postprogram period. Note that Job Corps students can leave centers during breaks and vacations; thus some treatments were arrested outside their centers.

As discussed, there is evidence from the literature that youth underreport their criminal activities. Huizinga and Elliott (1986), for instance, report that about 80 to 90 percent of youth with known offenses admit to them. Thus, the arrest rate impact findings for the NJCS may have been affected by underreporting. Furthermore, students in the 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, and treatment students may have been invested in showing that the program is effective.

Accordingly, we re-estimated the arrest rate impacts using the double hurdle model, where detailed baseline survey data were used to construct the model covariates. The clustering variable for the analysis was the Job Corps center to which a sample member was designated, which was obtained from Job Corps intake staff prior to random assignment, and thus, is available for both treatments and controls. The analysis sample contained 10,500 youths (6,350 treatments in the research sample and 4,150 controls) who (1) completed the baseline and 48-month follow-up interviews, (2) had nonmissing data for the binary arrest outcome, and (4) had nonmissing center designations for one of the 105 Job Corps centers in operation at the time of the evaluation. The NJCS weights were used for the analysis to obtain nationally representative estimates.

Table 1 displays descriptive statistics for the $\mathbf{Q}_{qij}$ covariates that were used in the probit models to predict arrest rates. The baseline covariates were measured at the time of random assignment and include measures of gender, race/ethnicity, age, education and employment experiences, family background, health, and prior arrests. These covariates were selected because the literature suggests that they are associated with juvenile arrest rates (see, for example, Steinberg, 2008) and exploratory analyses suggested that they had some explanatory power in the probit models.

As shown in Table 1, Job Corps serves disadvantaged youths; about three-quarters of students in the sample did not have a high school credential at random assignment, half grew up in female-headed households, and more than 40 percent were in households that received food stamps in the prior year. In addition, about one-quarter of students reported that they were ever arrested prior to random assignment. Because of random assignment, there were no statistically significant differences between the baseline characteristics of treatment and control students.

Note that Job Corps intake staff typically obtains official arrest records on program applicants as part of the program eligibility determination process. Thus, the self-reported prior arrest measure may not have been subject to as much measurement error as the follow-up arrest indicator that was used for the impact analysis. Nonetheless, we estimated models with and without the prior arrest measure as a covariate (see below), although our main results come from models that included the prior arrest measure.

Table 2 displays differences in *regression-adjusted arrest rates* for each baseline characteristic relative to the pertinent left-out characteristic for two binary choice models: (1) a standard random effects probit model and (2) the double hurdle model where the $\mathbf{X}_{qij}$ covariates included only an intercept. Consistent with the literature, we find that self-reported arrest rates

were significantly higher for males, the youngest students, those without a high school credential, those who were not working or in school prior to program application, those who received food stamps, those with health problems, and those with previous arrests. There were few differences between the findings for treatment and control students and between the double hurdle and standard random effects probit models, although the double hurdle model produced more statistically significant parameter estimates.

Table 3 displays estimated *misreporting rates* from the double hurdle model (among those who reported not being arrested during the follow-up period). The results are displayed for three model specifications that varied based on the included $\mathbf{Q}_{qij}$ and $\mathbf{X}_{qij}$ covariates: (1) Model 1, which included the highly predictive prior arrest indicator in $\mathbf{Q}_{qij}$ and only an intercept in $\mathbf{X}_{qij}$; (2) Model 2, which was similar to Model 1 except that the prior arrest indicator was not included in $\mathbf{Q}_{qij}$; and (3) Model 3, which included the prior arrest indicator variable in $\mathbf{Q}_{qij}$ as well as race/ethnicity indicators in $\mathbf{X}_{qij}$ to account for possible racial differences in the underreporting of criminal activities (see, for example, Hindelang, Hirshi & Weis, 1981). Misreporting rates for all models were estimated using Equation 21 and standard errors were estimated using Equation 24.

