Quasi-Experimental Evaluation of Alternative Sample Selection Corrections*

Robert Garlick† and Joshua Hyman‡

February 3, 2021

Abstract

Researchers routinely use datasets where outcomes of interest are unobserved for some cases, potentially creating a sample selection problem. Statisticians and econometricians have proposed many selection correction methods to address this challenge. We use a natural experiment to evaluate different sample selection correction methods' performance. From 2007, the state of Michigan required that all students take a college entrance exam, increasing the exam-taking rate from 64 to 99% and largely eliminating selection into exam-taking. We apply different selection correction methods, using different sets of covariates, to the selected exam score data from before 2007. We compare the estimated coefficients from the selection-corrected models to those from OLS regressions using the complete exam score data from after 2007 as a benchmark. We find that less restrictive semiparametric correction methods typically perform better than parametric correction methods but not better than simple OLS regressions that do not correct for selection. Performance is generally worse for models that use only a few discrete covariates than for models that use more covariates with less coarse distributions.

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. We thank the editor, associate editor, and three anonymous reviewers for helpful comments. 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. Address: Department of Economics, Amherst College, Amherst, MA 01002-5000; Email: jhyman@amherst.edu; Telephone: (413) 542-5537

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. From 2007, the state of Michigan required that all students take a college entrance exam, increasing the exam-taking rate from 64 to 99% and largely eliminating selection into exam-taking. We apply different selection correction methods, using different sets of covariates, to the selected exam score data from before 2007. We then compare the estimated coefficients from the selection-corrected models to those from OLS regressions using the complete exam score data from after 2007 as a benchmark. Our primary performance metric is the mean squared bias across all coefficients, but we also examine two coefficients of particular policy relevance: an indicator for Black student race and an indicator for free or reduced-price lunch receipt, representing respectively the race and income gaps in ACT scores.

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; Heckman and Robb, 1985a; 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 different sets of covariates, which we include in both the outcome and selection equations. These mimic the different types of data available to education researchers, ranging from sparse (student demographics) to rich (student demographics, lagged student test scores, and school and district characteristics). In the two-stage models requiring an exclusion restriction, we use two instruments that affect physical access to test-taking: the driving distance from a student's home to the nearest test center, and the number of ACT testing dates with a severe weather

event near the local testing center in the 24 hours leading up to the exam. We show that after controlling for other covariates, these instruments strongly predict ACT-taking but have little relationship with other measures of student achievement.

We find that less restrictive semiparametric methods typically perform better than parametric correction methods but not better than simple OLS regressions that do not correct for selection. No one correction stands out as the strongest performer in all cases, though the parametric bivariate normal correction without an instrument usually performs the worst. Performance is generally worse for models that use only a few discrete covariates than for models that use more covariates with less coarse distributions.

We consider several explanations for why the semiparametric corrections do not perform better than simply ignoring the selection problem and using OLS regressions. This is not explained by an absence of selection or weak instruments. The distributional assumptions of the parametric methods do not hold in our data, and the bivariate normal selection correction terms are almost colinear with the second stage covariates. This may explain their high bias relative to the semiparametric methods. The improved performance of most models when we add more detailed covariates is consistent with Angrist et al. (2013) and Angrist et al. (2017), who find that observational value-added models fairly reliably predict causal school effects, as well as with Newey et al. (1990), who conclude that the set of covariates matters more than the specification of the selection equation.

This is the first paper to evaluate the performance of selection correction methods for missing data against a quasi-experimental benchmark. Other papers whose main focus is studying missing data problems by comparing estimates across selection correction methods lack an external benchmark for evaluation (Mroz, 1987; Newey et al., 1990; Melenberg and Van Soest, 1996). These papers focus on evaluating the assumptions of different methods or examining how much estimates change across methods. In contrast, we examine how selection-corrected estimates compare to estimates using a benchmark measure of the complete data. We use Michigan's change in test-taking policy through time as a natural experiment to provide the external benchmark. The validity of our benchmark relies on a temporal stability assumption: that the distribution of unobserved student characteristics in Michigan does not change between cohorts. We present some evidence to indirectly support this assumption: differences in observed characteristics between cohorts are small, accounting for differences in observed characteristics does not change any of our main findings, and there is no difference in ACT

scores between cohorts within the pre-reform period (when we observe selected scores for all cohorts) or within the post-reform period (when we observe complete scores for all cohorts). Our approach is similar to the literature comparing different treatment effects methods, including some selection correction methods, against experimental benchmarks (LaLonde, 1986; Heckman et al., 1998; Dehejia and Wahba, 1999). Our approach is also related to an empirical literature that compares estimates using potentially selected survey data to estimates using more complete administrative data (Finkelstein et al., 2012; Meyer and Mittag, 2019).

We examine selection correction methods' performance in a single data set, and thus evaluate their performance using only a single empirical example. These patterns may not generalize to other empirical examples. Subject to this caveat, this exercise may be of interest to three audiences. First, our findings may be relevant for applied researchers using selection correction methods or adapting existing methods for new applications (e.g. Dahl 2002; Bonhomme et al. 2016; Krueger and Whitmore 2001; Card and Payne 2002; Angrist et al. 2006; Clark et al. 2009). Our findings provide an example where uncorrected OLS regressions perform better than the parametric corrections employed by some of these studies.

Second, our findings may be relevant to econometricians comparing selection correction performance (e.g., Mroz 1987; Goldberger 1983; Paarsch 1984; Newey et al. 1990; Vella 1998), developing selection correction methods, or extending methods to allow for features like nonparametric outcome models or dynamic selection (Das et al., 2003; Semykina and Wooldridge, 2013). We contribute to the work comparing correction performance by providing an example with an empirical benchmark that allows us to evaluate rather than compare performance.

Third, our findings may be of interest to researchers, practitioners, and policymakers who use college entrance exam scores to infer population achievement. For example, school district and state education administrators often compare scores over time or across different race and income groups, while researchers often use them as an outcome to examine the impact of some education treatment. Our results contribute to the literature on selection into college entrance exam-taking (Dynarski, 1987; Hanushek and Taylor, 1990; Dynarski and Gleason, 1993), by showing an example where college entrance exam scores come closer to correctly describing cross-group differences in population achievement when richer covariates are observed.

We describe the sample selection problem and correction methods in Section 2. In Section 3, we describe our data, our setting, and the extent of selection into test-taking in the pre-reform period. We report the main findings in Section 4 and discuss possible reasons for these findings

in Section 5. We conclude in Section 6 and reflect on some alternative approaches to selection correction not emphasized in our paper.

2 Sample Selection, Corrections, Evaluation Criteria

2.1 The Sample Selection Problem

We introduce the sample selection problem using a common application 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^* and ACT_i are respectively the latent and observed ACT score of student i. We assume throughout the paper that the conditional mean function $\mathbb{E}[ACT_i^*|X_i]$ is linear and that the objects of interest are the conditional means of ACT_i^* given X_i (i.e. the parameters from the population linear regression of ACT_i^* on X_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 "covariates" of test

scores rather than "determinants" or "causes."

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, conditional on X_i, influence the probability of taking the ACT and do not influence latent ACT scores (i.e. are independent of ϵ_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. 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, 1985b). This still poses a sample selection problem but can be addressed using simpler methods.

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 unobserved factors influencing test-taking are independent of latent test scores and the excluded instruments Z_i do not influence test-taking, because the omitted variable in equation (2) is zero under this assumption. 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 bivariate normal selection correction method (Gronau, 1974; Heckman, 1974). If $g\left(X_i, Z_i\right) = X_i\delta + Z_i\gamma$ and (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).[1] As our fourth model, we therefore estimate a bivariate normal selection correction model using two instruments: the driving distance from each student's home to the nearest ACT test center from the outcome model, and the number of ACT testing dates with a severe weather event near the closest ACT test center in the 24 hours leading up to the exam. 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 semiparametric 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 weighted mean ACT-taking rate among groups of students with similar covariate and instrument values, with more weight assigned to students with the most similar values of the covariates and instruments. We use the predicted probabilities of ACT-taking from these models 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 $T\hat{A}KE_i^*$ (Heckman and Robb, 1985a; Newey, 2009). The second differences out the bias term using pseudo-fixed effects for groups of students with similar values of $T\hat{A}KE_i$ (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.[2]

[1] 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.

[2] The differencing methods proposed by Ahn and Powell (1993) and Powell (1987) yield $\sqrt{n}$-consistent and asymptotically normal estimators of the regression coefficients in the ACT score model if the ACT-taking

We refer to these eight methods as OLS, Tobit, bivariate normal, bivariate normal with IV, semiparametric + polynomial, nonparametric + polynomial, semiparametric + differencing, and nonparametric + differencing. 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 factors influencing 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 bivariate normal corrections are appropriate if this interest is correlated with unobserved factors influencing her latent ACT score but the joint distribution of these unobserved characteristics satisfies specific parametric conditions. The polynomial and differencing corrections are appropriate if this interest is correlated with unobserved factors influencing her latent ACT score and the joint distribution of these unobserved characteristics satisfies weaker conditions.

2.3 Evaluating Alternative Selection Correction Methods

We evaluate each of the eight selection correction methods by comparing estimates of regression coefficients between selection-corrected pre-reform data and complete post-reform data. First, we regress the complete post-reform ACT test score data on a vector of covariates. Second, we regress the selected pre-reform ACT test score data on the same vector of covariates, using each of the eight selection correction methods in turn. We interpret the difference between the first and second vectors of coefficient estimates as the selection biases after applying the relevant correction method. Our primary evaluation criterion is the mean squared bias across the full vector of coefficients (excluding the intercept) in each selection-corrected regression. We also examine two individual coefficients that are particularly policy-relevant: an indicator for Black student race and an indicator for free or reduced-price lunch status, respectively measuring the race and income gap in ACT scores.

ACT-taking is almost universal in the post-reform period, so the coefficients from the post-reform regression models have little selection bias. If the latent ACT score distribution is stable

model is undersmoothed. In our context, this means identifying the "correct" series order using cross-validation and then deliberately choosing a higher series order, and similarly using smaller groups of students in the nonparametric matching estimator. To address this concern, we show in Appendix Figures XII and XIII that undersmoothing does not systematically improve performance in the cases we examine.

from the pre- to the post-reform period, then the difference between the coefficients captures the selection bias that remains after applying a selection correction method. We assess the cross-cohort stability of the ACT score distribution in Section 3. In brief, we show that there are only small differences between cohorts in observed characteristics, that our findings are robust to adjusting for these differences using reweighting, and that there is no difference in ACT scores across cohorts within either the pre- or post-reform period. Our main results compare the selection-corrected coefficient estimates to the post-reform estimates after reweighting the post-reform data to equate the distribution of observed covariates between cohorts. But we show in Appendix Figures III - V that the results are nearly identical when compared to the post-reform coefficient estimates without reweighting.

