Quasi-Experimental Evaluation of Alternative Sample Selection Corrections*

Robert Garlick† and Joshua Hyman‡

November 25, 2018

Abstract

We use a natural experiment to evaluate sample selection correction methods' performance. In 2007, Michigan began requiring that all students take a college entrance exam, increasing the exam-taking rate from 64 to 99%. We apply different selection correction methods, using different sets of predictors, to the pre-policy exam score data. We then compare the corrected data to the complete post-policy exam score data as a benchmark. We find that performance is sensitive to the choice of predictors, but not the choice of selection correction method. Using stronger predictors such as lagged test scores yields more accurate results, but simple parametric methods and less restrictive semiparametric methods yield similar results for any set of predictors. We conclude that gains in this setting from less restrictive econometric methods are small relative to gains from richer data. This suggests that empirical researchers using selection correction methods should focus more on the predictive power of covariates than robustness across modeling choices.

Keywords: education, sample selection, selection correction models, test scores

*We thank Susan Dynarski, John Bound, Brian Jacob, and Jeffrey Smith for their advice and support. We are grateful for helpful conversations with Peter Arcidiacono, Eric Brunner, Sebastian Calonico, John DiNardo, Michael Gideon, Shakeeb Khan, Matt Masten, Arnaud Maurel, Stephen Ross, Kevin Stange, Caroline Theoharides, Elias Walsh, and seminar participants at AEFP, Econometric Society North American Summer Meetings, Michigan, NBER Economics of Education, SOLE, and Urban Institute. Thanks to ACT Inc. and the College Board for the data used in this paper. In particular, we thank Ty Cruce, John Carrol, and Julie Noble at ACT Inc. and Sherby Jean-Leger at the College Board. Thanks to the Institute of Education Sciences, U.S. Department of Education for providing support through Grant R305E100008 to the University of Michigan. Thanks to our partners at the Michigan Department of Education (MDE) and Michigan's Center for Educational Performance and Information (CEPI). This research used data structured and maintained by MCER. MCER data are modified for analysis purposes using rules governed by MCER and are not identical to those data collected and maintained by MDE and CEPI. Results, information and opinions are the authors' and are not endorsed by or reflect the views or positions of MDE or CEPI.

†Department of Economics, Duke University

‡Corresponding author. Department of Public Policy, University of Connecticut. Address: 10 Prospect St., 4th Floor, Hartford, CT 06103; Email: joshua.hyman@uconn.edu; Telephone: (959) 200-3751; Fax: (860) 246-0334

1 Introduction

Researchers routinely use datasets where outcomes of interest are unobserved for some cases. When latent outcomes are systematically different for observed and unobserved cases, this creates a sample selection problem. Many canonical economic analyses face this challenge: wages are unobserved for the non-employed, test scores are unobserved for non-takers, and all outcomes are unobserved for attriters from panel studies or experiments. Statisticians and econometricians have proposed many selection correction methods to address this challenge. However, it is difficult to evaluate these methods' performance without observing the complete outcome distribution as a benchmark.

We use a natural experiment to evaluate the performance of different selection correction methods. In 2007, the state of Michigan began requiring that all students in public high schools take the ACT college entrance exam, raising the exam-taking rate from 64% to 99%. We apply different selection correction methods, using different sets of predictors, to the pre-policy exam score data. We then compare the corrected data to the complete post-policy exam score data as a benchmark.

We compare the performance of eight selection correction methods: linear regression (i.e. no correction), a one-stage parametric censored regression model (Tobin, 1958), a two-stage parametric selection model (Heckman, 1974), and several two-stage semiparametric selection models (Ahn and Powell, 1993; Newey, 2009; Powell, 1987). These make successively weaker assumptions about the economic or statistical model generating the latent outcomes and probability that the outcomes are missing. We evaluate each method using sets of predictors that range from sparse (student demographics) to rich (lagged student test scores and school characteristics) to mimic the different types of data available to researchers. We examine whether the performance of these correction methods varies by student race or poverty status and whether they can match gaps in the benchmark data in achievement by race and income.

We find that performance is not sensitive to the choice of selection correction method but is sensitive to the choice of predictors. Performance is similar for methods with weak assumptions (e.g. two-stage semiparametric methods) and methods with very restrictive assumptions (e.g. linear regression). All methods perform poorly when we use sparse predictors and well when we use rich predictors. We see the same patterns for subpopulations based on race and poverty, showing that our findings are not specific to one data generating process.

We consider several explanations for the similar performance across correction methods. This is not explained by an absence of selection, the assumptions of the parametric models holding, a weak instrument, or the data being too coarse to use semiparametric estimation. We conclude that the violations of the parametric models' assumptions are not quantitatively important in this setting. In contrast, the importance of detailed school- and student-level predictors is easy to explain. These characteristics strongly predict both latent test scores and test-taking and hence improve performance irrespective of the choice of selection method. This echoes ideas in Imbens (2003) and Oster (2017) that there is more scope for bias from unobserved predictors when observed predictors explain less outcome variation.

We believe this is the first paper to evaluate the performance of selection correction methods for missing data against a quasi-experimental benchmark. Other missing data papers comparing estimates across selection correction methods lack a quasi-experimental or experimental benchmark for evaluation (Mroz, 1987; Newey, Powell, and Walker, 1990; Melenberg and Van Soest, 1996). Our approach is similar to the literature comparing different treatment effects methods against experimental benchmarks (LaLonde, 1986; Heckman, Ichimura, Smith, and Todd, 1998; Dehejia and Wahba, 1999).[1]

Our findings are relevant to three audiences. First, our findings can inform methodological choices by applied researchers using selection correction methods or adapting existing methods for new applications (e.g. Dahl 2002; Bonhomme, Jolivet, and Leuven 2016). Many applied researchers establish that their results are robust across different selection correction methods (Krueger and Whitmore, 2001; Card and Payne, 2002; Angrist, Bettinger, and Kremer, 2006; Clark, Rothstein, and Whitmore Schanzenbach, 2009). Our findings show that results can be robust without being correct. This suggests researchers should focus more on the strength of the relationships between the observed predictors, the missing data indicator, and the non-missing outcomes than robustness across different methods.

Second, our findings are relevant to econometricians developing selection correction methods or adapting methods for new problems such as dynamic selection (e.g. Semykina and Wooldridge 2013). Most econometric work comparing selection correction methods' performance uses either simulated data or real data without a quasi-experimental benchmark (Mroz,

[1]LaLonde (1986) and Heckman, Ichimura, Smith, and Todd (1998) evaluate selection correction methods for treatment effects against experimental benchmarks. However, selection into treatment is a substantively different economic problem from selection due to missing data. Correction methods may work well for missing outcome data, the problem we consider, but poorly for treatment effects problems or vice versa.

1987; Goldberger, 1983; Paarsch, 1984; Newey, Powell, and Walker, 1990; Vella, 1998). We advance the comparisons based on real data by providing a quasi-experimental benchmark that allows us to evaluate rather than compare performance. We complement the comparisons based on simulations by examining a real-world application, as performance in simulations can be sensitive to how closely the simulation parameters match real-world data (Busso, DiNardo, and McCrary, 2014; Frölich, Huber, and Wiesenfarth, 2015).

Third, our findings are relevant to researchers, practitioners, and policymakers who want to use test scores to infer population achievement when test-takers are selected. Our results show that US college entrance exam scores predict population achievement if other test scores are observed. This contributes to the literature on selection into college entrance exam-taking (Dynarski, 1987; Hanushek and Taylor, 1990; Dynarski and Gleason, 1993). Our findings may be relevant to other education settings with selection into test-taking. For example, enrollment, and hence test-taking, is heavily selected in many developing countries, and even assessments used for accountability in the U.S. miss some students. Our findings can help researchers, practitioners, and policymakers in these settings learn about cohort-level achievement from assessments of enrolled, test-taking students.

We describe the sample selection problem in Section 2.1 and selection correction methods in Section 2.2. In Section 3, we describe our data, our setting, and the extent of selection into test-taking in the pre-policy period. We report the main findings in Section 4 and discuss reasons for the similar performance of different selection correction methods in Section 5. In Section 6, we conclude and discuss the extent to which our findings might generalize.