Across model specifications, the misreporting rate was about 2 to 4 percentage points higher for treatment than control students (Table 3). The misreporting rate for treatments was about 10 percent in Models 1 and 3 (that included the prior arrest variable), and 4.4 percent in Model 2; these misreporting rates are all statistically significant. The misreporting rate for controls was about 7 percent in Models 1 and 3, but 0 in Model 2. The estimated *impact* on the misreporting rate—that is, the treatment-control difference—was 3 percentage points in Model 1 (*p*-value =

.068), 4.4 percentage points in Model 2 (*p*-value = .035), and 2 percentage points in Model 3 (*p*-value = .197).[4]

Table 4 displays the *arrest rate ATE findings* for Models 1 to 3 using two double hurdle estimators: (1) the ratio estimator from Equations 6 and 8 and (2) the direct ML estimator from Equations 20 and 24. As discussed, the Job Corps evaluation found that 29 percent of treatments and 32.3 percent of controls reported ever being arrested during the follow-up period. Using the double hurdle model, these estimated arrest rates *increased* for both treatments and controls, reflecting the positive misreporting rate estimates that were found for most model specifications. Accordingly, the arrest rate *impacts* were typically *smaller* in absolute value than the original -3.3 percentage point impact. The estimated arrest rate impacts remain statistically significant using the ratio estimator, but not always for the direct ML estimator.

These results suggest that accounting for survey misreporting in the NJCS had some effect on the arrest rate impact findings. Misreporting occurred for both research groups, 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.

5. Summary and Conclusions

This article developed a parametric statistical framework to test and adjust for the misreporting of binary outcomes in the estimation of ATEs for school-based RCTs. We considered a realistic scenario where it was assumed that binary outcomes on sensitive topics can be misreported for students with truly undesirable outcomes, but not for those with truly

[4] Consistent with the literature, we found using Model 3 that blacks and Hispanics were more significantly likely to underreport their arrest rates than whites (not shown).

desirable outcomes. A latent index framework was employed where misreporting and binary outcome decision processes were 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 discussed 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.

In RCTs where suspicion exists that key outcomes might be misreported, education researchers could conduct data validation studies using gold-standard information from outside data sources (such as administrative records data) and data reliability studies. Pertinent data, however, may not always exist for such analyses and it may be prohibitively expensive to obtain them. In these instances, the double hurdle model could be used to conduct exploratory analyses to examine the extent to which the impact findings from standard HLM models might be sensitive to misreporting. Thus, analysts might consider adding the double hurdle model 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.

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. This article demonstrated, 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.

Appendix A. Details of the Quasi-Newton ML Estimation Procedure

The iterative quasi-Newton estimation method requires the gradient (first derivative) vector of the log likelihood in Equation 17, but not the Hessian matrix (which is very complex in our application). To derive the gradient vector for the analysis, we first reduce notational complexity by expressing the log likelihood function for research group $q \in (T, C)$ as follows:

$$(A1) \quad \log L_q \approx \sum_{i=1}^{n_q} wgt_{qi} \log(B_{qi}), \text{ where}$$

$$(A2) \quad B_{qi} = \frac{1}{\sqrt{\pi}} \sum_{d=1}^{D} w_d \prod_{j=1}^{m_{qi}} prob_{qij}^{\,y_{qij}} (1 - prob_{qij})^{(1-y_{qij})}, \text{ and}$$

$$(A3) \quad prob_{qij} = \Phi(\mathbf{Q_{qij}}\boldsymbol{\beta}_{\mathbf{q}} + \sqrt{2}\sigma_{\theta_q} a_d)\Phi(\mathbf{X_{qij}}\boldsymbol{\gamma}_{\mathbf{q}}) = \Phi_{qij}^{\theta}\Phi_{qij}^{X}.$$