We evaluate each of the eight selection correction methods using three different vectors of covariates, for a total of twenty-four estimates. In the main paper, we present the mean squared bias, and coefficients on the Black student race and free or reduced-price lunch receipt indicators, and their standard errors. In Appendix Tables 5 - 7, we report all of the estimated coefficients and their standard errors. We estimate the standard errors using a nonparametric bootstrap that replicates all estimation stages within each replication: estimating the reference model using the complete post-reform data, estimating the ACT-taking model using selected pre-reform data, and estimating the ACT score model using selected pre-reform data with the relevant selection correction method applied. We use bootstrap rather than analytical standard errors because our focus on mean squared bias requires standard errors for nonlinear combinations of estimates across multiple regressions.

3 Context, Data, and the Extent of Selection

We use student level data for two cohorts (2004/5 and 2007/8) of all first-time 11^{th} graders attending Michigan public high schools. Using the last pre-reform cohort (2005/6) and first post-reform cohort (2006/7) 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 2005/6 and 2006/7 cohorts, our main analysis uses the 2004/5 and 2007/8 cohorts. We refer to these as the 2005 and 2008 cohorts in the rest of the paper. Our main results are similar when we compare either of the two pre-reform cohorts to either of the two post-reform cohorts (Appendix Figures VI - VIII).

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 time-varying characteristics such as free and reduced-price lunch status and student home address. 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, to the driving distance between students' home during 11th grade and the nearest ACT test center, and to information about the timing and location of severe weather in Michigan during our sample period.[3] 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.[4] Second, mean ACT scores did not vary across years within each period: they change by only 0.1 points between 2005 and 2006 and between 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.

Third, ACT-taking rates increased more for student groups with lower pre-reform 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 group-level positive selection of students into ACT-taking based on their latent ACT scores in the pre-reform period, which was eliminated by the reform. 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-reform cohort to have the same distribution of observed characteristics as the pre-reform cohort (DiNardo et al., 1996).[5] This adjustment does not account for cross-cohort differences

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

[4]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.

[5]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

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
Number of Exam Dates with Severe Weather	0.82	0.88	0.25	0.99	0.76	-0.12	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.

in unobserved factors influencing latent ACT scores.

3.2 Modeling ACT-Taking

The two-stage selection correction methods are identified either by distributional and functional form assumptions, which are seldom viewed as credible in empirical work, or by an excluded instrument that predicts ACT-taking but not latent ACT scores. We propose two instrumental variables. The first is the driving distance from each student's home to the nearest ACT test center. The second is exposure to severe weather events occurring in the county of the nearest ACT test center in the 24 hours prior to a relevant ACT testing date. 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 covariates. Distance instruments have been widely used in research on education participation, including standardized test-taking (Bulman, 2015; Card, 1995; Kane and Rouse, 1995). Weather instruments are widely used in applied microeconomics work, showing poor weather affects outcomes such as labor supply, voter turn-out, and political protest participation (Krishnaswamy, 2019; Fujiwara et al., 2016; Madestam et al., 2013). We do not claim that the instruments are perfect, but rather that they are 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. See Appendix A for more information on the construction and distribution of both instruments.

Both instruments strongly predict ACT-taking, supporting the instrument strength condition. To show this, we use pre-reform data to estimate a probit regression of ACT-taking on a quadratic in distance and dummies for exposure to one and two severe weather events. A quadratic in distance allows the marginal cost of ACT-taking to vary with distance, accounting for fixed costs of travel or increasing marginal cost of time. We use indicators for one and two severe weather events to allow for possible nonlinear effects of having multiple affected testing dates, and because no student is exposed to more than two severe weather events in the relevant time window. We report both heteroskedasticity-robust standard errors, as distance varies at the individual level, and county-clustered standard errors, as the weather measure varies at the county level.

and all interactions; and district enrollment, pupil-teacher ratio, local unemployment rate and all interactions. Results are robust to alternative reweighting models.

We report the results in Table 2 columns 1-4. Both driving distance and severe weather events are associated with lower ACT-taking. The negative relationships grow stronger as we control for student demographics, school- and district-level characteristics, and student scores on other tests. Using either standard error type, the instruments are jointly statistically significant (with all covariates, $\chi^2_{\text{robust}} = 184.23$ and $\chi^2_{\text{clustered}} = 48.42$). The instruments pass the commonly used "$F > 10$" test for instrument strength, although this test is developed for linear two-stage least squares models and is not formally applicable to this setting (Stock and Yogo, 2005). The probability of ACT-taking drops by 10-13 percentage points (depending on the covariate set) with a move from the 1$^{\text{st}}$ to the 99$^{\text{th}}$ percentile of the instruments. Over half of this shift is due to exposure to two severe weather events. We return to the interpretation of the instrument in Section 5, including a discussion of identification at infinity.

Neither instrument strongly predicts latent achievement, supporting the exclusion condition. To show this, we conduct a placebo test using two other measures of latent achievement that are observed whether or not students take the ACT: their math and English scores on the in-school standardized exam that all students take in eleventh grade. We regress the average of these two measures on the instruments and various sets of covariates. We report results in Table 2 columns 5-8. Neither set of instruments is statistically significantly associated with the outcomes conditional on covariates ($\chi^2_{\text{robust}} = 4.65$, $\chi^2_{\text{clustered}} = 3.51$). Latent achievement shifts by only 0.05 standard deviations with a move from the 1$^{\text{st}}$ to the 99$^{\text{th}}$ percentile of the instruments. This placebo test provides reassurance that the instruments are unlikely to substantially shift ACT scores conditional on ACT-taking.

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

In this subsection, we compare the observed pre- and post-reform ACT score distributions to describe pre-reform selection into ACT-taking. Positive/Negative selection occurs if pre-reform scores are systematically higher/lower than post-reform scores. Researchers using selected test scores sometimes assume that all non-takers would score below some percentile in the observed distribution (Angrist et al., 2006) or below all takers (Krueger and Whitmore, 2001). We assess the plausibility 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-reform cohort to have the same number of students and

Table 2. Testing the Exclusion Restriction: the Relationship Between Test Center Access, 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)	-0.002	-0.008	-0.006	-0.006	0.031	-0.002	0.001	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)
	[0.003]	[0.003]	[0.001]	[0.001]	[0.013]	[0.008]	[0.002]	[0.002]
Distance Squared (/ 10)	-0.000	0.002	0.002	0.002	-0.014	-0.002	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
	[0.001]	[0.001]	[0.001]	[0.001]	[0.005]	[0.003]	[0.001]	[0.001]
One Severe Weather Event	0.016	-0.000	-0.023	-0.019	0.034	0.026	-0.005	0.000
	(0.004)	(0.004)	(0.005)	(0.004)	(0.009)	(0.009)	(0.010)	(0.006)
	[0.022]	[0.019]	[0.016]	[0.016]	[0.062]	[0.048]	[0.020]	[0.020]
Two Severe Weather Events	-0.069	-0.081	-0.078	-0.079	-0.018	-0.033	-0.042	-0.035
	(0.008)	(0.008)	(0.008)	(0.007)	(0.017)	(0.016)	(0.016)	(0.011)
	[0.024]	[0.025]	[0.018]	[0.018]	[0.063]	[0.066]	[0.024]	[0.018]
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	0.002	0.046	0.089	0.228	0.004	0.110	0.203	0.647
Chi-2 Statistic - Robust SEs	216.62	301.31	142.04	184.23	60.26	33.47	1.88	4.65
Chi-2 Statistic - Clustered SEs	14.64	18.72	44.04	48.42	3.83	2.89	1.20	3.51
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) are probit models and columns (5)-(8) are OLS models. We report average marginal effects for the probit models. We report heteroskedasticity-robust standard errors in parentheses and standard errors clustered at the county level in brackets. For each standard error type, we report the Chi-2 statistic from a test that the coefficients on all four IV terms equal zero. 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. Severe weather events are the number of events that occur within 24 hours prior to a ACT/SAT testing date in the county of the test center during the students' 11th and 12th grade years. 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.

distribution of observed characteristics. If the reweighting accounts for all factors influencing latent test scores that differ between periods, then the difference in the number of students at each ACT score equals the number of non-takers with that latent score.[6]

Figure I plots the frequency distribution of ACT scores pre-reform, the reweighted post-reform distribution of scores, and the difference, which proxies for the latent scores of non-takers pre-reform. 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-reform ACT score is 1.3 points or 0.27 standard deviations higher than the mean post-reform ACT score. ACT-takers tend to score higher than non-takers. 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. Appendix Table 2 reports moments and percentiles of the three distributions.

There is clear positive selection into ACT-taking, but less than that assumed in prior studies. Angrist et al. (2006) and Krueger and Whitmore (2001) use Tobit and bounding analyses that assume 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.[7]

4 Results

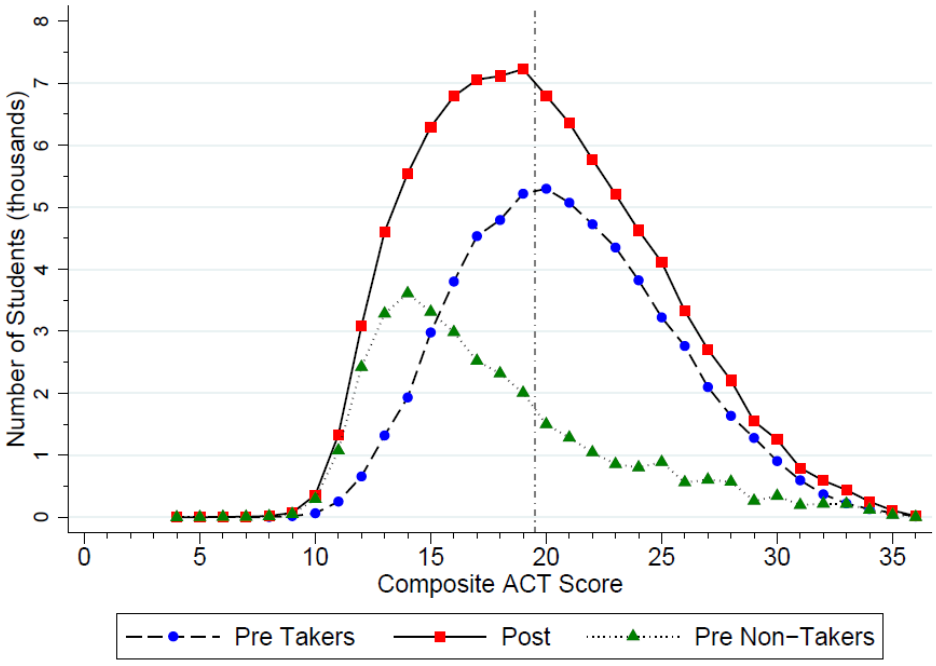
4.1 Comparing Mean Squared Bias of Different Selection Corrections

In this section, we evaluate the performance of multiple selection correction methods. We estimate selection-corrected regressions of ACT scores on covariates using the pre-reform ACT data and the methods described in Section 2 and Appendix B. We compare the coefficient estimates from these regressions to the coefficient estimates from the same regressions using the complete post-reform ACT data, weighted to adjust for the small differences in composition across the pre- and post-reform periods. We interpret the difference between the coefficient

[6]Hyman (2017) conducts a more extensive version of this analysis, measuring the number of students in the pre-reform 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.