We extend our main analysis in five appendices. In Appendix A, we describe the dataset construction and report additional summary statistics. In Appendix B, we elaborate on the selection correction methods and how we implement them. In Appendix C, we show that our findings are robust to evaluating selection correction methods using different criteria. In the main paper we evaluate corrections based on means: we compare the mean selection-corrected pre-policy test score to the mean score in the complete post-policy data. In the appendix we also evaluate selection corrections based on regression parameters and the full test score distribution.[2] In Appendix D, we show that our findings are robust to changes in regression

[2]Specifically, we estimate a selection-corrected regression of test scores on covariates using pre-policy data and compare the coefficients to the same regression estimated using the complete post-policy data. We then predict the full selection-corrected distribution of pre-policy test scores and compare this to the complete post-policy test score distribution. We also compare the predicted share of selection-corrected pre-policy ACT scores

specifications and sample definitions. In Appendix E, we replicate our analysis using aggregate data (e.g. mean test-taking rates and test scores by school), as many researchers observe only aggregate data. We show that performance improves as we aggregate data at lower levels but does not vary across selection correction methods, reinforcing the importance of richer data for selection correction.[3]

2 Sample Selection, Corrections, Evaluation Criteria

2.1 The Sample Selection Problem

We introduce the sample selection problem with an application common in education research. We want to analyze student achievement, using ACT scores to proxy for achievement. We observe scores for a subset of students, and the latent achievement distribution may differ for ACT-takers and non-takers. This is similar to the canonical selection problem in labor economics: wages are observed only for employed workers, and the latent wage distribution may differ by employment status (Gronau, 1974; Heckman, 1974). We focus on the case where selection into test-taking is determined by unobserved characteristics that are not independent of latent scores. Selection on only observed characteristics or on only unobserved characteristics independent of latent scores can be addressed with simpler methods.

All the selection correction models we consider are special cases of this framework:

$$ACT_i^* = X_i\beta + \epsilon_i \tag{1a}$$

$$TAKE_i^* = g\left(X_i, Z_i\right) + u_i \tag{1b}$$

$$TAKE_i = \begin{cases} 1 & \text{if} \quad TAKE_i^* \geq 0 \\ 0 & \text{if} \quad TAKE_i^* < 0 \end{cases} \tag{1c}$$

$$ACT_i = \begin{cases} ACT_i^* & \text{if} \quad TAKE_i^* \geq 0 \\ . & \text{if} \quad TAKE_i^* < 0 \end{cases} \tag{1d}$$

where ACT_i^* is the latent ACT score of student i with observed score ACT_i. The objects of interest are the conditional means of ACT_i^* given X_i (i.e. the parameters from the population

meeting a college readiness threshold to the share in the complete post-policy data. These comparisons may be useful for many applied researchers. But the corrections we evaluate are designed recover conditional outcome means, not distributions. Hence these comparisons should thus be interpreted with caution.

[3]Similarly, Clark, Rothstein, and Whitmore Schanzenbach (2009) study selection into ACT-taking in Illinois. They also find that parametric corrections using group-level data can approximate group-level latent ACT scores when other group-level test scores are observed.

linear regression of ACT_i^* on X_i) and the unconditional mean of ACT_i^*. We draw a distinction between the sample selection problem due to missing values of ACT_i^*, and the more general identification problem due to correlation between X_i and ϵ_i. We abstract away from the latter problem by assuming that the object of interest is the conditional mean of ACT_i^* given X_i, rather than some causal effect of X_i on ACT_i^*. The ordinary least squares estimator of β consistently estimates this conditional mean in the absence of sample selection. We therefore refer to "predictors" of test scores rather than "determinants" or "causes." In the main paper we restrict attention to models where the functional form of $X_i\beta$ is known and where X_i and i are additively separable.[4]

Equation (1b) models the sample selection problem. Selection depends on a vector of observed characteristics (X_i, Z_i) and an unobserved scalar term u_i, which has an unknown distribution and may be correlated with ϵ_i. There may exist instrumental variables Z_i that are independent of ϵ_i, influence the probability of taking the ACT, and do not influence latent ACT scores (all conditional on X_i). We do not assume that the functional form of $g(.,.)$ is known. Equations (1c) and (1d) show the relationships between latent and observed ACT-taking and scores. Note that we observe the vector X_i for students who do not take the ACT.

Selection bias arises because the expectation of the observed ACT score conditional on X_i depends on the conditional expectation of the error term:

$$\mathbb{E}\left[ACT_i|X_i, TAKE_i = 1\right] = X_i\beta + \mathbb{E}\left[\epsilon_i|g(X_i, Z_i) + u_i > 0, X_i\right] \tag{2}$$

If u_i and ϵ_i are not independent, the compound error term is correlated with X_i, creating an omitted variable problem.[5]

2.2 Selection Correction Methods

We evaluate eight selection correction methods. All are discussed in more detail in Appendix B and summarized in Appendix Table 3. First, we estimate $ACT_i = X_i\beta + \epsilon_i$ using ordinary least squares and the sample of ACT-takers. This approach provides a consistent estimator of β if

[4]The additive separability assumption is common in the empirical and theoretical literature on sample selection. See Altonji, Ichimura, and Otsu (2012) and Arellano and Bonhomme (2017) for exceptions. In Appendix D we implement an informal test of additive separability and fail to reject this assumption. We also show in Appendix D that our results are robust to alternative parametric specifications of $X_i\beta$.

[5]If ϵ_i and u_i are independent, then we describe the data as missing conditionally at random (Rubin, 1976) or selected on observed characteristics (Heckman and Robb, 1985). This still poses a sample selection problem but is straightforward to address.

unobserved predictors of test-taking are independent of latent test scores, because the omitted variable in equation (2) is zero under this assumption.[6] Second, we estimate $ACT_i = X_i\beta + \epsilon_i$ using a Type 1 Tobit maximum likelihood estimator and the sample of ACT-takers (Tobin, 1958). If ϵ_i is normally distributed and equal to u_i, we can estimate equation (2) by maximum likelihood, allowing consistent estimation of β. This method assumes that ACT-taking and ACT scores are jointly determined by the same unobserved student characteristic. If students with high latent ACT scores do not take the ACT (or vice versa), this assumption fails.

Third, we jointly estimate the score and test-taking models using a parametric selection correction method and assuming that $g(X_i, Z_i) = X_i\delta + Z_i\gamma$ (Heckman, 1974). If (ϵ_i, u_i) are jointly normally distributed, the omitted variable in equation (2) can be estimated and included as a control variable, allowing consistent estimation of β. This does not impose the Tobit model's restrictive assumption that student selection into ACT-taking is based on latent scores. However, this approach relies on specific distributional assumptions and may perform poorly if there is no excludeable instrument Z_i that predicts ACT-taking but not latent ACT scores (Puhani, 2002).[7] As our fourth model, we therefore estimate a Heckman selection correction model excluding the driving distance from each student's home to the nearest ACT test center from the outcome model. This follows Card (1995), among others, and we justify the exclusion restriction in Section 3.2.

We also estimate four semiparametric models, which relax the assumptions that (ϵ_i, u_i) are jointly normally distributed and that the functional form of $g(.,.)$ is known. Each model combines one of two ACT-taking models, estimated for all students, and one of two selection-corrected ACT score models, estimated for only ACT-takers. The first ACT-taking model is a series logit: a logit regression of $TAKE_i$ on polynomial functions of X_i and Z_i, with the polynomial order chosen using cross-validation. The second ACT-taking model is a nonparametric matching estimator that calculates the mean ACT-taking rate among group of students with similar predictor values. We use the predicted probabilities of ACT-taking from these models

[6]This OLS approach relates to a broader literature in statistics on imputation. Imputation methods replace student i's missing ACT score with the ACT score for a randomly chosen student with similar values of the predictors to i or the mean ACT score for a group of students with similar values of the predictors (Rubin, 1987). Like OLS, these methods assume there are no unobserved predictors of ACT-taking that also predict latent ACT scores. These methods differ from OLS by using different functional forms of equation (1a). Rather than evaluating a range of imputation methods, we show in Appendix Table 11 that our results are robust to alternative functional forms of equation (1a).

[7]Joint normality of (ϵ_i, u_i) is a sufficient but not necessary condition for this selection correction model to provide a consistent estimator of β. There are other assumptions on the joint distribution that are sufficient.

to construct two selection corrections for the ACT score model.

The first selection-corrected ACT score model approximates the bias term in equation (2) with a polynomial in $TA\hat{K}E_i^*$, following Heckman and Robb (1985) and Newey (2009). The second removes the bias term using pseudo-fixed effects for groups of students with similar values of $TA\hat{K}E_i$ (Ahn and Powell, 1993; Powell, 1987). These approaches do not rely on specific distributional assumptions. But they do impose some restrictions on the joint distribution of (ϵ_i, u_i) and the function $g(.,.)$ and may have poor statistical performance in even moderately large samples. We discuss the assumptions and implementation of the semiparametric models in Appendix B.

We refer to these eight methods as OLS, Tobit, Heckman, Heckman with IV, semiparametric Newey, nonparametric Newey, semiparametric Powell, and nonparametric Powell. In the body of the paper we only vary the ACT-taking equation and selection correction term; in the appendices we also vary the functional form of the latent ACT score model. We summarize the differences between these methods by describing a hypothetical student's ACT-taking choice. Assume that her decision to take the ACT depends on her unobserved (to the econometrician) interest in attending college. The OLS correction is appropriate if this interest is uncorrelated with unobserved predictors of her latent ACT score. The Tobit Type I correction is appropriate if this interest predicts her ACT-taking decision only through her latent test score, so she will take the ACT if and only if she has a high latent score, conditional on her observed characteristics. The Heckman corrections are appropriate if this interest is correlated with unobserved predictors of her latent ACT score but the joint distribution of these unobserved characteristics satisfies specific parametric conditions. The Newey and Powell corrections are appropriate if this interest is correlated with unobserved predictors of her latent ACT score and the joint distribution of these unobserved characteristics satisfies weaker conditions.