After some algebra, we find that the gradient for the *k*th parameter β_{qk} in the row vector $\boldsymbol{\beta}_{\mathbf{q}}$ is as follows:

$$(A4) \quad \frac{\partial \log L_q}{\partial \beta_{qk}} = \sum_{i=1}^{n_q} \frac{wgt_{qi}}{B_{qi}} \frac{1}{\sqrt{\pi}} \sum_{d=1}^{D} w_d \sum_{j=1}^{m_{qi}} D_{qijk} H_{qij} \prod_{l \neq j}^{m_{qi}} prob_{qil}^{\,y_{qil}} (1 - prob_{qil})^{(1-y_{qil})}, \text{ where}$$

$$(A5) \quad D_{qijk} = \Phi_{qij}^{X} \phi_{qij}^{\theta} Q_{qijk} \text{ and}$$

$$(A6) \quad H_{qij} = \frac{[\Phi_{qij}^{\theta}\Phi_{qij}^{X}]^{y_{qij}}[1 - \Phi_{qij}^{\theta}\Phi_{qij}^{X}]^{1-y_{qij}}[y_{qij} - \Phi_{qij}^{\theta}\Phi_{qij}^{X}]}{\Phi_{qij}^{\theta}\Phi_{qij}^{X}[1 - \Phi_{qij}^{\theta}\Phi_{qij}^{X}]},$$

where Q_{qijk} is the *k*th covariate in the student's covariate vector $\mathbf{Q_{qij}}$, and ϕ_{qij}^{θ} is the normal density function associated with Φ_{qij}^{θ}.

Using parallel notation, the gradient $(\partial \log L_q / \partial \sigma_{\theta_q})$ is the same as above except that D_{qijk} in Equation A5 changes to $\sqrt{2}\Phi_{qij}^{X}\phi_{qij}^{\theta}a_d$. Finally, the gradient $(\partial \log L_q / \partial \gamma_{qk})$ is the same as above except that D_{qijk} changes to $\Phi_{qij}^{\theta}\phi_{qij}^{\eta}X_{qijk}$.

The quasi-Newton ML estimation approach was conducted using SAS Proc IML programs written by the author. The programs used the SAS NLPQN quasi-Newton function. The weights w_d and abscissas a_d for the GHQ method were obtained using the Mathematica Software for $D = 13$ evaluation points. The ML algorithm was applied for different starting parameter values to assess whether global versus local maxima were found.

The SAS NLPQN function provides ML estimates for the full set of parameters $\boldsymbol{\lambda}$ but not variance estimates. Thus, the variance estimates were estimated numerically using diagonal elements of the Hessian matrix $\mathbf{H} = -\hat{\boldsymbol{\Omega}}^{-1}$, where $\hat{\boldsymbol{\Omega}}$ was calculated using the gradients $g_{qk} = \partial \log L_q / \partial \gamma_{qk}$ as follows:

$$(A7) \quad \hat{\boldsymbol{\Omega}}(\mathbf{k},\mathbf{k}') = \frac{g_{qk}(\hat{\boldsymbol{\lambda}} + h\mathbf{J}_{\mathbf{k}}) - g_{qk}(\hat{\boldsymbol{\lambda}})}{2h} + \frac{g_{qk'}(\hat{\boldsymbol{\lambda}} + h\mathbf{J}_{\mathbf{k}}) - g_{qk'}(\hat{\boldsymbol{\lambda}})}{2h},$$

where $\mathbf{J}_{\mathbf{k}}$ is a column vector that equals one in row k and zero elsewhere, and h is a very small number (see Dennis & Schnabel, 1983).