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

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-reform period (dashed line with blue circles) and the number of students attaining each ACT score in the post-reform period (solid line with red squares) after reweighting the post-reform data to have the same distribution of observed covariates as the pre-reform data (DiNardo et al., 1996). The difference between the two numbers (dotted line with green triangles) is an estimate of how many pre-reform 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.

estimates as measures of selection bias after applying each selection correction.

Table 3, row 1 reports the mean squared bias (MSB) for each of the eight selection correction methods, taking the mean over all coefficient estimates except the intercept. The MSB is shown separately for each of the three covariate vectors. We summarize these results in Figure II. The simple OLS regression, which ignores selection, has the second lowest MSB of any method: 0.380 (standard error 0.045). The MSB is larger for the other parametric selection corrections: 1.249 for Tobit (s.e. 0.096), 7.357 for the bivariate normal model without instruments (s.e. 3.475), and 2.264 for the bivariate normal model with the instruments (s.e. 0.356).

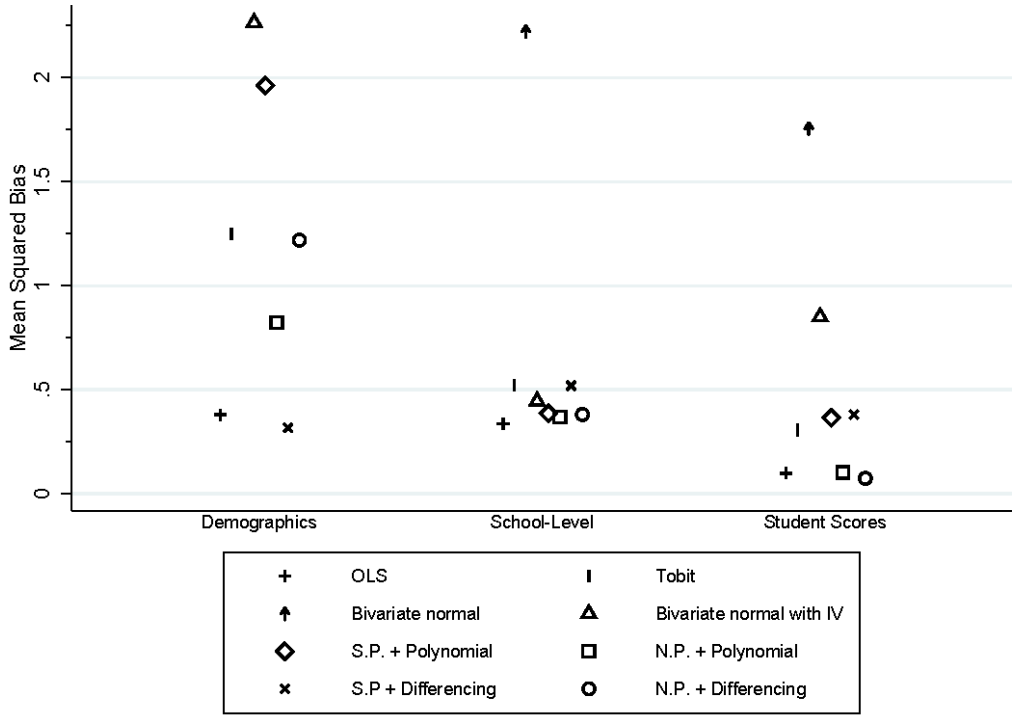
The semiparametric correction methods have substantially lower MSB than the parametric correction methods. The semiparametric differencing method has a slightly lower MSB than OLS: 0.317 with standard error 0.322, which is neither significantly not substantively different than OLS. The semiparametric polynomial, nonparametric polynomial, and nonparametric differencing methods have MSB of respectively 1.962 (s.e. 0.320), 0.823 (s.e. 0.095), and 1.219 (s.e. 0.154).

Table 3. Mean Squared Bias by Correction Method and Covariate Set

			Bivariate Normal		Polynomial		Differencing	
	OLS	Tobit	No IV	With IV	SP	NP	SP	NP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student Demographics	0.380	1.249	7.537	2.264	1.962	0.823	0.317	1.219
	(0.045)	(0.096)	(3.475)	(0.356)	(0.320)	(0.095)	(0.322)	(0.154)
Plus School/District-Level Covs	0.336	0.522	2.221	0.445	0.386	0.369	0.519	0.381
	(0.077)	(0.110)	(0.437)	(0.109)	(0.080)	(0.080)	(0.095)	(0.083)
Plus Student Test Scores	0.099	0.307	1.754	0.850	0.365	0.103	0.380	0.074
	(0.018)	(0.071)	(0.183)	(0.121)	(0.051)	(0.016)	(0.055)	(0.026)

Notes: The sample is as in Table 1, except only the 2005 cohort. Cells report the mean squared bias calculated over all coefficients except the intercept from a selection-corrected regression of ACT scores on covariates, where the bias is the difference between the coefficient and the reference value, estimated from a regression of post-reform ACT scores on the same covariates. Standard errors estimated using 500 bootstrap replications reported in parentheses.

Figure II. Mean Squared Bias by Selection Correction and Covariate Set



Notes: Figure shows the mean squared bias for every selection correction and covariate set, taken over all coefficient estimates except the intercept. We omit the bivariate normal correction without instruments with the student demographics covariate set. Including this estimate, which has MSB of 7.53, compresses the other estimates and makes them difficult to read.

We now examine whether these patterns change when the econometrician has access to school- and district-level covariates (such as demographic composition, average 8^{th} and 11^{th} grade test scores, class sizes, and local unemployment). We report these results in the second row of Table 3. Adding these covariates reduces the MSB for almost all of the methods, with larger reductions for the methods with the highest MSB over the sparse set of covariates. The improvement is particularly large for the very biased parametric methods: MSB for Tobit drops from 1.249 to 0.522, for bivariate normal without instruments drops from 7.537 to 2.221, and for bivariate normal with instruments falls from 2.264 to 0.445. Among the semiparametric methods, the bias reduction is also larger for the previously more biased semiparametric polynomial and nonparametric differencing methods, while the bias on previously least biased semiparametric differencing method rises by a small and statistically insignificant amount.

OLS has the lowest MSB for this covariate set. Three of the semiparametric methods (semiparametric polynomial and differencing, nonparametric differencing) have only slightly higher

MSB and the differences are not statistically significant. The bivariate normal method without instruments continues to have the largest bias, almost four times higher than any other method.

Finally, we add student-level 8^{th} and 11^{th} grade test scores to the set of covariates. We might expect these two measures to behave differently than the other covariates. These two measures explain 59% of the variation in ACT scores in a linear regression of selected scores, compared to 17% for all other covariates we observe. The pseudo-R^2 from regressing ACT-taking on these two measures is 0.19, compared to 0.06 for all other measures. In other education research, conditioning on lagged student test scores is particularly important for eliminating biases in value-added models (Angrist et al., 2013, 2017). Adding in these two covariates reduces MSB for all methods except the bivariate normal method with instruments, but the changes are small. The nonparametric differencing method now has the lowest mean squared bias, but this is almost identical to OLS and semiparametric differencing. The bivariate normal models with and without instruments have substantially and significantly higher MSB than any of the other methods.

Comparing MSB across different covariate sets is complicated by the different scales of the covariates. To address this concern, we rerun all of our analyses after standardizing all covariates to have mean zero and standard deviation one, and report the results in Appendix Figure II. Using standardized covariates, MSB is still substantially lower for the intermediate covariate set (using school- and district-level characteristics) than the basic covariate set (using only student demographics). MSB is marginally higher for the rich covariate set (using other student test scores) than the intermediate set. But the rise is explained mainly by the two bivariate normal models, with small changes in the mean squared biases for other methods that are small relative to the standard errors. We conclude that going from the basic to the intermediate covariate set robustly reduces MSB, but going from the intermediate to rich covariate set does not produce a robust reduction in bias. With standardized covariates, mean squared bias continues to be relatively high for the parametric estimators, and not systematically different between OLS and the less-biased semiparametric estimators.

The 95% confidence intervals on mean squared bias estimates exclude zero for every correction, using every set of covariates, except the semiparametric differencing model using the basic covariates. Although mean squared bias over multiple covariates does not have a natural quantitative interpretation, the statistical significance provides one metric to conclude that none of the correction methods entirely eliminates selection bias.

We summarize these results in Figure II. The figure shows the mean squared bias for each correction method on the vertical axis. The left-hand set of estimates are for the basic set of covariates, the middle set of estimates are for the intermediate covariates including school- and district-level information, and the right-hand set of estimates are for the rich covariate set including student-level test scores. Within each of the left, middle, and right blocks, the estimates for the different correction methods are staggered horizontally for legibility. But the horizontal axis has no cardinal or ordinal interpretation.

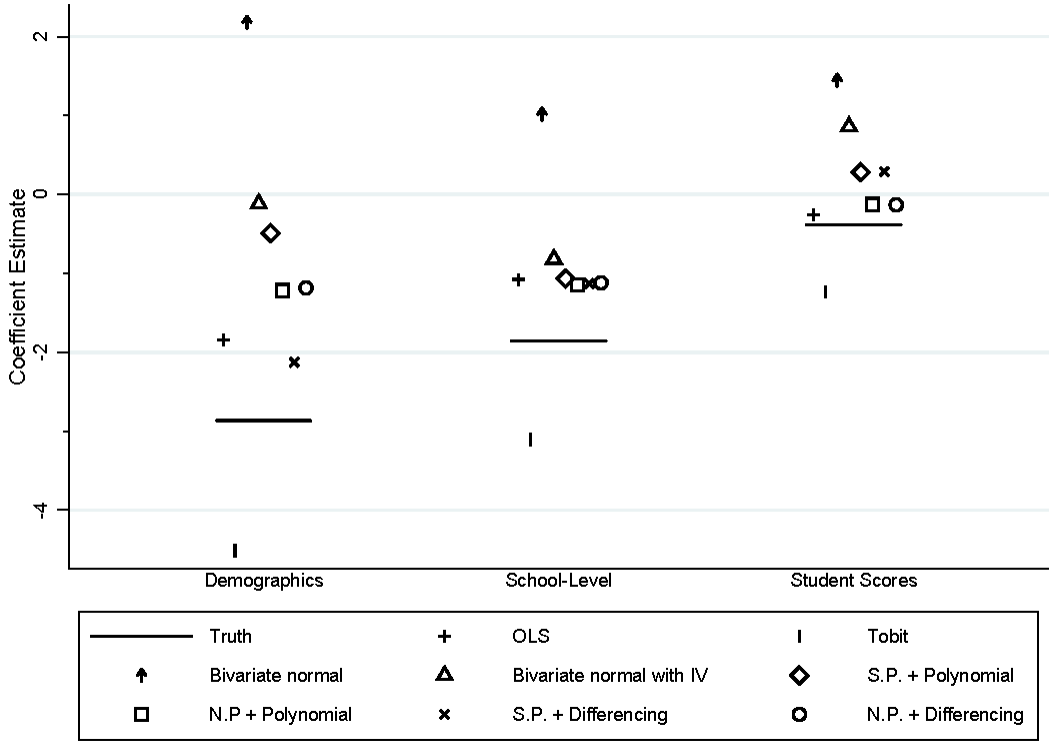
The figure clearly shows the three main patterns discussed above. First, OLS has lower MSB than most correction methods, including the semiparametric methods, for each of the covariate sets. Second, the semiparametric correction methods mostly have lower MSB than the parametric correction methods. Third, the MSB is lower for the intermediate and rich covariate sets than the basic covariate set for almost all correction methods.