All these methods aim to point identify β. Another set of methods aims to derive bounds on possible values of β. These methods assume that non-takers have either very high or very low latent ACT scores and use these two extreme assumptions to construct bounds on the distribution of ACT scores (Manski, 1990; Lee, 2009). These methods yield bounds that are too wide to be informative in our application.[8]

[8]Manki's least restrictive bounding method assumes that all non-takers score either the maximum or minimum ACT score. This approach estimates bounds of [13.40, 26.32] points for the mean ACT score, which only excludes the top and bottom deciles of the complete post-policy ACT score distribution. Lee's more restrictive approach derives bounds for the difference in means between groups with higher and lower test-taking rates,

2.3 Evaluating Alternative Selection Correction Methods

We evaluate each of the eight selection correction methods by how closely they predict the mean ACT scores in the post-policy period, which we call the reference mean. For each correction method, we regress the selected pre-policy ACT scores on predictors to estimate $\hat{\beta}$ and then predict $\overline{ACT}_i = \overline{\hat{\beta} X_i}$, using the predictors for the full population. We compare this to the mean of the reference distribution. We construct the reference distribution from the observed post-policy score distribution in two stages. First, we adjust for small differences in the distribution of observed student predictors of ACT scores between the two time periods (shown in Table 2) using inverse probability weights. Second, we account for the fact that 1.5% of students do not take the ACT in the post-policy period by replacing their missing scores with predicted values from estimating equation (1a) by OLS on the post-policy data. We show in Appendix Figures 10 and 11 and Appendix Table 11 that our findings are not affected by these adjustments.

In Appendix C, we report results from evaluating selection correction methods on three additional criteria. First, we compare the estimated parameter vector $\hat{\beta}$ to the parameter vector from regressing the post-policy ACT scores on the same student predictors. Second, we compare the selection-corrected pre-policy ACT score distribution to post-policy ACT score distribution. Third, we compare the selection-corrected pre-policy student share passing a minimum ACT score typically interpreted as "college-ready" to the same share in the post-policy period.[9] Our main findings are robust across all these criteria.

For all evaluation criteria, we interpret the difference between the selection-corrected pre-policy statistic and the post-policy statistic as a measure of the correction method's bias, conditional on the predictors. We report this bias and the variance of the selection-corrected pre-policy statistic, estimated using a nonparametric cluster bootstrap, clustering by school.[10]

rather than bounds for the population mean. For example, the ACT-taking rate differs by 7.7 percentage points between black and white students and Lee's method yields bounds of [3.66, 5.56] points for the black-white ACT score gap, or roughly 0.4 standard deviations.

[9]The first additional criterion is similar to our primary comparison of the predicted mean, but does not use β_0, the constant term in equation (1a). Identification of the constant term in semiparametric correction methods is a challenge that we discuss in section V. The selection correction methods we evaluate are not designed to perform well on the second and third additional criteria. However, these criteria are of interest to many applied researchers and we show how selection correction methods can be informally adapted for this purpose.

[10]To the best of our knowledge, the literature has not proposed an analytical variance estimator for two-stage semiparametric selection correction models with clustered data. We follow typical empirical practice by using the bootstrap, though this is problematic for our nonparametric first stage model (Abadie and Imbens, 2008).

3 Context, Data, and the Extent of Selection

We use student level data for two cohorts (2005 and 2008) of all first-time 11th graders attending Michigan public high schools.[11] Using the last pre-policy cohort (2006) and first post-policy cohort (2007) would minimize demographic differences between the samples. However, the policy was piloted in some schools in 2006, and not all districts implemented the reform in 2007. Given these challenges with the 2006 and 2007 cohorts, our main analysis uses the 2005 and 2008 cohorts. Our results are robust to using the 2006/2007, 2006/2008, and 2005/2007 cohort combinations (see Appendix Figures 12, 13, and 14).

3.1 Data

We use student-level administrative data from the Michigan Department of Education (MDE) that cover all first-time 11th grade students in Michigan public schools. The data contain the time-invariant demographics sex, race, and date of birth, as well as the time-varying characteristics free and reduced-price lunch status, limited-English-proficiency status (LEP), special education status (SPED), and student home addresses. The data also contain 8th and 11th grade state assessment results in multiple subjects. We match the MDE data to student-level ACT and SAT information over the sample period and to the driving distance between students' home during 11th grade and the nearest ACT test center.[12] See Appendix A for more information about our data and sample definition.

Table 1 shows sample means for the combined sample (column 1) and separately for the two cohorts of interest (columns 2 and 5). Four patterns are visible. First, the fraction of students taking the ACT jumped discontinuously from 2006 to 2007 when the policy was introduced. The ACT-taking rate rose from 64.1% in 2005 to 98.5% in 2008.[13] Second, mean ACT scores did not vary across years within each policy period: they are almost identical in 2005 and 2006 and in 2007 and 2008. This suggests that cohort-level latent achievement was stable through time, supporting our claim that differences in observed ACT scores reflect changes in ACT-taking rather than changes in composition.

[11] Throughout the paper, we refer to academic years using the spring year (e.g., we use 2008 for 2007-08).

[12] If a student took the ACT multiple times, we use their first score. If a pre-policy student took the SAT but not the ACT, we convert their score into ACT scale using the standard concordance table.

[13] Michigan's policy required 95% of students in each school to take the ACT for accountability purposes but did not require that individual students took the exam to graduate high school. This explains why 1.5% of students did not take the ACT exam even after the policy change.

Table 1. Sample Means of Michigan 11th Grade Cohorts

	2005 and 2008	2005 Cohort	2006 Cohort	2007 Cohort	2008 Cohort	08-05 Diff (5) - (2)	P-Value (6)=0
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Demographics							
Female	0.516	0.514	0.515	0.517	0.517	0.003	0.226
White	0.790	0.805	0.792	0.782	0.775	-0.030	0.000
Black	0.145	0.132	0.148	0.154	0.158	0.026	0.000
Hispanic	0.029	0.027	0.027	0.029	0.031	0.004	0.000
Other race	0.035	0.036	0.033	0.034	0.035	0.000	0.600
Free or reduced lunch	0.242	0.204	0.231	0.256	0.279	0.075	0.000
Local unemployment	7.518	7.285	7.064	7.329	7.745	0.460	0.000
Driving miles to nearest ACT test center	3.71	4.87	4.61	2.59	2.58	-2.29	0.000
Took SAT	0.058	0.076	0.069	0.047	0.039	-0.037	0.000
SAT Score	25.2	24.8	24.6	25.6	25.9	1.0	0.000
Took SAT & ACT	0.054	0.070	0.064	0.046	0.039	-0.031	0.000
Took ACT or SAT							
All	0.815	0.641	0.663	0.971	0.985	0.345	0.000
Male	0.793	0.598	0.621	0.969	0.984	0.387	0.000
Female	0.836	0.681	0.702	0.973	0.986	0.305	0.000
Black	0.780	0.575	0.608	0.905	0.947	0.372	0.000
White	0.822	0.652	0.674	0.985	0.993	0.341	0.000
Free or reduced lunch	0.749	0.434	0.483	0.936	0.970	0.536	0.000
Not free/reduced lunch	0.838	0.693	0.717	0.983	0.991	0.299	0.000
Low grade 8 scores	0.747	0.474	0.513	0.972	0.979	0.505	0.000
High grade 8 scores	0.875	0.778	0.789	0.971	0.991	0.213	0.000
First ACT or SAT Score							
All	19.9	20.9	20.8	19.2	19.3	-1.6	0.000
Male	19.9	21.0	20.9	19.1	19.2	-1.8	0.000
Female	19.9	20.7	20.6	19.2	19.3	-1.4	0.000
Black	16.0	16.8	16.6	15.8	15.6	-1.2	0.000
White	20.6	21.4	21.5	19.8	20.0	-1.5	0.000
Free or reduced lunch	17.1	18.3	18.0	16.7	16.8	-1.5	0.000
Not free/reduced lunch	20.7	21.3	21.3	20.0	20.2	-1.1	0.000
Low grade 8 scores	16.8	17.8	17.6	16.4	16.3	-1.4	0.000
High grade 8 scores	22.1	22.4	22.5	21.6	21.8	-0.6	0.000
Number of Students	197,014	97,108	99,441	101,344	99,906		

Notes: The sample is first-time 11th graders in Michigan public high schools during 2004-05 through 2007-08 who graduate high school, do not take the SPED 11th grade test, and have a non-missing home address. Free or reduced-price lunch lunch status is measured as of 11th grade. Low (high) grade 8 scores are below (above) the median score in each sample.