Appendix B. Monte Carlo Simulations

We conducted simulations to examine the performance of the double hurdle model for correctly specified models and misspecified models where relevant X_{ij} covariates were excluded from the estimation models. Simulated datasets were obtained using the following model:

$$(B1) \quad Y_{ij}^{*} = \beta_0 + Q_{1ij}\beta_1 + Q_{2ij}\beta_2 + (\theta_i + u_{ij})$$

$$(B2) \quad R_{ij}^{*} = \gamma_0 + X_{1ij}\gamma_1 + \varepsilon_{ij},$$

where $Q_{1ij} = \lambda_i + \delta_{ij}$ is a $N(0,1)$ random variable where λ_i and δ_{ij} were drawn from $N(0,.1)$ and $N(0,.9)$ distributions, respectively; Q_{2ij} was drawn from a *Uniform*(-2,2) distribution; the

covariate X_{1ij} was set to Q_{1ij} in some specifications and omitted in others; the error terms θ_i and u_{ij} were drawn from $N(0,.1)$ and $N(0,.9)$ distributions, respectively; and ε_{ij} was drawn from a $N(0,1)$ distribution. The observed binary outcome Y_{ij} was then set to 1 if $Y_{ij}^* > 0$ and $R_{ij}^* > 0$ and to 0 otherwise. The model parameters were estimated using the ML procedures discussed in this article for various model specifications, assuming 30 schools and 75 students per school (typical treatment or control group sample sizes used in education RCTs). Note that Y_{ij}^* is not normally distributed, and thus, the simulations allowed for some specification error in the model distributional assumptions.

The β parameters were set to yield true failure rates that (1) averaged to 50 or 80/20 percent and (2) fell within a preset percentage point range for 95 percent of the sample—which we refer to as the "*95-percent range*"—and which determines the predictive power of the two covariates; this predictive power was assumed to be the same for each covariate. The β parameters were all positive so that increases in each covariate were associated with increased failure probabilities. A similar procedure was used to set the γ parameters. For example, to simulate (1) failure rates with a mean of 80 percent and a 95-percent range of ±20 percentage points and (2) a constant misreporting rate of 10 percent, we set $\beta_0 = 1.073$, $\beta_1 = \beta_2 = .518$, and $\gamma_0 = 1.282$. For each model specification, we obtained 1,000 simulated data sets, estimated the model parameters using ML methods for each one, and computed estimated failure and misreporting rates.

Table B.1 displays simulation results for *correctly specified* models where both the data generating and estimation models included the *same* covariates. In these simulations, the misreporting models either included no covariates (the first row in each sequence) or the

$X_{1ij} = Q_{1ij}$ covariate with various 95-percent ranges. The table entries display average simulated failure rates (SFRs) and average simulated misreporting rates (SMRs) across the simulations.

The results from Table B.1 suggest that the SFRs and SMRs are close to true values, although the SMRs stray further from the truth as the variance of misreporting rates increases across the sample. Importantly, these results hinge critically on the predictive power of the covariates in the failure rate model; SFR and SMR biases become noticeable if the 95-percentage point range for the true failure rates are smaller than those displayed in the table (not shown). Thus, the double hurdle model should only be used if the study has available baseline covariates that have considerable explanatory power in the failure rate models. One way to check this is to examine the distribution of predicted failure probabilities from the probit model.

Table B.2 displays simulation results for *misspecified* models where the data were generated using models that included the $X_{1ij} = Q_{1ij}$ covariate, but where the estimation models excluded this covariate. The results indicate that even with misspecification, the SFRs are close to true values unless X_{1ij} has significant predictive power in the misreporting model. The SMR results, however, are more sensitive to misspecification, and SMR biases become large when the misreporting rate varies substantially across the sample (that is, when the omitted X_{1ij} covariate matters).

TABLE B.1
Simulated Failure Rates (SFRs) and Simulated Misreporting Rates (SMRs) for Correctly Specified Models