4.2 Comparing Selection Corrections' Performance for Specific Coefficients

Thus far we focused on the mean squared bias (MSB) across all coefficient estimates in each selection correction model, excluding the intercept. MSB provides a useful summary measure of selection correction methods' performance. However, many researchers using selection correction methods are particularly interested in a subset of coefficient estimates. In this section, we examine correction performance focusing on estimates of two particularly policy-relevant coefficients: an indicator variable for free or reduced-price lunch receipt, which proxies for low-income status, and an indicator for Black student race. These parameters are of interest to researchers and policy-makers who care about income and race gaps in student achievement. White students are the reference group for the latter comparison, and the model also includes indicators for Hispanic students and students of other races (who collectively make up only about 6% of the sample).

We show estimates of the income gap from each selection correction method, along with the reference estimate from the complete post-reform data, in Figure III. This has the same structure as Figure II: estimated values on the vertical axis with different sets of covariates in the left, middle, and right blocks on the horizontal axis. We also report the point estimates, standard errors, and bias on each selection-corrected point estimate in Table 4. The estimated post-reform income gap is 2.87 ACT points with basic covariates, which shrinks to 1.86 when we add school- and district-level covariates and to 0.38 when we add student-level test scores.

Figure III. Coefficient on Free Lunch Receipt Indicator by Selection Correction and Covariate Set



Notes: Figure shows the estimated coefficient on an indicator for free or reduced-price lunch receipt for every covariate set, for every selection correction and for the reference model that uses complete post-reform data.

We see the same three patterns for estimates of the income gap as for estimates of MSB. First, OLS has smaller bias than most all other correction methods for the basic and intermediate covariate sets, and has the smallest bias for the rich covariate set. Second, most semiparametric methods have lower biases than most parametric methods for most of the covariate sets. For the intermediate and rich covariate sets, all semiparametric methods have lower biases than all parametric methods, although the differences are not always statistically significant. Third, most methods have lower biases with the intermediate than the basic covariate set, and most methods have even lower biases with the rich than intermediate covariate set. The third pattern is not mechanically explained by the smaller estimated income gap when more covariates are included in the model - it is possible to have large biases on estimates of even small coefficients.

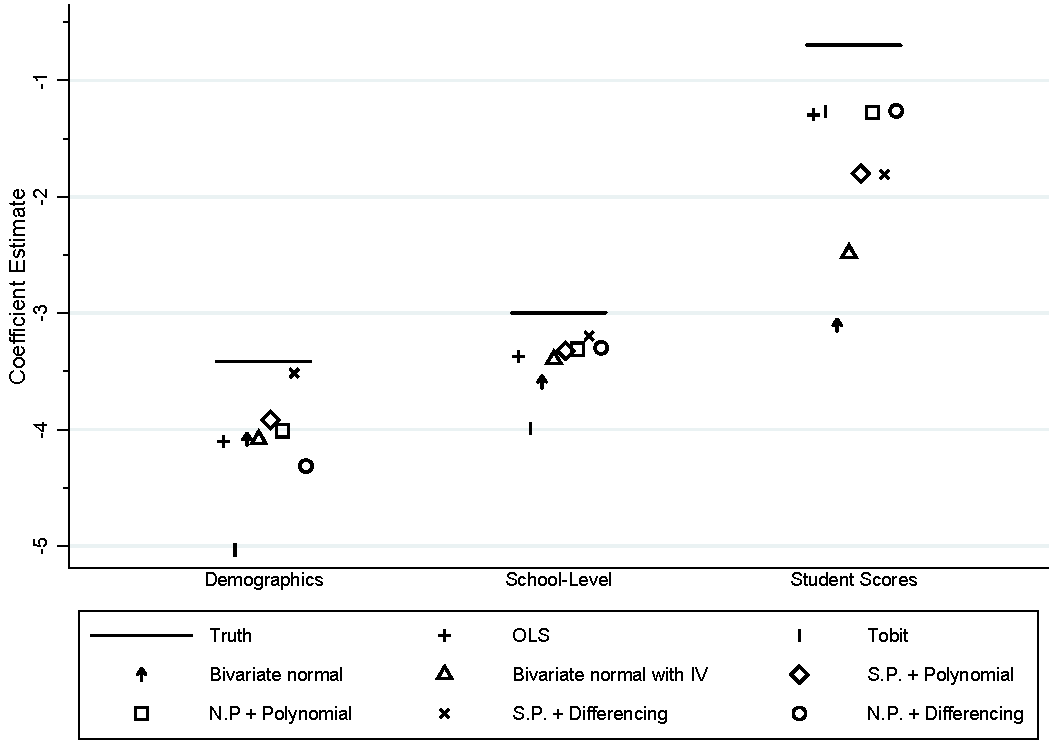
Figure IV and Table 5 show the same results for estimates of the coefficient on Black student race. White students, the reference group, have mean scores 3.41 points higher than Black students in the model with basic covariates, 3.00 in the model with intermediate covariates,

Table 4. Coefficient and Bias on Free Lunch Receipt Indicator by Correction Method and Covariate Set

	Post-Reform (Uncensored)		Pre-Reform, by Correction Method							
	OLS				Bivariate Normal		Polynomial		Differencing	
	Unweighted	Weighted	OLS	Tobit	No IV	With IV	SP	NP	SP	NP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student Demographics	-2.551	-2.866	-1.841	-4.515	2.180	-0.116	-0.492	-1.217	-2.127	-1.185
	(0.033)	(0.04)	(0.046)	(0.067)	(1.117)	(0.209)	(0.192)	(0.091)	(0.398)	(0.123)
			[1.025]	[-1.649]	[5.046]	[2.750]	[2.374]	[1.649]	[0.739]	[1.680]
Plus School/District-Level Covs	-1.703	-1.858	-1.078	-3.106	1.016	-0.819	-1.065	-1.148	-1.140	-1.120
	(0.033)	(0.039)	(0.048)	(0.071)	(0.282)	(0.154)	(0.076)	(0.056)	(0.100)	(0.070)
			[0.780]	[-1.248]	[2.874]	[1.039]	[0.792]	[0.710]	[0.717]	[0.738]
Plus Student Test Scores	-0.369	-0.383	-0.254	-1.239	1.445	0.860	0.283	-0.128	0.291	-0.131
	(0.019)	(0.022)	(0.033)	(0.054)	(0.083)	(0.081)	(0.050)	(0.036)	(0.057)	(0.043)
			[0.130]	[-0.856]	[1.828]	[1.244]	[0.666]	[0.256]	[0.675]	[0.253]

Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. Cells report coefficients on an indicator for free or reduced-price lunch receipt from a regression of ACT scores on covariates. Standard errors estimated using 500 bootstrap replications reported in parentheses. The bias on each coefficient, estimated as the difference between its value and the reference value in column value, reported in brackets.

Figure IV. Coefficient on Black Race Indicator by Selection Correction and Covariate Set



Notes: Figure shows the coefficient on an indicator for Black student race for every covariate set, for every selection correction and for the reference model that uses complete post-reform data.

and 0.70 in the model with rich covariates.

The pattern of results differs for biases on the race gap estimates as compared to the mean squared bias and biases on the income gap. The first pattern changes slightly: OLS now has higher biases than most semiparametric methods for all three covariate sets, although these differences are generally small and seldom statistically significant. The second pattern is largely unchanged: most semiparametric methods still have lower bias than most parametric methods, although the differences are not as large as for MSB or the income gap biases. The third pattern changes slightly: biases are still lower for the intermediate than basic covariate set, but they are now larger for the rich than intermediate covariate set for all methods except Tobit.

5 Explaining Results

The results in Section 4 show that the four semiparametric correction methods mostly have similar mean squared biases to OLS regressions that ignore selection, and that parametric

Table 5. Coefficient and Bias on Black Race Indicator by Correction Method and Covariate Set

	Post-Reform (Uncensored)		Pre-Reform, by Correction Method							
	OLS				Bivariate Normal		Polynomial		Differencing	
	Unweighted	Weighted	OLS	Tobit	No IV	With IV	SP	NP	SP	NP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student Demographics	-3.444	-3.414	-4.102	-5.036	-4.087	-4.085	-3.918	-4.010	-3.515	-4.314
	(0.034)	(0.046)	(0.044)	(0.08)	(0.098)	(0.058)	(0.06)	(0.045)	(0.195)	(0.075)
			[-0.688]	[-1.622]	[-0.673]	[-0.671]	[-0.504]	[-0.596]	[-0.101]	[-0.900]
Plus School/District-Level Covs	-2.870	-2.998	-3.371	-3.988	-3.593	-3.398	-3.324	-3.309	-3.196	-3.298
	(0.055)	(0.072)	(0.083)	(0.129)	(0.13)	(0.087)	(0.085)	(0.084)	(0.104)	-0.096
			[-0.373]	[-0.990]	[-0.595]	[-0.400]	[-0.326]	[-0.311]	[-0.199]	[-0.301]
Plus Student Test Scores	-0.648	-0.696	-1.295	-1.267	-3.106	-2.485	-1.801	-1.276	-1.807	-1.261
	(0.032)	(0.042)	(0.060)	(0.094)	(0.136)	(0.115)	(0.081)	(0.061)	(0.092)	(0.063)
			[-0.599]	[-0.571]	[-2.410]	[-1.789]	[-1.105]	[-0.580]	[-1.111]	[-0.565]

Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. Cells report coefficients on an indicator for black race from a regression of ACT scores on covariates. Standard errors estimated using 500 bootstrap replications reported in parentheses. The bias on each coefficient, estimated as the difference between its value and the reference value in column (2), reported in brackets.

correction methods mostly have higher mean squared biased. In this section we explore possible reasons for this combination of results.