Third, ACT-taking rates increased more for student groups with lower pre-policy rates: black students, free lunch-eligible students, and students with low 8th grade test scores. These same groups saw weakly larger drops in their mean scores. This shows that groups of students pre-policy positively selected into ACT-taking based on their latent ACT scores, and that the policy largely eliminated this selection. Fourth, student demographics changed smoothly through time with no jump at the policy change. The percentage of black and free lunch-eligible students rose, as did the unemployment rate. Our comparisons account for this shift by reweighting the post-policy cohort to have the same distribution of observed characteristics as the pre-policy cohort (DiNardo, Fortin, and Lemieux, 1996).[14] This adjustment does not account for cross-cohort differences in unobserved latent ACT score predictors.

3.2 Modeling ACT-Taking

The two-stage selection correction methods are identified either by functional form assumptions, which are seldom viewed as credible in empirical work, or by an exclusion restriction, a variable that predicts ACT-taking but not latent test scores. We use the driving distance from each student's home to the nearest ACT test center to provide an exclusion restriction. We assume that students with easier access to a test center have a lower cost and hence higher probability of taking the test but do not have systematically different latent test scores, conditional on the other test score predictors.[15] We show below that driving distance strongly predicts test-taking and does not predict scores on non-ACT tests, supporting the exclusion restriction. Appendix Table 1 shows percentiles of the distance distribution by period and by urban/rural status. This exclusion restriction follows closely from prior research on education participation (Card, 1995; Kane and Rouse, 1995). We do not claim that the exclusion restriction is perfect, but rather that it is consistent with common empirical practice. This is the appropriate benchmark if we aim to inform empirical researchers' choice of selection correction methods, conditional on the type of instruments typically available.

We test if distance robustly predicts ACT-taking. Using pre-policy data, we estimate a probit regression of ACT-taking on a quadratic in distance. A quadratic allows the marginal cost of

[14]Our reweighting model includes indicators for individual race, gender, special education status, limited English proficiency, and all interactions; school means for the same four variables, urban/suburban/rural location and all interactions; and district enrollment, pupil-teacher ratio, local unemployment rate and all interactions. Results are robust to alternative reweighting models or not reweighting.

[15]Bulman (2015) finds SAT-taking rises when schools offer the SAT, supporting the first assumption.

ACT-taking to vary with distance, accounting for fixed costs of travel or increasing marginal cost of time. We report the results in Table 2. Without controlling for any other predictors, the distance variables are jointly but not individually significant ($\chi^2 = 12.54$, $p = 0.002$). The relationship grows stronger as we control for student demographics, school- and district-level characteristics, and student scores on other tests ($\chi^2 = 25.15$, $p < 0.001$). The controls account for low test-taking by disadvantaged students who live in dense urban areas where distances to test centers are small. The probability of ACT-taking falls with distance, dropping by 4 percentage points with a move from the 5^{th} to the 95^{th} percentile of driving distance to the nearest ACT test center (14.1 miles). The instrument passes standard tests for instrument strength, though these tests are developed for linear two-stage least squares models (Stock and Yogo, 2005). We return to the interpretation of the instrument in Section 5, including a discussion of identification at infinity.

We also use a placebo test to assess whether distance predicts latent achievement. We regress the average of students' 11th grade math and English test scores on the quadratic in distance, reporting results in columns 5-8 of Table 2. Distance to a test center is associated with higher scores but this relationship disappears when we control for other student characteristics (χ^2=1.30, p=0.480). This shows that distance predicts ACT-taking but not latent academic performance, providing reassurance about the exclusion restriction's validity.

3.3 Describing Selection by Comparing Pre- & Post-Policy Score Distributions

In this subsection, we compare the observed pre- and post-policy ACT score distributions to describe pre-policy selection into ACT-taking. Positive/Negative selection occurs if pre-policy scores are systematically higher/lower than post-policy scores. Researchers using selected test scores often assume that all non-takers would score below some percentile in the observed distribution (Angrist, Bettinger, and Kremer, 2006) or below all takers (Krueger and Whitmore, 2001). We assess the plausbility of these assumptions in our setting.

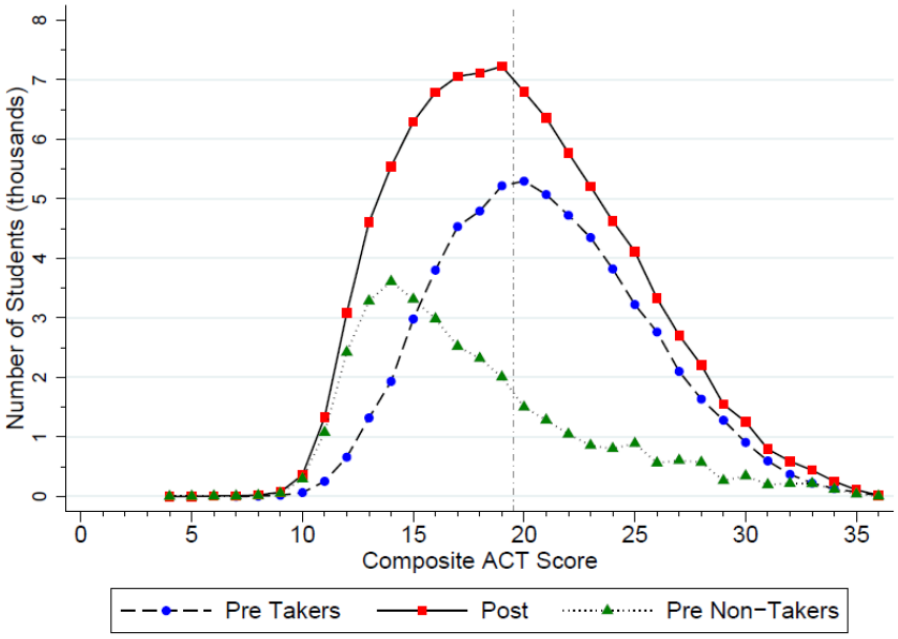
We estimate the latent ACT score distribution for non-takers by subtracting the number of test-takers with each ACT score in the pre-period from the number with each score in the post-period. We reweight the post-policy cohort to have the same number of students and distribution of observed characteristics. If the reweighting accounts for all latent test score predictors that differ between periods, then the difference in the number of students at each

Table 2. Testing the Exclusion Restriction: the Relationship Between Test Center Proximity, Test-Taking, and Achievement

	Dependent Variable = Took the ACT				Dependent Variable = 11th Grade Test Score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Distance (miles)	-0.003	-0.009***	-0.006***	-0.007***	0.030***	-0.003	0.001	0.002
	(0.002)	(0.002)	(0.001)	(0.001)	(0.007)	(0.005)	(0.002)	(0.002)
Distance Squared (/ 10)	-0.000	0.003***	0.003***	0.003***	-0.014***	-0.002	-0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.001)	(0.001)
Student-Level Demographics	N	Y	Y	Y	N	Y	Y	Y
School- & District-Level Covs	N	N	Y	Y	N	N	Y	Y
Student-Level Test Scores	N	N	N	Y	N	N	N	Y
R-Squared	0.001	0.045	0.088	0.223	0.003	0.110	0.203	0.647
Chi-2 Statistic	12.54	22.38	19.87	25.15	31.82	20.86	0.16	1.30
Sample Size	97,108	97,108	97,108	97,108	86,679	86,679	86,679	86,679

Notes: The sample is as in Table 1 but includes only the 2005 11th grade cohort. Columns (1)-(4) report marginal effects from probit and columns (5)-(8) are OLS. Distance is driving distance in miles from the student's home address during 11th grade to the nearest ACT test center. The distance squared term is divided by 10 for interpretability. The dependent variable in columns (1)-(4) is a dummy for taking the ACT (mean = 0.64), and in columns (5)-(8) is the average of 11th grade math and English test scores standardized to have mean zero and SD 1. The drop in sample size between columns (1)-(4) and (5)-(8) is due to missing 11th grade test scores. Student-level test scores included as covariates are average math and English 8th grade score and 11th grade social studies score. See text for the complete list of covariates. Standard errors clustered at the school-level.
*** indicates statistical significance at the 0.01 level, ** at the 0.05 level, and * at the 0.10 level.

Figure I. Frequency Distribution of Observed and Latent ACT Scores by Period



Notes: Figure shows the number of students attaining each ACT score in the pre-policy period (dashed line with blue circles) and the number of students attaining each ACT score in the post-policy period (solid line with red squares) after reweighting the post-policy data to have the same distribution of observed covariates as the pre-policy data (DiNardo et al., 1996). The difference between the two numbers (dotted line with green triangles) is a possible measure of how many pre-policy non-takers would attain each ACT score. We display frequencies rather than densities to demonstrate the change in the number of ACT takers from the pre- to post-policy period.

ACT score equals the number of non-takers with that latent score.[16]

Figure 1 plots the frequency distribution of ACT scores pre-policy, the reweighted post-policy distribution of scores, and the difference, which proxies for the latent scores of non-takers pre-policy.[17] The observed test score distribution is approximately normal, reflecting the test's design. The non-takers' test score distribution is shifted to the left. The mean pre-policy ACT score is 1.3 points or 0.27 standard deviations higher than the mean post-policy ACT score. Almost 60% of takers achieve the ACT's "college-readiness" score, while less than 30% of the non-takers would do so. However, some non-takers have high latent scores: 68% and 24% of the latent scores exceed the 10th and 50th percentiles of the observed score distribution.