True Misreporting Rate (Percents) and the ± 95-Percent Range (Percentage Points)	True Failure Rate (Percents) and the 95-Percent Range (Percentage Points)							
	50%				80 or 20%			
	±35		±50		±20		±40[b]	
	SFR	SMR	SFR	SMR	SFR	SMR	SFR	SMR
5%[a]	50.4	5.6	49.9	4.8	79.9	4.9	80.0	5.0
5% ±5	51.5	6.3	50.1	5.1	80.0	4.7	80.0	4.9
5% ±10	51.6	6.4	50.1	5.0	80.1	4.9	79.9	4.7
10%[a]	50.1	9.9	49.9	9.9	79.9	9.8	79.9	9.9
10% ±5	50.8	10.5	50.1	9.9	79.5	9.2	79.9	9.7
10% ±10	51.6	12.0	50.0	9.6	79.8	9.2	79.8	9.2
20%[a]	49.7	19.1	50.0	20.0	79.6	19.5	80.0	20.0
20% ±5	50.1	18.8	50.0	19.8	78.9	18.5	79.5	19.3
20% ±10	50.4	18.9	49.9	19.4	79.3	18.6	79.8	19.2
20% ±20	50.9	18.4	50.0	18.0	79.0	16.5	79.4	17.2

Notes: SFR (SMR) figures pertain to average estimated failure (misreporting) rates across the simulations for correctly specified models where the data generating and estimation models contained the same covariates. The simulations assumed 30 treatment or control schools and 75 students per school. See text for formulas and other assumptions.

[a]Figures in these rows pertain to models with constant misreporting rates across sample members.

[b]The upper value of the 95-percent range is 100 percent for this specification, whereas the lower value is 40 percent.

TABLE B.2
Simulated Failure Rates (SFRs) and Simulated Misreporting Rates (SMRs) for Misspecified Models

True Misreporting Rate (Percents) and the ± 95-Percent Range (Percentage Points)	True Failure Rate (Percents) and the 95-Percent Range (Percentage Points)							
	50%				80 or 20%			
	±35		±50		±20		±40[b]	
	SFR	SMR	SFR	SMR	SFR	SMR	SFR	SMR
5% ±5	49.6	4.1	49.1	2.9	77.8	2.1	78.0	2.3
5% ±10	49.8	4.6	49.0	2.5	77.8	2.0	77.7	1.8
10% ±5	48.7	7.3	49.1	8.1	77.0	6.3	77.6	7.0
10% ±10	48.2	5.6	47.9	4.7	74.7	2.8	74.9	3.0
20% ±5	47.9	15.9	49.0	18.2	76.8	16.5	77.4	17.2
20% ±10	45.4	10.9	47.9	15.9	72.9	11.8	74.3	13.4
20% ±20	44.0	5.3	45.1	7.3	67.3	2.9	67.9	3.4

Notes: SFR (SMR) figures pertain to average estimated failure (misreporting) rates across the simulations for misspecified models where the data generating model included $X_{1ij} = Q_{1ij}$ as a covariate, but where the estimation model excluded this covariate. The simulations assumed 30 schools and 75 students per school. See text for formulas and other assumptions.

[b]The upper value of the 95-percent range is 100 percent for this specification, whereas the lower value is 40 percent.

References

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.

TABLE 1

Descriptive Statistics for the Baseline Covariates Used to Predict Arrest Rates, by Treatment Status (Percentages)

Baseline Covariate Pertaining to the Random Assignment Date (Program Eligibility Date)	Treatment Group	Control Group	*p*-value of Significance Test of Difference
Male	58.9	59.4	0.212
Age			
16 to 17	40.8	41.4	0.417
18 to 19	32.0	32.1	0.986
20 to 24	27.1	26.5	0.418
Race/Ethnicity			
White, non-Hispanic	28.4	27.2	0.468
Black, non-Hispanic	47.1	47.3	0.468
Hispanic	16.9	17.4	0.906
Other	7.7	8.0	0.853
PMSA or MSA Residence Status			
In PMSA	30.2	30.8	0.864
In MSA	47.2	46.6	0.778
In neither	22.7	22.7	0.866
Has Natural Children	18.3	18.3	0.795
Lives with Spouse or Partner	6.4	6.6	0.697
Education			
High school diploma	18.6	18.5	0.745
GED	4.8	5.3	0.219
Neither	76.6	76.2	0.744
Not in School in the Prior Year	39.1	38.2	0.317
Employment			
Never had a full-time or part-time job	20.2	21.5	0.149
Employed in the past year	64.9	64.2	0.597
Not currently employed	78.0	78.9	0.294
Mother Was the Head of the Household When the Student Was 14 Years Old	47.5	48.4	0.351
Mother Has a High School Credential	54.0	54.2	0.701
Family Was on Welfare Most or All of the Time When Growing Up	19.3	19.3	0.941
Household Received Food Stamps in Past Year	42.6	43.0	0.710
Had Physical or Emotional Problems That Limited the Amount of Work That Could Be Done	4.7	5.3	0.168
Ever Arrested or Charged with a Delinquency or Criminal Complaint	26.3	26.4	0.689
Sample Size	**6,350**	**4,150**	**10,500; 105 Centers**