First, our data do not satisfy the distributional assumptions of the parametric selection correction methods, which may help explain their high biases. Joint normality of the unobserved factors determining ACT scores and ACT-taking is a sufficient, though not necessary, condition for the bivariate normal models to address selection bias. The latent test score distribution in Figure I is not normal, and we verify this with parametric (skewness-kurtosis) and nonparametric (Kolmogorov-Smirnov) normality tests.[8] The latent distribution is also non-normal conditional on demographic characteristics (see Appendix Figure I) and the threshold censoring assumed by the Tobit model clearly does not hold. We also test the assumption that the unobserved factors that affect latent test scores are normally distributed: we regress the complete post-reform test scores on each of the three sets of covariates, 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).[9]

Second, there are some differences in the predicted probabilities of test-taking across different first stage models, which may explain part of the differences in performance between parametric and semiparametric methods. Table 6 reports percentiles of the distribution of predicted probabilities from each test-taking model and correlations between these predicted probabilities. Predictions from the probit first stages with and without instruments have correlations > 0.97 for all sets of covariates. They are slightly less correlated with predictions from the series logits (0.91-0.98). The predicted probabilities from the nonparametric matching are less strongly correlated with the predictions from the series logit models (0.89-0.93) and particularly the probit models (0.83-0.90). There are also some differences in the percentiles of the predicted distributions, particularly in the left tail.

Third, the relationship between the selection correction terms and covariates in the test score models differs substantially between the bivariate normal and polynomial correction methods, which helps to explain the higher bias of the parametric methods. The R^2 from regressing the

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

[9]This contrasts with the conclusions from Vella (1998), who finds that parametric and semiparametric selection models produce similar results in real data even when the assumptions of the parametric models fail. However, it is consistent with Goldberger (1983), Heckman et al. (2003), and Paarsch (1984), who show that some parametric models perform poorly in simulations when their assumptions are violated.

Table 6. Cross-Model Comparison of First Stage Predicted Probabilities by Covariate Set

	Basic Student Demographics					Plus School & District Covariates					Plus Prior Student Test Scores				
	OLS	Probit No IV	Probit With IV	Semi-Parametric	Non-Parametric	OLS	Probit No IV	Probit IV	Semi-Parametric	Non-Parametric	OLS	Probit No IV	Probit IV	Semi-Parametric	Non-Parametric
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Percentiles															
1%	0.386	0.380	0.344	0.315	0.298	0.217	0.199	0.197	0.127	0.185	0.041	0.051	0.051	0.049	0.159
5%	0.386	0.381	0.386	0.377	0.370	0.358	0.340	0.338	0.330	0.335	0.232	0.171	0.170	0.149	0.294
10%	0.475	0.478	0.441	0.441	0.430	0.421	0.411	0.410	0.403	0.412	0.329	0.271	0.269	0.236	0.373
25%	0.645	0.646	0.614	0.599	0.574	0.552	0.551	0.551	0.532	0.533	0.486	0.471	0.470	0.439	0.521
50%	0.645	0.646	0.665	0.670	0.670	0.656	0.665	0.668	0.663	0.665	0.647	0.684	0.684	0.691	0.683
75%	0.735	0.735	0.736	0.746	0.740	0.742	0.753	0.757	0.767	0.772	0.800	0.847	0.848	0.875	0.829
90%	0.735	0.735	0.754	0.761	0.780	0.829	0.828	0.829	0.851	0.867	0.942	0.938	0.939	0.954	0.924
95%	0.735	0.735	0.759	0.761	0.800	0.877	0.869	0.867	0.890	0.917	1.029	0.968	0.969	0.976	0.958
99%	0.836	0.822	0.818	0.808	0.850	0.973	0.917	0.919	0.939	0.965	1.194	0.993	0.993	0.994	0.986
Correlations															
OLS	1.000					1.000					1.000				
Probit, No IV	1.000	1.000				0.994	1.000				0.974	1.000			
Probit, With IV	0.975	0.975	1.000			0.988	0.994	1.000			0.971	0.997	1.000		
Series Logit	0.952	0.952	0.976	1.000		0.912	0.915	0.921	1.000		0.910	0.937	0.940	1.000	
Nonparametric	0.880	0.880	0.902	0.923	1.000	0.834	0.834	0.839	0.896	1.000	0.834	0.839	0.842	0.894	1.000
Fraction Correct Predictions		0.658	0.634	0.637	0.632		0.663	0.663	0.674	0.672		0.745	0.745	0.759	0.745

Notes: Table reports descriptive statistics and correlations of the first stage predicted probabilities across selection models and across covariate sets. The fraction correct predictions is the fraction of non-takers with predicted probabilities below 0.36 and takers with predicted probabilities above 0.36, as the test-taking rate is 0.64. For the basic covariate set, the bivariate normal model without an IV has a slightly higher correct prediction rate than the other models. This is sensitive to whether we assign values at the 34th percentile of the latent index as predicted 1s or predicted 0s. The other combinations of models and covariate sets do not have this sensitivity, because they all use continuous covariates and/or instruments, resulting in less coarse predicted values. Using the basic covariate set, no predicted values from OLS are above one and or below zero. Using the school demographics covariate set, 0.4% of predicted values from OLS are above one and 0.2% are below zero. Using the covariate set that includes student test scores, 6.4% of predicted values from OLS are above one and 0.7% are below zero.

inverse Mills ratio on the covariates used in the second stage model is 0.98-0.99 without instruments and 0.94-0.98 with instruments (ranges over the three different covariate sets). This illustrates that the nonlinearity of the inverse Mills ratio in the covariates generates almost no independent variation in the selection correction term conditional on the second stage covariates. The instruments generate some additional variation, but the bivariate normal correction terms are colinear enough with the second stage covariates to potentially cause problems. In contrast, the relationship between the polynomial selection correction terms and the second stage covariates are much weaker. The R^2 from regressing the second and third order terms in the polynomial corrections on second stage covariates range from 0.43 to 0.74 (over the three covariate sets and two different first stage estimators).

To explore the relative importance of the first and second stages, we estimate three models that mix together elements of different models. We use the series logit first stage to construct an inverse Mills ratio for the second stage, use the probit first stage with instruments to construct a polynomial selection correction for the second stage, and use the probit first stage with instruments to implement a differencing estimator in the second stage. The first two mixed approaches have on average almost half the mean squared bias of the bivariate model correction with instruments. This shows that the poor performance of the bivariate normal model can be alleviated by introducing more variation in the selection correction terms conditional on the second stage covariates, either through higher-order covariate terms in the first stage or through a less linear selection correction term than the inverse Mills ratio. Similar gains might be possible using parametric methods with different distributional assumptions that lead to less linear selection correction terms. In contrast, the third mixed approach (probit + differencing) performs no better than the bivariate normal model. This might occur because the probit first stage yields fewer unique values for the predicted probability of test-taking than the series logit or matching first stages, which leads to ties when implementing the differencing estimator.

The preceding points suggest that the semiparametric methods outperform the parametric methods because the latter rely on incorrect distributional assumptions and (potentially because of this) the selection correction terms have little variation conditional on the second stage covariates. But why do the semiparametric methods not systematically outperform OLS, which ignores the selection problem? To answer this, we examine two further features of our data.

First, there is a sample selection problem, so the relatively low bias of the OLS estimates is surprising. Figure I shows substantial differences between distributions of observed scores

for test-takers in the pre- and post-reform periods. Test-takers in the pre-reform period are clearly positively selected on latent scores. This pattern of positive selection also holds within race and income subgroups (Appendix Figure I). Furthermore, the selection correction terms in both the bivariate normal and polynomial selection correction models are large and statistically significant predictors of ACT scores (Appendix Tables 5 - 7).[10]

Second, the instruments have a strong but limited association with test-taking, which may limit the capacity of the semiparametric methods to correct for selection. The driving distance and weather instruments are negatively associated with test-taking: the point estimates are jointly statistically significant at conventional levels and moving from the 1st to the 99th percentile of the instrument distribution shifts the probability of taking the ACT by 10-13 percentage points (depending on the covariate set). The instruments are also not associated with scores on other tests taken by all students, supporting the exclusion restriction. However, the instruments do not shift the probability of test-taking from 0 to 100, so they do not satisfy "identification at infinity," as we discuss in Appendix B (Andrews and Schafgans, 1998; Chamberlain, 1986; Heckman, 1990). This means we can identify the slope coefficients in equation (1a) but cannot separately identify the intercept coefficient β_0 from the level of the selection correction term. We view this as a natural feature of semiparametric selection analysis in many settings, rather than a feature specific to our setting. The relationship between our instrument and participation measure is at least as strong as in classic education applications (Bulman, 2015; Card, 1995; Kane and Rouse, 1995). And non-identification of the intercept is not necessarily a problem for our analysis, which examines mean squared bias of the slope coefficients. However, we acknowledge that the relative performance of different selection models may differ when researchers have instruments that shift the probability of selection closer to 0 and 100.[11]

[10]The inverse Mills ratio term in the bivariate normal 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 covariates and instruments ($p = 0.069$ for the model with instruments and school- and district-level covariates, $p < 0.001$ for all other models). The coefficients are large: moving from the 1st to the 99th percentile of the predicted probability of test-taking shifts the test score by 5.3 points, averaging over the models. We also test if the coefficients on the polynomial correction terms in the polynomial model are jointly zero. We reject this hypothesis for all three combinations of covariates and both approaches to estimating the first stages ($p < 0.001$).

[11]We show in Appendix Figures IX and X that our main findings hold when we use only the weather instruments or only the distance instruments.

6 Conclusion

Sample selection arises when outcomes of interest are not observed for part of the population and the latent outcomes differ between 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 limited evidence on their relative performance outside of simulation studies. We use a Michigan policy that changed ACT-taking for 11^{th} graders from voluntary to mandatory to observe selected ACT scores for one cohort and complete scores for another cohort. We evaluate how well different selection corrections, applied to the selected data, can recover the coefficients of regression models estimated using the complete data.

We find that OLS, which ignores the selection problem, and several semiparametric methods perform similarly well. No one semiparametric method dominates the others or dominates OLS. Parametric corrections that rely on specific distributional and functional form assumptions perform worse than OLS and the semiparametric methods, with the bivariate normal correction without instruments performing particularly poorly. Mean squared bias is generally higher for models that use only a few discrete covariates than for models that use more covariates with less coarse distributions.

We examine selection correction methods' performance in a single data set, and thus evaluate their performance using only a single empirical example. These patterns may not generalize to other empirical examples. However, our findings may be of interest to those trying to use exam scores from selected samples of exam takers to infer population achievement. School district and state education administrators, as well as education researchers, often use college entrance exam scores to construct proxies for mean achievement for different groups of students. In our application, this exercise is subject to greater selection bias when few control variables are used and when using parametric selection correction methods.

Our main finding that semiparametric corrections tend to perform no better than simple OLS may be of interest to applied researchers using selection correction methods or adapting existing methods for new applications (e.g. Dahl 2002; Bonhomme et al. 2016). In our application, there is a sample selection problem and our instruments are comparable in strength to other widely-used instruments. This is a setting where we would expect semiparametric models to outperform OLS. However, the gains from using these more flexible methods are minimal.