There is clear positive selection into ACT-taking, but less than that assumed in prior stud-

[16]Hyman (2017) conducts a more extensive version of this analysis, measuring the number of students in the pre-policy cohort who have college-ready latent scores but do not take a college entrance test. He also examines the effect of the mandatory ACT policy on postsecondary outcomes.

[17]Appendix Table 2 reports moments and percentiles of the three distributions.

ies. Angrist, Bettinger, and Kremer (2006) and Krueger and Whitmore (2001) use Tobit and bounding analyses by assuming that all non-takers would score below specific quantiles of the observed distribution. In our data, this type of assumption would hold only at very high quantiles, generating uninformative bounds. We conclude that selection corrections relying on strong assumptions about negative selection are not justifiable in this setting.

The substantial number of individuals with high latent outcomes selecting out of participation is not unique to our setting. For example, Bertrand, Goldin, and Katz (2010) show that women who temporarily leave the labor market are not negatively selected on predictors of latent wages. Similarly, high-income respondents routinely decline to report incomes on surveys, generating positive selection. We do not believe that the pattern of selection shown in Figure 1 weakens the generalizability of our results.

4 Results

4.1 Comparing Sample Selection Corrections

In this section, we evaluate the performance of multiple selection correction methods. We estimate the selection-corrected means from the pre-policy ACT score distribution using the methods described in Section 2.2 and Appendix B. We construct the benchmark distribution from the post-policy ACT score distribution using the methods described in Section 2.3. We report all results in Table 3 and summarize these results in Figure 2.

In Table 3, we report the mean for the raw post-policy ACT score distribution (column 1), the reweighted post-policy distribution (column 2), and the reweighted post-policy distribution with missing scores replaced by predicted scores (column 3). These provide three measures, as discussed in Section 2.3, of the benchmark latent ACT distribution to which we compare the selection-corrected pre-policy ACT distribution. For example, the mean ACT score is 19.25 in the raw post-policy data, 19.73 after reweighting, and 19.56 after predicting missing values.[18] We report the mean from the observed distribution in column 4 and from the selection-corrected distributions in columns 5-12. Readers can directly compare these selection-corrected statistics to their preferred benchmark in columns 1-3.

Our first selection correction method uses a simple linear regression adjustment: we regress

[18]Reweighting raises the mean because the fraction of students eligible for free and reduced-price lunch is higher post-policy. The predicted mean is slightly lower than the reweighted mean because the 1.5% of students who do not take the ACT post-policy period are negatively selected on observed characteristics.

Table 3. Mean Latent ACT Score by Correction Method and Predictor Set

	Post-Policy ("Truth")			Pre-Policy (Biased)		Pre-Policy, by Correction Method						
							Heckman		Newey		Powell	
	Raw	DFL	OLS	Raw	OLS	Tobit	No IV	With IV	Series Lgt	N.P.	Series Lgt	N.P.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Student Demographics	19.25	19.73	19.56	20.86	20.67	20.62	20.67	20.67	20.66	20.67	20.65	20.71
			(0.11)		(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.12)	(0.10)
…Plus School-Level Covs	19.25	19.73	19.77	20.86	20.48	20.37	20.50	20.48	20.49	20.49	20.52	20.49
			(0.13)		(0.09)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.10)
…Plus Student Test Scores	19.25	19.73	19.69	20.86	19.52	19.22	19.66	19.63	19.64	19.55	19.95	19.67
			(0.13)		(0.09)	(0.10)	(0.09)	(0.09)	(0.09)	(0.10)	(0.11)	(0.09)

Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. Columns (1) and (4) give raw mean ACT scores for each sample, and column (2) uses the DFL-weighted post-policy score distribution. Cells in columns (3) and (5) - (12) report the mean predicted ACT score from regressions of ACT scores on covariates. The predicted ACT score is calculated for ACT-takers and non-takers. Standard errors calculated using 500 bootstrap replications resampling schools.

observed test scores on a vector of student demographics and use the coefficients to predict test scores. The mean of the predicted values using OLS is 20.67 (standard error 0.10), shown in column 5. So OLS closes only 11% of the gap between the observed mean of 20.86 and the reference mean of 19.56. The poor predictive fit is unsurprising, as there is substantial heterogeneity within each conditional mean cell (e.g., within race groups) that we do not yet model.[19]

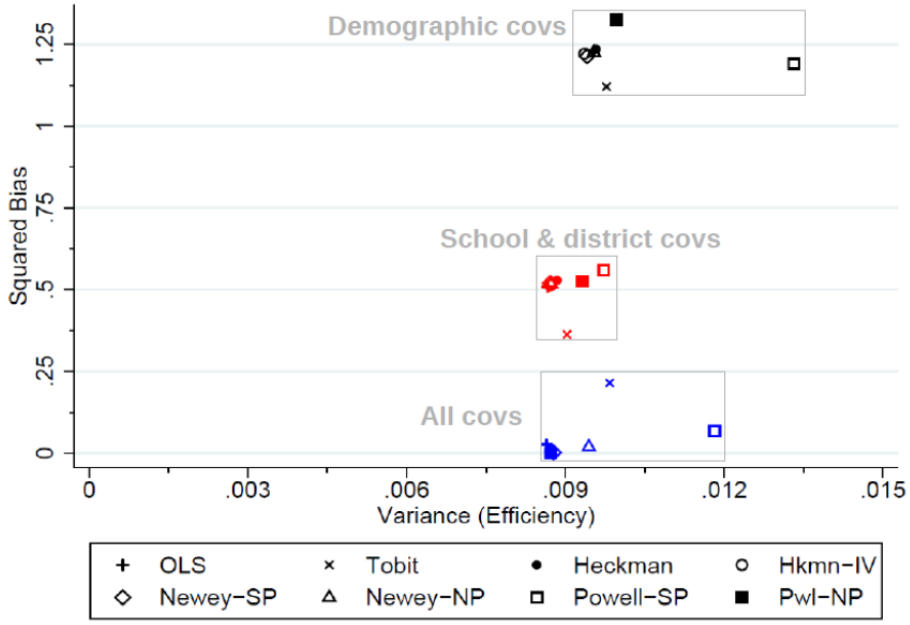
Our second selection correction is a Type 1 Tobit model, censoring at the 36th percentile of the post-policy ACT score distribution, as the test-taking rate in the pre-policy period is 64%. The predicted mean is similar to that from OLS. We next show results from the Heckman two-stage correction procedure in columns 7 and 8. When the test-taking model does not use an exclusion restriction, the mean predicted score is essentially identical to that predicted by OLS. Adding driving distance from students' home to the nearest ACT test center as a predictor of test-taking does not change the predicted mean ACT score. Finally, we implement the two-stage semiparametric sample selection corrections: the Newey and Powell models, each estimated using both the semiparametric and nonparametric first stages, including the driving distance instrument in all cases. See Appendix B for details on how we implement these estimators, including the data-driven choice of predictors in the series logit and functional form of the Newey correction term. We report the results using the Newey correction in columns 9 (semiparametric first stage) and 10 (nonparametric first stage). These results are almost identical to those from the Heckman correction, very similar to those from the OLS and Tobit corrections, and robust across different orders of polynomial selection correction terms. The Powell model yields similar results (with semiparametric first stage in column 11 and nonparametric first stage in column 12) and is marginally more biased with the nonparametric than the semiparametric first stage.

4.2 Comparing Selection Corrections' Performance with Different Predictors

We now examine whether a researcher who has access to school- and district-level covariates (such as demographics, urbanicity, and average 8th and 11th grade test scores) can do a better job at correcting for selection in ACT scores. We report these results in the second row of Table 3. Adding these controls moves the predicted mean closer to the reference mean for all methods. However, the predicted means still exceed the reference mean by at least 0.6 ACT

[19] Appendix Figure 1 shows the complete, selected, and latent test score distributions for subsamples by race and poverty, using the same approach as Figure 1. The latent score distributions for all subsamples span a similar range to the full sample, and remain quite skewed.

Figure II. MSE Comparison Across Correction Methods and Covariate Sets



Notes: Figure shows the mean squared error of each combination of correction method and covariate set from Table 3. Black (top): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the difference between the statistic predicted by a) the correction method applied to the pre-policy data and b) the post-policy, DFL-weighted, fitted distribution.

points (equal to 0.27 standard deviations). There is again no evidence that the semiparametric models outperform the parametric or single-equation models.