Source: NJCS Baseline Survey Data

Notes: All figures are weighted to adjust for the sample and survey designs and interview nonresponse, and *p*-values are adjusted for center-level clustering and design effects due to unequal weighting.

*Treatment-control difference in the percentages is statistically significant at the 5 percent level, two-tailed test.

TABLE 2

Regression-Adjusted Marginal Arrest Rates, by Treatment Status and Model (Percentage Points; Standard Errors in Parentheses)

	Random Effects Probit Model		Double Hurdle Model	
Baseline Covariate	Treatments	Controls	Treatments	Controls
Male	20.4 (1.2)*+	24.6 (1.5)*	22.0 (3.2)*	25.7 (4.3)*
Age (20 to 24 is left-out state)				
16 and 17	13.8 (1.9)*	13.9 (2.4)*	17.6 (3.5)*	15.4 (3.6)*
18 and 19	5.3 (1.6)*	7.0 (2.1)*	7.3 (1.9)*	7.9 (2.3)*
Race/Ethnicity (Hispanic is left-out state)				
White, non-Hispanic	3.1 (1.9)	5.2 (2.4)*	4.9 (2.0)*	7.0 (2.6)*
Black, non-Hispanic	2.8 (1.8)	4.1 (2.2)	4.5 (1.6)*	5.2 (2.0)*
Other	0.8 (2.6)	5.5 (3.5)	1.7 (2.3)	2.8 (2.9)
In PMSA or MSA (In PMSA is left out)				
In MSA	2.0 (1.4)	0.7 (1.9)	0.8 (1.2)	-0.2 (1.6)
In neither PMSA or MSA	2.5 (1.7)	3.2 (2.3)	3.7 (1.6)*	2.7 (2.0)
Has Natural Children	-1.9 (1.7)	3.5 (2.4)	-1.3 (1.5)	3.0 (2.0)
Does Not Live with Spouse or Partner	0.3 (2.5)	-0.2 (3.3)	3.3 (2.2)	-0.4 (2.8)
Education (High school diploma is left out)				
GED	4.7 (3.1)	8.4 (4.0)*	2.3 (2.6)	9.8 (3.4)*
No high school credential	7.5 (1.7)*	7.6 (2.2)*	6.7 (1.7)*	9.2 (2.2)*
Not in School in the Prior Year	1.8 (1.3)	0.5 (1.7)	4.9 (1.5)*+	0.7 (1.4)
Employment				
Never had a full-time or part-time job	-0.1 (2.0)	2.5 (2.6)	2.3 (1.9)	1.7 (2.2)
Employed in the past year	1.4 (1.7)	1.1 (2.3)	2.9 (1.6)	2.4 (1.9)
Not currently employed	2.7 (1.5)	3.1 (2.0)	3.7 (1.5)*	4.7 (1.9)*
Mother Was the Head of the Household When the Student Was 14 Years Old	0.3 (1.2)	0.2 (1.6)	0.7 (1.1)	-0.2 (1.3)
Mother Has a High School Credential	1.0 (1.2)	3.0 (1.5)	2.0 (1.1)	4.2 (1.7)*
Family Was on Welfare Most or All of the Time When Growing Up	0.8 (1.6)	0.9 (2.0)	1.1 (1.5)	2.7 (1.9)
Received Food Stamps in Past Year	3.8 (1.3)*	4.3 (1.7)*	5.5 (1.5)*	5.3 (1.8)*
Had Physical or Emotional Problems That Limited the Work That Could Be Done	4.7 (2.9)	4.1 (3.5)	6.2 (3.2)*	8.0 (3.8)*
Ever Arrested or Charged with a Delinquency or Criminal Complaint	17.2 (1.5)*	15.5 (1.8)*	20.5 (3.6)*	17.4 (3.9)*
Variance of Random Effect (Theta)	.008 (.06)	.017 (.01)	0.001 (0.08)	.014 (.04)
-2*Log Likelihood Value	**10,560**	**7,130**	**10,540**	**7104**
Sample Size	**6,350**	**4,150**	**6,350**	**4,150**