Researchers using similar datasets who believe there is a sample selection problem should not necessarily conclude that semiparametric correction methods will fully solve the problem.

We focus on selection bias in the parameters of the conditional mean function $\mathbb{E}[ACT_i^*|X_i]$. But researchers may be interested in other features of the latent test score distribution. In an earlier version of this paper, we examine how well selection methods can recover the mean, subgroup means, and distribution of latent ACT scores (Garlick and Hyman, 2018). We find that performance does not differ substantially across selection correction methods (including methods that ignore selection) but improves substantially when we use more covariates with less coarse distributions.

Another strand of the selection correction literature studies conditional quantiles of the latent outcome distribution. In Appendix D, we compare the biases of uncorrected and selection-corrected quantile regression estimates, following the correction approach in Arellano and Bonhomme (2017). We find suggestive evidence that the selection-corrected quantile method has lower bias than the uncorrected method. But we do not examine all possible implementations of Arellano and Bonhomme's method. Another strand of the selection correction literature aims to derive bounds on possible values of the conditional mean or conditional quantile functions. 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). In our data, these bounds cover the reference estimates from the post-reform data, but they are also wide enough that some researchers will not view them as informative.[12] We view thorough analysis of these different approaches to selection correction using empirical examples as a topic for possible future work.

References

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

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

[12]For example, Manski's least restrictive bounding method assumes that all non-takers score either the maximum or minimum ACT score. This approach estimates a range of [-4.43, 3.10] for the income gap in ACT scores, with a width equal to 1.54 standard deviations of the complete ACT score distribution. Lee's more restrictive approach derives bounds for the difference in means between groups with higher and lower test-taking rates. For example, the income gap in the ACT-taking rate is 25.9 percentage points and Lee's method yields bounds of [-3.63, -0.28] for the income gap in ACT scores, with a width equal to 0.69 standard deviations of the complete ACT score distribution.

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, 847–862.

Angrist, J., P. Hull, P. Pathak, and C. Walters (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics*, 132, 871–919.

Angrist, J., P. Pathak, and C. Walters (2013): "Explaining Charter School Effectiveness." *American Economic Journal: Applied Economics*, 5, 1–27.

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

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

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

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, Toronto: University of Toronto Press.

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, 2367–2420.

Das, M., W. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

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

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

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

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

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. Newhouse, H. Allen, K. Baicker, and Oregon Health Study Group (2012): "The Oregon Health Insurance Experiment: Evidence From the First Year," *Quarterly Journal of Economics*, 127, 1057–1106.

Fujiwara, T., K. Meng, and T. Vogl (2016): "Habit Formation in Voting: Evidence from Rainy Elections," *American Economic Journal: Applied Economics*, 8, 160–188.

Garlick, R. and J. Hyman (2018): "Quasi-Experimental Evaluation of Alternative Sample Selection Corrections," Mimeo, Duke University.

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

Gronau, R. (1974): "Wage Comparisons – A Selectivity Bias," *Journal of Political Economy*, 82, 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, 179–201.

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

——— (1990): "Variation of Selection Bias," *American Economic Review*, 80, 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 R. Robb (1985a): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and S. Burton, Econometric Society Monograph Series.

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

Heckman, J. J. and R. Robb, Jr. (1985b): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of Econometrics*, 30, 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, 281–311.

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

Krishnaswamy, N. (2019): ""Missing and Fired: Worker Absence, Labor Regulation, and Firm Outcomes"," Mimeo, University of Southern California.

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.

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

Madestam, A., D. Shoag, S. Veuger, and D. Yanagizawa-Drott (2013): "Do Political Protests Matter? Evidence from the Tea Party Movement," *Quarterly Journal of Economics*, 128, 1633–1685.

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

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

Meyer, B. and N. Mittag (2019): "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net," *American Economic Journal: Applied Economics*, 11, 176–204.

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, 324–328.

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, 53–68.

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

Semykina, A. and J. Wooldridge (2013): "Estimation of Dynamic Panel Data Models with Sample Selection," *Journal of Applied Econometrics*, 28, 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, 24–36.

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

Online Appendices: "Quasi-Experimental Evaluation of Alternative Sample Selection Corrections: Online Appendices"

Robert Garlick and Joshua Hyman

A Data Construction and Additional Statistics

This appendix provides more information on how we construct the dataset and shows additional summary statistics.

Matching data sources: We matched data from the Michigan Department of Education (MDE) with four other data sources. First, using student name, date of birth, sex, race, and 11^{th} grade home zip code stored on a restricted access computer at the MDE, we match the student-level Michigan data to microdata from ACT Inc. and The College Board on every ACT-taker and SAT-taker in Michigan over the sample period. For the pre-reform cohorts, we use students' first ACT score, which is typically from 11^{th} grade, but in some cases is from 12^{th} grade. For students taking the SAT but not the ACT pre-reform, we convert their first SAT score into the ACT scale following published concordance tables.

Second, we acquired from ACT Inc. a list of all ACT test centers in Michigan over the sample period, including their addresses and open and close dates. Again using the restricted access data computer at MDE, we geocode student home addresses during 11^{th} grade and the addresses of these test centers to construct a student-level driving distance from 11^{th} grade home to the nearest ACT test center. When a student has multiple addresses during 11^{th} grade, we use the one with the shortest distance to a center. When 11^{th} grade home address is missing, we use home address during the surrounding grades. The $\approx$2% of students with a missing address during every high school grade are dropped from the pre- and post-reform samples. Appendix Table 1 shows detailed summary statistics for driving distance.

Third, we obtained historical weather data for Michigan during our sample period from the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI) *Storm Events Database*. These data, submitted monthly to the NCEI by the National Weather Service, record the time, location, and type (e.g., hail, winter storm, blizzard, thunderstorm wind) of all severe weather events in the U.S. We drop event types that we expect could have no impact on test-taking (e.g., heat, high surf, drought), and then merge the storm data with our historic data on the location of ACT test centers by county. We use

Appendix Table 1. Summary Statistics of Distance from Student Home to Nearest Test Center

	Overall			Urban		Rural	
	Total	Pre	Post	Pre	Post	Pre	Post
Mean	3.71	4.87	2.58	2.32	1.33	8.54	4.01
SD	3.89	4.67	2.47	1.79	0.90	5.90	3.29
Percentiles							
1st	0.2	0.3	0.2	0.3	0.2	0.4	0.2
5th	0.5	0.7	0.4	0.6	0.3	1.1	0.4
10th	0.7	1.0	0.6	0.7	0.4	1.8	0.7
25th	1.2	1.7	1.0	1.2	0.7	4.0	1.6
Median	2.4	3.1	1.8	1.9	1.1	7.5	3.3
75th	4.7	6.5	3.4	2.9	1.7	12.0	5.5
90th	8.6	11.5	5.7	4.2	2.4	16.6	8.1
95th	11.9	14.8	7.4	5.3	3.0	19.5	9.8
99th	18.7	21.1	11.2	9.7	4.6	26.7	15.1
Sample Size	197,014	97,108	99,906	20,434	20,859	25,194	25,856

Notes: The sample is as in Table 3. Distance, measured in miles, is the driving distance from the student's home address during 11th grade to the nearest ACT-test center. In the post-reform period, this is the driving distance from a student's home to his or her high school. If a student has multiple addresses during 11th grade, then the smallest distance is used. Urban and rural sample sizes do not sum to the total, because approximately half of students live in suburban areas and towns.

data on the timing of national ACT testing dates during our sample period (which took place only five times a year, once in October, December, February, April, and June) to include only severe weather events that occurred within the 24 hours prior to a test date. Over 90% of college entrance exams in our pre-reform period are taken between April of 11^{th} grade and February of 12^{th} grade, so we code a student as exposed to a severe weather event if they experience an event affecting an exam date in their county during that period.

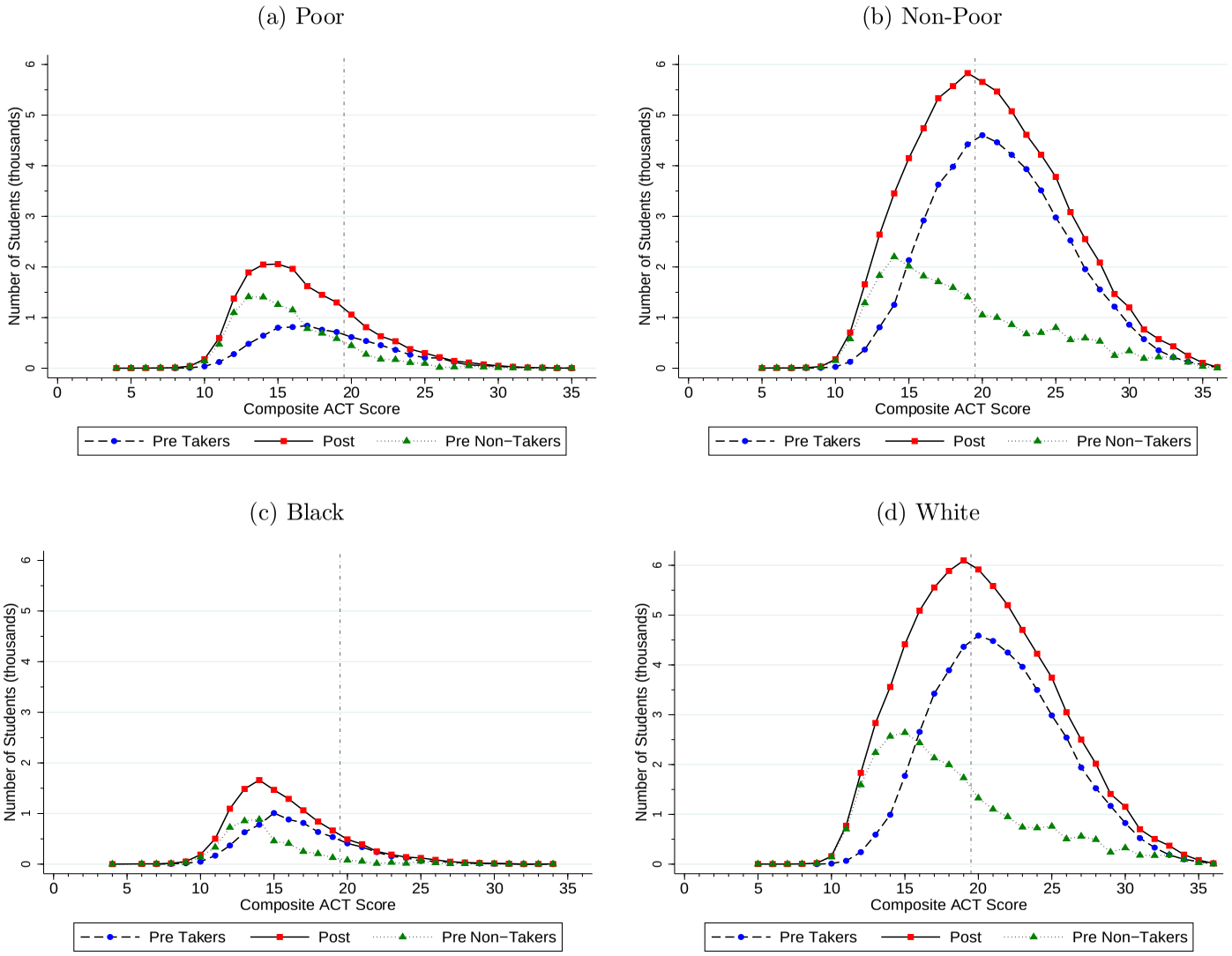
Finally, we matched unemployment rates at the city (when available) or county level from the Bureau of Labor Statistics onto the school-level data.