Finally, we include student-level 8th and 11th grade test scores in the prediction model. These data are often available to state education administrators, though researchers seldom have them matched to students' college entrance test scores. We report these results in the third row of Table 3. All the corrections perform much better using the student-level scores in the prediction. This reflects the strong relationship between students' past and contemporaneous achievement, ACT-taking, and ACT scores. The predicted means are mostly within 0.2 ACT points of the reference mean, though the Tobit and semiparametric Powell correction perform worse. Although the skewness of the latent test score distribution visible in Figure 1 caused serious problems for models with few predictors, this problem is much less serious with richer predictors.[20] The predicted mean is closest to the reference value for the nonparametric Powell

[20]We also implement this exercise using 8th grade test scores as predictors but omitting 11th grade test scores. The relative predictive accuracy of different models, reported in Appendix Table 11, is unchanged.

model. The more flexible models do not robustly outperform the parametric models, single-equation models, or even simple OLS.[21]

We summarize these results in Figure 2. We show each of the 24 predicted ACT means generated by the 8 selection correction models and 3 predictor sets in a bias-variance scatterplot. This allows us to visually compare the bias and variance of the model-predictor combinations. Points closer to the origin estimate the mean with lower mean squared error. The predictions relying on only student demographics (black points) or student demographics and school-/district-level characteristics (red points) are consistently high on the bias axis, reflecting their poor ability to replicate the benchmark ACT mean. The predictions that include student test scores are less biased and have similar variance. Within each covariate set, there is little variation in bias or variance across different selection correction methods, except the semiparametric Powell correction, which has consistently higher variance. This figure clearly demonstrates that if we seek to minimize mean-squared error (or any reasonable weighted average of bias and variance), better data is valuable and more flexible methods are less so. In particular, our results show that robustness of results to different modeling choices, a common feature of empirical papers, is not necessarily reassuring.

4.3 Comparing Selection Corrections on Other Criteria

We estimate the parameter vector from a linear regression of the non-missing pre-policy ACT scores on the predictors for each of the 8 selection correction models and 3 predictor sets. We compare each of these to the estimated parameter vector using the complete test scores from the post-policy period and interpret this as a measure of how well each correction model addresses selection-induced bias in parameter estimates. We discuss these comparisons in detail in Appendix C. In brief, we find that the mean-squared bias across all parameter estimates is lower with richer sets of predictors but not for more flexible econometric models. We observe the lowest mean squared bias with OLS (i.e. without any selection correction). We conclude that for both prediction and parameter estimation, the gains from using less restrictive econometric methods are small relative to the gains from seeking richer or more disaggregated data. We

[21] We conduct several robustness checks where we further vary the set of predictors. We describe these checks in Appendix D with results presented in Appendix Table 12. Our main findings are robust to including squared and interacted predictors in the ACT-taking and ACT score models, using different combinations of the individual, school-, and district-level predictors, and relaxing the assumption that the predictors and selection correction terms are additively separable in the ACT score model.

find a similar result when we compare the full distribution of selection-corrected pre-policy test scores to the post-policy distribution.

4.4 Comparing Selection Corrections for Different Subgroups

We also evaluate how well individual-level selection correction models predict the mean latent test score for four subgroups. This is of interest for two reasons. First, researchers, administrators, and policymakers are interested in latent scores for key student subgroups in addition to the full population. Second, econometricians, applied and theoretical, are interested in how well selection correction models perform across different data generating processes. The latent ACT score distributions, ACT-taking rates, and the distributions of predictors differ substantially for black, white, low-income, and higher-income students (see Appendix Figure 1). If the main pattern of results that we find for the overall sample holds across these subgroups, this shows that the results are not specific to a single data generating process and may be more generalizable. Reassuringly, we find that the main results hold across different subgroups. For all subgroups, as in the overall sample, we find that the choice of correction method makes little difference, but that corrections perform substantially better when including richer covariates (see Appendix Figure 7 and Appendix Table 10). This robustness across different data generating processes addresses some concerns about the generalizability of our findings.

We present results for all eight selection correction models estimated separately by race and free-lunch status, using the full set of predictors, in Table 4 (summarized in Appendix Figure 2). There are large gaps in mean observed ACT scores between black and white students and between low-income and higher-income students in the pre-policy period. In the post-policy period, the test-taking rate rises for all groups. The gap in the test-taking rate between low-income and higher-income students narrows, but the gap between black and white students remains approximately constant. The rise in test-taking rates is associated with a fall in mean test scores for all four subgroups. All selection correction models, applied to all four subgroups, raise the predicted mean score relative to the observed data. However, many of the models overestimate the predicted mean, particularly for black and low-income students. The gaps in performance by race and by income are therefore underestimated; some models actually estimate gaps that are farther from the truth than the observed gap. This pattern is more pronounced for the income gap than the race gap.

What might explain this result? Recent research shows that past achievement is less pre-

Table 4. Race and Poverty Gaps in Mean Latent ACT Scores by Correction Method

	Black	White	Gap	Poor	Non-Poor	Gap
	(1)	(2)	(3)	(4)	(5)	(6)
Post-Policy						
Raw	15.61	19.98	4.38	16.77	20.19	3.42
DFL	15.95	20.28	4.33	16.84	20.46	3.62
OLS	15.86	20.27	4.41	16.78	20.43	3.65
	(0.26)	(0.11)	(0.28)	(0.08)	(0.12)	(0.12)
Pre-Policy						
Raw	16.76	21.44	4.68	18.29	21.28	3.00
OLS	16.04	20.07	4.03	17.21	20.12	2.91
	(0.19)	(0.08)	(0.20)	(0.08)	(0.09)	(0.10)
Tobit	15.87	19.79	3.92	16.94	19.90	2.95
	(0.19)	(0.08)	(0.20)	(0.09)	(0.09)	(0.11)
Heckman	16.08	20.18	4.10	17.31	20.22	2.91
(with IV)	(0.18)	(0.08)	(0.19)	(0.09)	(0.09)	(0.11)
Newey -	16.05	20.22	4.17	17.31	20.25	2.93
Series Logit	(1.42)	(0.08)	(1.43)	(0.10)	(0.09)	(0.11)
Newey -	16.00	20.16	4.16	17.15	20.18	3.03
Nonparametric	(0.18)	(0.08)	(0.19)	(0.09)	(0.09)	(0.11)
Powell -	16.27	20.41	4.14	17.39	20.46	3.06
Series Logit	(0.20)	(0.11)	(0.22)	(0.11)	(0.10)	(0.13)
Powell -	16.12	20.17	4.05	17.41	20.22	2.80
Nonparametric	(0.19)	(0.08)	(0.21)	(0.09)	(0.09)	(0.11)

Notes: The sample is as in Table 3. The table reports means of the predicted ACT score from regressions of ACT scores on the full set of covariates, including student-level 8th and 11th grade test scores. The predicted ACT score is calculated for ACT-takers and non-takers. Poverty status is proxied for using free or reduced-price lunch receipt measured during 11th grade. Standard errors calculated using 500 bootstrap replications resampling schools.

dictive of college application behavior among disadvantaged groups (Avery and Hoxby, 2013; Hyman, 2017; Dillon and Smith, 2017). This is consistent with our results. Among white and higher-income students we find that the corrections perform quite well after conditioning on student test scores, suggesting that such test scores are strongly predictive of ACT-taking and ACT scores. The fact that the models perform substantially worse among black and lower-income students even after conditioning on student test scores, suggests that such scores are less predictive of ACT-taking, which is a critical piece of the college application process.

Alternatively, the worse prediction among disadvantaged groups may reflect the nature of the quasi-experiment we study. Students required to take the ACT by a mandatory testing policy who do not anticipate applying to a four-year college may not exert as much effort as students who take the test voluntarily. Our selection corrections predict latent scores for these students using observed characteristics and a distance instrument that shifts the cost of taking the ACT but not the value of performing well in the ACT. This selection correction strategy will imperfectly account for heterogeneity in effort on the ACT. We risk predicting incorrectly high ACT scores for non-takers, particularly non-takers from disadvantaged groups with lower probabilities of attending college conditional on observed characteristics. This hypothesis would explain both our overprediction of ACT scores for disadvantaged subgroups (see Table 4) and our slight overprediction of ACT scores on average (see Table 3). However, we find no difference between periods in the share of students with the precise score they would obtain by random guessing. This shows that the students induced to take the ACT by the mandatory testing policy are not more likely to exert very low effort on the test. Even if this hypothesis holds, it does not explain why we see similar performance across different selection correction methods.

5 Explaining Similar Results across Different Corrections

Section 4 shows that different selection corrections methods predict similar mean ACT scores despite their different assumptions. In this section, we explore possible economic and statistical explanations for the similarities.

We begin by noting that different methods predict similar student-level ACT-taking and scores as well as similar mean ACT scores. Table 5 reports summary statistics for the predicted probabilities of taking the ACT for all first stages (probit with and without instruments, series logit, nonparametric) and the three predictor sets. The student-level predicted probabilities are

very similar across the series logit and the two probit models, with correlation coefficients ≥ 0.93. The correlations between the nonparametric model and other models are still ≥ 0.84. These high correlations help to explain the similarity of the predicted ACT score distributions across the different corrections. The student-level predicted ACT scores are also very highly correlated across models (see Appendix Table 5). The different correction models generate predicted ACT scores with correlations ≥ 0.97 when using only student demographics as predictors. Including student test scores and school- and district-level characteristics leaves all correlations ≥ 0.95. Table 5 also shows that the predicted probabilities cover the whole unit interval only if we use the richest set of predictors. When only student demographics are used as predictors, the predicted values from all models are coarse and seldom near 0 or 1. This limited variation in the predicted probabilities of ACT-taking contributes to the poor performance of selection corrections using weak predictors.