Source and Notes: See Table 1

*Difference between the regression-adjusted arrest rate for the subgroup relative to the rate for left-out subgroup is statistically significant at the 5 percent level, two-tailed test.

+Difference between the regression-adjusted arrest rates for the treatment and control groups is statistically significant at the 5 percent level, two-tailed test.

TABLE 3

Estimated Misreporting Rates from the Double Hurdle Model for the Self-Reported Non-Arrestees, by Treatment Status and Model Specification (Percentages; Standard Errors in Parentheses)

Model Specification	Misreporting Rates		
	Treatment Group	Control Group	Estimated Impact
1: All $\mathbf{Q}_{qij}$ covariates shown in Table 1; No $\mathbf{X}_{qij}$ covariates beyond the intercept	9.4 (1.1)*	6.4 (1.3)*	3.0 (1.7)
2: Same as Model 1 except $\mathbf{Q}_{qij}$ excludes the prior arrest indicator	4.4 (2.1)*	0.0 (0.0)	4.4 (2.1)*
3: Same as Model 1 except $\mathbf{X}_{qij}$ includes race/ethnicity indicators	10.3 (1.0)*	8.3 (1.2)*	2.0 (1.3)
Sample Size	**6,350**	**4,150**	**10,540; 105 Centers**

Source: NJCS baseline and follow-up interview data.

Notes: See Table 1 and text for estimation details.

*The estimated misreporting rate or impact is statistically significant at the 5 percent level, two-tailed test.

TABLE 4
Estimated ATEs on the Arrest Rate, by Model Specification (Percentages; Standard Errors in Parentheses)

Model Specification and Estimator	Treatment Group Mean Arrest Rate	Control Group Mean Arrest Rate	Estimated Impact
Standard Random Effects Model	29.0	32.3	-3.3 (0.84)*
Double Hurdle Model			
1: All $\mathbf{Q}_{qij}$ covariates shown in Table 1; No $\mathbf{X}_{qij}$ covariates beyond the intercept			
Ratio Estimator[a]	31.9	34.5	-2.6 (0.60)*
Direct ML Estimator[a]	32.7	33.3	-0.6 (0.60)
2: Same as Model 1 except $\mathbf{Q}_{qij}$ excludes the prior arrest indicator			
Ratio Estimator[a]	30.3	32.3	-2.0 (0.67)*
Direct ML Estimator[a]	30.7	32.3	-1.6 (0.67)
3: Same as Model 1 except $\mathbf{X}_{qij}$ includes race/ethnicity indicators			
Ratio Estimator[a]	32.3	35.3	-3.0 (0.59)*
Direct ML Estimator[a]	33.0	34.2	-1.2 (0.58)*
Sample Size	**6,350**	**4,150**	**10,540; 105 Centers**

Source: NJCS baseline and follow-up interview data.

Notes: See Table 1 and text for estimation details.

[a] The ratio estimator uses Equations 6 and 8 and the direct estimator uses Equations 20 and 24.

*The estimated impact is statistically significant at the 5 percent level, two-tailed test.