Test scores: For the pre-reform cohorts, we measure students' ACT scores using their first attempt. This is typically from 11^{th} grade, but in some cases is from 12^{th} grade. For students taking the SAT but not the ACT pre-reform, we convert their first SAT score into the ACT scale following published concordance tables. Appendix Table 2 shows detailed summary statistics for ACT scores. Appendix Figure I shows the distribution of observed pre- and post-reform test scores and the difference between these, interpreted as a measure of the latent scores of non-takers. Unlike Figure I in the main paper, this figure shows the distributions for subgroups based on race and free and reduced-price lunch (in)eligibility.

We construct student-level 8^{th} and 11^{th} grade test scores from in-school, state-wide assessments. For the 8^{th} grade test score, we use the average of a student's standardized math and English scores. For 11^{th} grade, we use standardized social studies scores because post-reform math and English scores are in part determined by a student's ACT score. If a student has missing test scores, we replace the scores with zeros and include indicator variables for missing test scores as covariates.

Sample restrictions: Our main analysis sample, which is conditional on students taking the state-wide 11^{th} grade test, excludes the small number of such students who do not complete high school or who take the special education version of the state-wide 11^{th} grade test. These students are not suited for our analysis because they are not required to take the ACT in either period. Our results are robust to including them. The 2006 cohort includes students in some schools where the mandatory ACT policy was piloted. When we analyze the 2006 cohort in Appendix C, we exclude these schools.

Appendix Figure I: Observed and Latent ACT Scores, By Subgroup



Notes: Figures show (1) the distribution of pre-reform ACT scores, 2) the distribution of post-reform ACT scores reweighted following DiNardo, Fortin, and Lemieux (1996) to resemble the pre-reform cohort, and (3) the difference between (1) and (2), which estimates the latent score distribution among non-takers in the pre-reform period. Weights are estimated separately for each subgroup.

Appendix Table 2. ACT Score Distributions Pre- and Post-Reform

	2005 Cohort		
	Takers	Non-Takers	2008 Cohort
	(1)	(2)	(3)
Moments			
Mean	20.85	17.65	19.73
Variance	4.54	5.11	4.98
Skewness	0.31	1.01	0.42
Kurtosis	2.72	3.56	2.65
Percentiles			
1st	12	10	11
5th	14	12	12
10th	15	12	14
25th	17	14	16
Median	21	16	19
75th	24	20	23
90th	27	25	27
95th	29	28	29
99th	32	33	32
Fraction Scoring>=20	0.588	0.285	0.482
K-S Test vs Column 1			
D-Stat		0.335	0.117
P-Value		0.000	0.000
Number of Students	62,186	33,475	95,661

Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. The reported number of students in the 2008 cohort is adjusted to match the size of the 2005 cohort and also includes only the 98.5% of the sample who take the ACT. Column (2) reports the distribution of latent ACT scores of students not taking the exam calculated using the methodology described in the text. The K-S test statistic and p-value are from a Kolmorogov-Smirnov test of the equality of the distributions.

B Selection Correction Models

This appendix elaborates on Section 2.2 of the main paper. We discuss each of the selection correction models in more detail, explaining the different assumptions under which they yield consistent estimators of β, and discuss implementation of the four semiparametric models. We summarize differences between these models' assumptions in Appendix Table 3. We do not evaluate imputation methods, bounding methods, or methods focused on identification using large support conditions on the outcome or covariates rather than parametric assumptions or instruments (D'Haultfoueille and Maurel, 2013; Lewbel, 2007). The large support conditions in the latter literature are unlikely to hold in our setting.

We estimate parameter variances for all models using a nonparametric bootstrap. The bootstrap replicates all estimation stages within each replication, including the first stage test-taking and second stage test score models. We use bootstrap rather than analytical standard errors because our focus on mean squared bias requires standard errors for nonlinear combinations of estimates across multiple regressions. We follow most applied researchers in using a bootstrap approach but acknowledge that our variance estimates should be interpreted with caution.

B.1 Single-Equation Corrections for Sample Selection Bias ("OLS" and "Tobit")

We begin with a simple single equation approach using ordinary least squares, which ignores sample selection. Specifically, we estimate the model

$$ACT_i = X_i\beta + \epsilon_i \tag{3}$$

for the test-takers. This is a special case of system (1) where u_i and ϵ_i are independent and the instruments Z_i do not influence latent test scores so the omitted variable in equation (2) is zero. In this case, the probability of taking the ACT score may depend on observed and unobserved characteristics, but these are independent of ϵ_i and so there is no sample selection problem. Differences between the observed and latent distributions occur only because the probability of test-taking and test scores jointly vary across observed characteristics. For example, students from low-income households have both lower rates of test-taking (in the pre-reform period) and lower test scores (in the post-reform period). The assumptions for this special case will be violated if test-taking decisions and latent test scores are jointly influenced by any unobserved characteristics, such as motivation.

Appendix Table 3. Comparison of Assumptions Made by Different Selection Correction Models

Selection Correction Model	Joint Distribution of Unobserved Scalar Characteristics Predicting Test-Taking and Test Scores, F(ε,u)	Instrumental Variable	Functional Form of Test-Taking Model	Functional Form of Selection Correction	Functional Form of Test Score Model Absent Selection
OLS	ε and u independent	Irrelevant	Irrelevant	Irrelevant	Linear in observed and unobserved predictors
Tobit	ε = u is univariate normal	Unnecessary	Probit	Irrelevant	
Bivariate Normal	F(ε,u) is bivariate normal	Unnecessary	Probit	Inverse Mills ratio	
Bivariate Normal with IV		Necessary			
Semiparametric Polynomial	No restriction on joint distribution	Necessary	Series logit ("semiparametric")	Polynomial approximation	
Nonparametric Polynomial			Matching ("nonparametric")		
Semiparametric Differencing			Series logit ("semiparametric")	Differenced out	
Nonparametric Differencing			Matching ("nonparametric")		

Notes: Table reports assumptions made by each of the eight selection correction models for individual data used in this paper. For all models, we assume that (1) all unobserved characteristics predicting test-taking and test scores can be summarized in two scalars, respectively denoted ε and u, and (2) the observed predictors of test scores are additively separable from the unobserved scalar predictor in the absence of selection. Note that the Heckman model is identified under weaker parametric assumptions than joint normality of ε and u, but we focus on this case for clarity.

We next estimate a single equation adjustment for sample selection bias adapted from Tobin (1958). This "Type 1 Tobit" adjustment assumes that ϵ_i is homoskedastic and normally distributed and that students take the ACT if and only if their latent scores exceed some threshold value $\overline{\overline{ACT}}$. Under these assumptions, we can assign the threshold score $\overline{\overline{ACT}}$ to all students who do not take the ACT, where $\overline{\overline{ACT}}$ is the lowest score obtained by any test-taker. In practice, researchers generally set $\overline{\overline{ACT}}$ higher than the minimum observed value and then assign the score $\overline{\overline{ACT}}$ to both students with missing scores and students with non-missing scores below $\overline{\overline{ACT}}$. This necessarily discards information for some test-takers, and discards more information as $\overline{\overline{ACT}}$ is set higher. Under these assumptions, the parameter vector equals the minimizer of the likelihood function

$$L\left(\beta, \sigma^2\right) = \prod_{i=1}^{n} \left(\frac{1}{\sigma}\phi\left(\frac{TAKE_i - X_i\beta}{\sigma}\right)\right)^{TAKE_i} \cdot \left(1 - \Phi\left(\frac{X_i\beta - \overline{\overline{ACT}}}{\sigma}\right)\right)^{1-TAKE_i} \tag{4}$$

where the first and second terms of the likelihood reflect the observed ACT scores and the probability of taking the ACT respectively. $\phi(.)$ and $\Phi(.)$ are the standard normal density and distribution functions respectively. Differences between the observed and latent distributions occur because no students with latent scores below $\overline{\overline{ACT}}$ take the test. This set of assumptions allows test-taking to depend on the unobserved characteristic ϵ_i but in a very restrictive way. These assumptions will be violated if students with low latent scores take the test and/or students with high latent scores do not take the test, perhaps due to heterogeneity in preferences for going to college. The assumptions will also be violated if ϵ_i is not homoskedastic and normally distributed, or if the threshold $\overline{\overline{ACT}}$ is incorrectly specified. We set $\overline{\overline{ACT}}$ equal to the 36^{th} percentile of the post-reform distribution of test scores, as the test-taking rate in the pre-reform period is 64%.

B.2 Parametric Multiple-Equation Corrections for Sample Selection Bias ("Bivariate Normal" and "Bivariate Normal with Instruments")

We estimate two variants of the bivariate normal selection model proposed by Gronau (1974) and Heckman (1974, 1976, 1979). Both consider the system

$$ACT_i = X_i\beta + \sigma_u\rho_{\epsilon,u}\lambda\left(Z_i\gamma\right) + \epsilon_i \text{ if } TAKE_i^* \geq 0 \tag{5a}$$

$$TAKE_i^* = X_i\delta + Z_i\gamma + u_i \tag{5b}$$

$$TAKE_i = \begin{cases} 1 & \text{if} \quad TAKE_i^* \geq 0 \\ 0 & \text{if} \quad TAKE_i^* < 0 \end{cases} \tag{5c}$$

where ϵ_i and u_i are jointly normally distributed and homoskedastic, and $\phi(.)$ and $\Phi(.)$ are the standard normal density and distribution functions respectively. Under the assumption of joint normality, the non-zero conditional mean error function $\mathbb{E}\left[ACT_i|X_i\right] = X_i\beta + \mathbb{E}\left[u_i > -X_i\delta - Z_i\gamma\right]$ is a linear function of the inverse Mills ratio. Hence, estimating a probit regression of $TAKE_i$ on (X_i, Z_i) and equation (5a) by ordinary least squares provides a consistent estimator of β. We estimate equation (5b) using only X_i as covariates ("bivariate normal") and also including a set of instruments Z_i that are excluded from equation (5a) and assumed not to affect test scores directly ("bivariate normal with instruments"). The former approach generally performs poorly in Monte Carlo simulations because the inverse Mills ratio is approximately linear for most of its support (Puhani, 2002).