This shows that the similarity in predicted mean ACT scores and coefficients in ACT regressions is explained by similar student-level predicted test-taking probabilities and scores. But why do the different corrections deliver such similar predictions? We consider and reject four possible explanations. First, there may be no sample selection problem. If test-taking is not influenced by unobserved characteristics that also influence test scores, then the selection corrections are unnecessary. We can reject this explanation. The distributions of observed and latent scores in Figure 1 show clear evidence of negative selection into test-taking. Further, the selection correction terms in both the Heckman and Newey models are large and significant predictors of ACT scores (see Appendix Tables 6, 7, and 8).[22]

Second, there may be a sample selection problem, but the structure of the problem may satisfy the parametric assumptions of the Tobit or Heckman models. In particular, the Heckman model is appropriate if the unobserved factors determining ACT scores and ACT-taking are jointly normally distributed. The latent test score distribution in Figure 1 is not normal, and we verify this with parametric (skewness-kurtosis) and nonparametric (Kolmogorov-Smirnov) normality tests.[23] The latent distribution is also non-normal conditional on demographic characteristics

[22]The inverse Mills ratio term in the Heckman model has a zero coefficient if the unobserved determinants of test-taking and test scores are uncorrelated. We reject the hypothesis of a zero coefficient for models with all combinations of the predictors and the instrument ($p < 0.001$). The coefficients are large: moving from the 5th to the 95th percentile of the predicted probability of ACT-taking shifts the ACT score by 10-13 points. We also test if the coefficients on all of the polynomial correction terms in the Newey model are zero. We reject this hypothesis for all combinations of predictors ($p < 0.005$).

[23]The rejection of normality is not explained by our large sample size. We also consistently reject normality for random 1% subsamples of the data.

Table 5. Cross-Model Comparison of First Stage Predicted Probabilities by Covariate Set

	Basic Demographics				Plus School Demographics				Plus Individual Test Scores			
	Probit No IV	Probit With IV	Series Logit	Non-Parametric	Probit No IV	Probit IV	Series Logit	Non-Parametric	Probit No IV	Probit IV	Series Logit	Non-Parametric
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Percentiles												
1%	0.380	0.341	0.316	0.300	0.199	0.199	0.153	0.172	0.051	0.051	0.062	0.143
5%	0.381	0.385	0.377	0.370	0.340	0.338	0.333	0.331	0.171	0.171	0.164	0.284
10%	0.478	0.438	0.439	0.430	0.411	0.412	0.406	0.409	0.271	0.270	0.254	0.365
25%	0.646	0.614	0.600	0.580	0.551	0.550	0.529	0.532	0.471	0.471	0.447	0.517
50%	0.646	0.665	0.669	0.670	0.665	0.665	0.663	0.665	0.684	0.684	0.681	0.684
75%	0.735	0.734	0.741	0.740	0.753	0.754	0.765	0.771	0.847	0.847	0.868	0.829
90%	0.735	0.751	0.759	0.780	0.828	0.828	0.849	0.866	0.938	0.939	0.953	0.925
95%	0.735	0.758	0.761	0.800	0.869	0.865	0.886	0.916	0.968	0.969	0.976	0.958
99%	0.822	0.817	0.806	0.840	0.917	0.919	0.937	0.965	0.993	0.993	0.994	0.986
Correlations												
Probit, No IV	1.000				1.000				1.000			
Probit, With IV	0.985	1.000			0.998	1.000			0.999	1.000		
Series Logit	0.962	0.976	1.000		0.930	0.932	1.000		0.962	0.963	1.000	
Nonparametric	0.886	0.899	0.922	1.000	0.842	0.843	0.896	1.000	0.849	0.850	0.885	1.000
Fraction Correct Predictions	0.642	0.647	0.650	0.652	0.665	0.666	0.672	0.672	0.727	0.727	0.738	0.704

Notes: Table reports descriptive statistics and correlations of the first stage predicted probabilities across selection models and by covariate set included as regressors. The fraction of correct predictions is the fraction of predicted probabilities that are rounded to 0 and 1 using a cutoff of 0.36, which is 1 minus the fraction taking a college entrance exam, match their observed test-taking indicator.

(see Appendix Figure 1) and the threshold censoring assumed by the Tobit model clearly does not hold, even conditional on demographic characteristics. We also test the assumption that the unobserved factors that affect latent test scores are normally distributed: we regress post-policy test scores on each of the three sets of predictors, generate the fitted residuals, and test whether they are normally distributed. We reject normality of all three sets of residuals using both Kolmogorov-Smirnov and skewness-kurtosis tests ($p < 0.001$ in all cases). We conclude that the structure of the selection problem, given the specification of the predictors, does not satisfy the joint normality assumption.[24]

Third, there may be sample selection that violates the parametric models' assumptions, but the test-taking predictors may be too coarse for the semiparametric models to perform well. Some semiparametric models are identified only if at least one predictor is strictly continuous (Ichimura, 1993; Klein and Spady, 1993). The series logit and Mahalanobis matching models we use do not have this requirement but their performance may still be poor if the data are all discrete or coarse. Coarse data may generate predicted probabilities that do not span the unit interval, limiting the effective variation in the selection correction terms.[25] This can explain the similarity in the ACT scores predicted by different models using only the discrete student demographics. But it does not explain the similarity across models using the richer set of predictors. The 8th and 11th grade student test scores are relatively continuous variables, which have respectively 1270 and 213 unique values, with no value accounting for more than respectively 1.3% and 2.5% of all observations.

Fourth, there may be a sample selection problem whose structure violates the assumptions of the parametric models, but the instrument may not be strong enough for the semiparametric models to perform well. The instrument satisfies conventional instrument strength conditions and does not predict other 11th grade test scores. However, the instrument does not satisfy "identification at infinity" (see Appendix B).[26] This means we can identify the slope coefficients

[24] As joint normality is a sufficient but not necessary condition for identification in the Heckman model, this test should be viewed as only partial evidence against the validity of the model assumptions.

[25] We show in Appendix Figure 3 that the predicted probability of ACT-taking has a narrow distribution and is linear in the predictors when we use only student demographics. This helps explain why two-stage corrections using only student demographics perform poorly: the correction terms are highly colinear with the predictors in the ACT regression. This relationship becomes nonlinear when we use richer predictor sets.

[26] In the probit model with the full set of predictors, moving from the 5th to the 95th percentile of the instrument (14.1 miles) lowers the probability of test-taking by 4 percentage points. The relationship is similar for the series logit model. Standard identification arguments for β_0 require an instrument that shifts the test-taking probability from 0 to 100 (Andrews and Schafgans, 1998; Chamberlain, 1986; Heckman, 1990).

in equation (1a) but cannot separately identify the intercept coefficient β_0 from the level of the selection correction term. This is not necessarily a problem for our analysis, which examines the mean predicted test score and is not interested in separating the intercept coefficient from the selection correction term. We view this as a natural feature of semiparametric selection models in many settings, rather than a feature specific to this application. The relationship between our instrument and participation measure is at least as strong as in many classic education applications (Card, 1995; Kane and Rouse, 1995). However, we acknowledge that the relative performance of different selection models may differ when an extremely strong instrument is available that permits identification of β_0.

We conclude that there is a selection problem whose structure is not consistent with the assumptions of the parametric models and that the data are continuous enough to use semiparametric analysis. The instrument does not support identification of the intercept coefficient in the ACT model but this does not explain why parametric and semiparametric methods perform similarly well at estimating slope coefficients. It appears that the violations of the more restrictive models' assumptions are not quantitatively important in this setting.[27]

6 Conclusion

Sample selection arises when outcomes of interest are not observed for part of the population and the latent outcomes differ for the cases with observed and unobserved values. Econometricians and statisticians have proposed a range of parametric and semiparametric methods to address sample selection bias, and applied researchers routinely implement these methods, but there is little evidence on their relative performance. We use a Michigan policy that changed ACT-taking for 11th graders from voluntary to required to observe partially missing outcomes for one cohort and complete outcomes for another cohort. We evaluate how well different selection corrections, applied to the partially missing outcomes, can match the complete outcomes.

We show that none of the sample selection corrections perform well when using only basic demographic information as predictors. With more information about students, particularly scores on state-administered standardized tests, simple OLS regressions perform well and there are few gains from using more flexible selection correction methods. This result holds when

[27]Vella (1998) also finds that parametric and semiparametric selection models produce similar results even when the assumptions of the parametric models fail. He uses real data but without a quasi-experimental benchmark. However, Goldberger (1983), Heckman, Tobias, and Vytlacil (2003), and Paarsch (1984) show that some parametric models perform poorly in simulations when their assumptions are violated.

we evaluate selection corrections on their ability to predict the mean outcome, predict the complete outcome distribution, or match the parameters of regression models estimated with the complete data. Predictions are more accurate for white and higher-income students than for black and lower-income students, leading to incorrect predictions of latent achievement gaps. Finally, group-level correction methods perform poorly across different model specifications. Aggregating the groups to increasingly refined cells, in particular cells defined by prior test scores, substantially improves performance.

What, if any, more general implications can be drawn from our findings? Our results may not generalize to very different settings, such as selection into wage employment (Heckman, 1974), selection into education levels (Willis and Rosen, 1979), or selection into different occupations or industries (Roy, 1951). However, two aspects of our results may be useful for other researchers. First, we find that performance depends heavily on the richness of the predictors. Regressing pre-policy ACT scores on the three sets of predictors – basic, district/school, and student test scores – yields R^2 values of respectively 0.134, 0.198, and 0.614. Regressing ACT-taking on the instrument and the three sets of predictors yields pseudo-R^2 values of 0.045, 0.088, and 0.223 respectively. Researchers estimating selection corrections with models that explain only a small fraction of the variation in the outcome should be very cautious. In a labor economics context, our results suggest that correcting wage distributions or regressions for selection will work better when lagged wage data is available as a predictor.[28] This reinforces findings in the treatment effects literature emphasizing the importance of rich data for estimating treatment effects in non-experimental settings (Heckman, Ichimura, Smith, and Todd, 1998; Heckman and Smith, 1999).

Second, our findings are not limited to settings where the assumptions of parametric selection correction models hold. We find strong evidence of quantitatively important selection on latent test scores, in a form that does not satisfy the assumptions of the parametric models we implement. The predictors are continuous enough to allow semiparametric estimation, and the instrument is comparable in strength to other widely-used instruments. This is a setting where we would expect semiparametric models to outperform parametric models. However, the gains from using these more flexible methods are minimal. Researchers who believe that parametric

[28]This echoes results in labor economics that lagged earnings are a particularly important control variable when matching methods are used for program evaluation (Andersson, Holzer, Lane, Rosenblum, and Smith, 2016; Lechner and Wunsch, 2013). Though see Heckman and Smith (1999) for a cautionary discussion.

model assumptions do not fit their application should not necessarily conclude that they will do better by estimating more flexible methods.

References

Abadie, A., and G. Imbens (2008): "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76(6), 1537–1557.

Ahn, H., and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

Altonji, J., H. Ichimura, and T. Otsu (2012): "Estimating Derivatives in Nonseparable Models with Limited Dependent Variables," *Econometrica*, 80(4), 1701–1719.

Andersson, F., H. Holzer, J. Lane, D. Rosenblum, and J. Smith (2016): "Does Federally-Funded Job Training Work? Nonexperimental Estimates of WIA Training Impacts Using Longitudinal Data on Workers and Firms," Working Paper 6071, CESifo.

Andrews, D., and M. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65(3), 497–517.

Angrist, J., E. Bettinger, and M. Kremer (2006): "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*, 96(3), 847–862.

Arellano, M., and S. Bonhomme (2017): "Quantile Selection Models with an Application to Understanding Changes in Wage Inequality," *Econometrica*, 85(1), 1–28.

Avery, C., and C. Hoxby (2013): "The Missing 'One-Offs': The Hidden Supply of Low-Income, High-Achieving Students for Selective Colleges," *Brookings Papers on Economic Activity, Economic Studies Program, The Brookings Institution*, 46(1), 1–65.

Bertrand, M., C. Goldin, and L. Katz (2010): "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics*, 2(3), 228–255.

Bonhomme, S., G. Jolivet, and E. Leuven (2016): "School Characteristics and Teacher Turnover: Assessing the Role of Preferences and Opportunities," *Economic Journal*, 126(594), 1342–1371.

Bulman, G. (2015): "The Effect of Access to College Assessments on Enrollment and Attainment," *American Economic Journal: Applied Economics*, 7(4), 1–36.

Busso, M., J. DiNardo, and J. McCrary (2014): "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *Review of Economics and Statistics*, 96(5), 885–897.

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Returns to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by C. Louis, K. Grant, and R. Swidinsky. University of Toronto Press, Toronto.

Card, D., and A. Payne (2002): "School Finance Reform, the Distribution of School Spending, and the Distribution of Student Test Scores," *Journal of Public Economics*, 83, 49–82.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

Clark, M., J. Rothstein, and D. Whitmore Schanzenbach (2009): "Selection Bias in College Admissions Test Scores," *Economics of Education Review*, 26, 295–307.

Dahl, G. (2002): "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica*, 70(6), 2367–2420.

Dehejia, R., and S. Wahba (1999): "Reevaluating the Evaluation of Training Programmes," *Journal of the American Statistical Association*, 94(448), 1053–1062.

Dillon, E., and J. Smith (2017): "The Determinants of Mismatch between Students and Colleges," *Journal of Labor Economics*, 35(1), 45–66.

DiNardo, J., N. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64(5), 1001–1044.

Dynarski, M. (1987): "The Scholastic Aptitude Test: Participation and Performance," *Economics of Education Review*, 6(3), 263–273.

Dynarski, M., and P. Gleason (1993): "Using Scholastic Aptitude Test Scores as Indicators of State Educational Performance," *Economics of Education Review*, 12(3), 203–211.

Frölich, M., M. Huber, and M. Wiesenfarth (2015): "The Finite Sample Performance of Semi- and Nonparametric Estimators for Treatment Effects and Policy Evaluation," Discussion Paper 8756, IZA.

Goldberger, A. (1983): "Abnormal Selection Bias," in *Studies in Econometrics, Time-Series and Multivariate Statistics*, ed. by S. Karlin, T. Amemiya, and L. Goodman. Academic Press, New York.

Gronau, R. (1974): "Wage Comparisons – A Selectivity Bias," *Journal of Political Economy*, 82(6), 1119–1143.

Hanushek, E., and L. Taylor (1990): "Alternative Assessments of the Performance of Schools: Measurement of State Variations in Achievement," *Journal of Human Resources*, 25(2), 179–201.

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42(4), 679–694.

——— (1990): "Variation of Selection Bias," *American Economic Review*, 80(2), 313–318.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J., and J. Smith (1999): "The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies," *Economic Journal,*, 109, 313–348.

Heckman, J., J. Tobias, and E. Vytlacil (2003): "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85(3), 748–755.

Heckman, J. J., and R. Robb, Jr. (1985): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of Econometrics*, 30(1-2), 239–267.

Hyman, J. (2017): "ACT for All: The Effect of Mandatory College Entrance Exams on Postsecondary Attainment and Choice," *Education Finance and Policy*, 12(3), 281–311.

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120.

Imbens, G. (2003): "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review Papers and Proceedings*, 93(2), 126–132.

Kane, T., and C. Rouse (1995): "Labor Market Returns to Two-Year and Four-Year Colleges," *American Economic Review*, 85(3), 600–614.

Klein, R., and R. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61(2), 387–421.

Krueger, A., and D. Whitmore (2001): "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal*, 111, 1–28.

LaLonde, R. . (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

Lechner, M., and C. Wunsch (2013): "Sensitivity of Matching-based Program Evaluations to the Availability of Control Variables," *Labour Economics*, 21, 111–121.

Lee, D. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Manski, C. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80(2), 319–323.

Melenberg, B., and A. Van Soest (1996): "Parametric and Semi-Parametric Modeling of Vacation Expenditures," *Journal of Applied Econometrics*, 11, 59–76.

Mroz, T. (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–800.

Newey, W. (2009): "Two Step Series Estimation of Sample Selection Models," *Econometrics Journal*, 12, S217–S229.

Newey, W., J. Powell, and J. Walker (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review Papers and Proceedings*, 80(2), 324–328.

Oster, E. (2017): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business and Economic Statistics*, forthcoming.

Paarsch, H. (1984): "A Monte Carlo Comparison of Estimators for Censored Regression Models," *Journal of Econometrics*, 24, 197–213.

Powell, J. (1987): "Semiparametric Estimation of Bivariate Latent Variable Models," Working Paper 8704, Social Systems Research Institute, University of Wisconsin, Madison.

Puhani, P. (2002): "The Heckman Correction for Sample Selection and its Critique," *Journal of Economic Surveys*, 14(1), 53–68.

Roy, A. D. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3(2), 135–46.

Rubin, D. (1976): "Inference and Missing Data," *Biometrika*, 63(3), 581–592.

——— (1987): *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York.

Semykina, A., and J. Wooldridge (2013): "Estimation of Dynamic Panel Data Models with Sample Selection," *Journal of Applied Econometrics*, 28(1), 47–61.

Stock, J., and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, ed. by J. Stock, and D. Andrews. Cambridge University Press.

Tobin, J. (1958): "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26(1), 24–36.

Vella, F. (1998): "Abnormal Selection Bias," *Journal of Human Resources*, 33(1), 127–169.

Willis, R., and S. Rosen (1979): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Journal of Political Economy,*, 87(5), S7–S36.