This approach allows ACT-taking and ACT scores to depend jointly on both observed and unobserved characteristics. Unlike the Tobit model, the bivariate normal model allows the threshold score to vary with X_i, u_i, and potentially Z_i. This imposes few behavioral or economic assumptions but requires a strong statistical assumption on the joint distribution of ϵ_i and u_i. The approaches discussed in Appendix B.3 are all attempts to relax these distributional assumptions.[13]

[13]Several authors propose extensions of the bivariate normal selection model that yield consistent estimators under alternative parametric assumptions: uniform (Olsen, 1980) or Student-t (Lee, 1982, 1983) error distributions, or normal but heteroskedastic error distributions (Donald, 1995). Results for alternative parametric models, not reported in this version of the paper, are very similar to those from the bivariate normal model.

Appendix Table 4: First Stage Results

	Coef.	Standard Error	
		Robust	Clustered
	(1)	(2)	(3)
Student-Level			
Distance (Miles)	-0.006	0.001	0.001
Distance Squared (/ 10)	0.002	0.000	0.001
One Severe Weather Event	-0.019	0.004	0.016
Two Severe Weather Events	-0.079	0.007	0.018
Free Lunch	-0.111	0.003	0.011
Male	-0.068	0.003	0.003
Black	0.105	0.007	0.005
Hispanic	-0.005	0.008	0.014
Other Race	0.082	0.008	0.017
8th Grade Test Score	0.114	0.002	0.004
11th Grade Test Score	0.147	0.002	0.004
School-Level			
Average Class Size	0.000	0.000	0.000
Percent Free Lunch	-0.009	0.015	0.032
Percent Black	-0.004	0.026	0.045
Grade 11 Enrollment	0.000	0.000	0.000
Average 8th Grade Score	0.122	0.011	0.024
Average 11th Grade Score	0.022	0.007	0.011
District-Level			
Suburb	0.007	0.004	0.010
Town	0.016	0.006	0.020
Rural	0.027	0.006	0.017
Grade 11 Enrollment	0.000	0.000	0.000
Average Class Size	-0.005	0.001	0.002
Percent Free Lunch	-0.086	0.017	0.053
Percent Black	0.174	0.027	0.038
Student-Counselor Ratio	0.000	0.000	0.000
Local Unemployment Rate	-0.003	0.001	0.002

Notes: Table shows average marginal effects (column 1) from the first stage probit regression of a dummy for whether a student takes the ACT or SAT on the test access instruments, and student, school, and district demographics and test scores. We report heteroskedasticity-robust standard errors in column 2 and standard errors clustered at the county level in column 3.

B.3 Semiparametric Multiple-Equation Corrections for Sample Selection Bias ("Semiparametric + Polynomial," "Semiparametric + Differencing," "Nonparametric + Polynomial," and "Nonparametric + Differencing")

We now consider models of the form

$$ACT_i^* = X_i\beta + h\left(\hat{g}\left(X_i, Z_i\right)\right) + \epsilon_i \tag{6a}$$

$$TAKE_i^* = g\left(X_i, Z_i\right) + u_i \tag{6b}$$

$$TAKE_i = \begin{cases} 1 & \text{if} \quad TAKE_i^* \geq 0 \\ 0 & \text{if} \quad TAKE_i^* < 0 \end{cases} \tag{6c}$$

where $g(.,.)$ and $h(.)$ are potentially unknown functions, and we do not assume a specific distribution for ϵ_i or u_i. There are a wide range of semiparametric sample selection correction models (Pagan and Ullah, 1999). All use some "flexible" procedure to estimate the first stage model $Pr(TAKE_i = 1|X_i, Z_i)$ and to approximate the selection correction function $h(\hat{g}(X_i, Z_i))$. We consider two approaches to estimating the first stage and two approaches to dealing with the selection correction function.

Our first ACT-taking model is a series logit model, which we call the "semiparametric" first stage. We assume that we can approximate $g\left(X_i, Z_i\right)$ using polynomial expansions in X_i and Z_i, inside a logistic link function:

$$Pr\left(TAKE_i = 1\right) = L\left(\sum_{p=1}^{P}\left(\sum_{k=1}^{K}\theta_k X_{i,k}\right)^p + \sum_{q=1}^{Q}\left(\sum_{j=1}^{2}\psi_j Z_{i,j}\right)^q\right) \tag{7}$$

We observe multiple covariates $X_{i,1}, \ldots, X_{i,K}$ and two instruments, so we include polynomial terms in each element interactions between the elements. Higher values of P and Q achieve a closer fit to the data and hence reduce the bias of the coefficient estimator, but at the cost of higher variance.

We choose the orders P and Q of the two series to minimize the mean squared prediction error of the logistic regression using 10-fold repeated cross-validation.[14] We first randomly sort the data and estimate a logit model with a linear specification inside the logit ($P = Q = 1$) on deciles 2-10 of the sample and predict the outcomes for decile 1. We then estimate the model

[14]There does not appear to be a consensus on how to choose the order of series estimators in nonlinear regression models, even though series logit models are used in important econometric theory papers such as Hirano et al. (2003). We use repeated 10-fold cross-validation because leave-one-out cross-validation with a nonlinear model is computationally burdensome in large datasets like ours.

for deciles 1 and 3-10 and predict the outcomes for decile 2 and repeat this process to obtain predictions for all deciles. We calculate the mean squared difference between the observed binary values of $TAKE_i$ and the predicted values. We then resort the data and repeat this process 10 times, averaging the mean squared prediction error over repetitions. This repetition reduces the sensitivity of the prediction error to the initial ordering of the data and performs well in simulations (Borra and Di Ciaccio, 2010). We repeat this process for different values of P and Q and select the pairs of values that minimize the mean squared prediction error. We consider values of $P \in \{1, 2, 3\}$ and $Q \in \{1, \ldots, 10\}$, as higher values of P generate too many interaction terms to estimate the logit without dimension reduction techniques.

This cross-validation algorithm selects a second-order polynomial in the covariates for all three sets of covariates. This polynomial contains linear terms in all covariates, quadratic terms in all continuous variables, and all pairwise interaction terms.[15] Some pairwise interaction terms are omitted because they are mutually exclusive (e.g. Black and Hispanic). The cross-validation algorithm selects sixth-, eighth-, and third- order polynomials in the instrument when using respectively the basic, school/district, and student test score sets of covariates.

This semiparametric model therefore differs from the probit model used in the bivariate normal selection correction in two ways: the semiparametric model includes quadratic and interaction terms in the covariates and covariates, and uses a logit instead of a probit link function. Nonetheless, we see in Table 6 that the predicted probabilities of ACT-taking are similar, with correlations of 0.91 - 0.98.

Our second ACT-taking model uses a weighted K-nearest neighbor matching approach, which we call the "nonparametric" first stage. We directly estimate the conditional expectation $\mathbb{E}[X_i, Z_i] = g(X_i, Z_i)$ rather than approximating it with a regression model. We start by calculating the Mahalanobis distance between every pair of observations i and j: $D_{i,j} = \sqrt{(W_i - W_j)(V_W)^{-1}(W_i - W_j)'}$, where $W_i = (X_i, Z_i)$. Mahalanobis distance generalizes Euclidean distance by weighting the differences between the elements of the vectors W_i and W_j by the inverse of the sample covariance matrix V_W. This takes into account the different variances of different covariates/instruments and the covariances between covariates/instruments. We then identify the K nearest neighbors of each observation with respect

[15]The series model includes the interaction and polynomial terms in the ACT-taking model but not in the ACT score model. This effectively treats them as instruments for ACT-taking, though we do not claim they are excludable from the ACT score model. Our results are robust to including these terms in the ACT score model as well.

to the Mahalanobis distance and calculate the weighted average outcome amongst these K observations: $\hat{TAKE}_i = \sum_{k=1}^{K} \omega_{i,k} TAKE_k$. The weighting function $\omega_{i,k} = \frac{1}{1+d_{i,k}} / \sum_{k=1}^{K} \frac{1}{1+d_{i,k}}$ assigns more weight to observations with a lower Mahalanobis distance to i.[16] This estimator directly estimates the conditional mean $\mathbb{E}[W_i = w]$ at each value w without making assumptions about the function $g(.)$. We report results in this paper using $K = 100$, but we find similar results with $K = 10$ and $K = 1000$. Increasing the value of K past 10 has little effect on results because the estimator assigns very low weight $\omega_{i,k}$ to high values of k. This also smoothes the estimator relative to unweighted K-nearest neighbor matching with low K, making the bootstrap more appropriate (Abadie and Imbens, 2008). Code for implementing this estimator is available on the authors' websites.

Our first selection-corrected ACT score model approximates $h(.)$ using a series model in $\hat{TAKE}_i$, the predicted probability of test-taking (Heckman and Robb, 1985a; Newey, 2009).[17] We call this the "polynomial" second stage. We select the order of the series using leave-one-out cross-validation. We then estimate equation (6a) including a polynomial with the selected order as a control. This approach yields a consistent estimator of β when the selection correction term is a sufficiently smooth function of the predicted probabilities of test-taking. The cross-validation algorithm selects fourth, ninth, and fourth order polynomials for the selection term when we use a semiparametric first stage with respectively basic, school/district, and student test score sets of covariates. The cross-validation algorithm selects third, fourth, and fourth order polynomials for the selection term when we use a nonparametric first stage with respectively basic, school/district, and student test score sets of covariates.

Second, we remove $h(.)$ from equation (6a) using a differencing approach, which we call the "differencing" second stage (Ahn and Powell, 1993; Powell, 1987). We calculate $dACT_i = ACT_i - \frac{1}{N-1} \sum_{j \neq i} w(i,j) ACT_j$ and $dX_i = X_i - \frac{1}{N-1} \sum_{j \neq i} w(i,j) X_j$, where $w(i,j)$ is a kernel or weighting function that is decreasing in the difference between i and j's predicted probability of ACT-taking. For appropriate choices of the weighting function, $dh_i = h_i - \frac{1}{N-1} \sum_{j \neq i} w(i,j) h_j \approx$

[16] We use $\frac{1}{1+d_{i,k}}$ in the weighting function rather than $\frac{1}{d_{i,k}}$ to avoid zero-valued denominators for pairs of observations with $d_{i,k} = 0$.

[17] Newey (2009) proposes using polynomials in either the predicted probability $TAKE_i$ or the latent index $TAKE_i^*$. Our nonparametric matching estimator generates only predicted probabilities of test-taking so we use this in the ACT-taking model. Our series logit estimator generates both predicted index values and predicted probabilities. We report results in this paper using predicted index values, after censoring the top and bottom percentiles. Results are very similar using predicted probabilities. Note that concerns about "forbidden regression" are not necessarily applicable here, as the series is simply an approximating function and not an exact replacement for the selection bias term $\mathbb{E}[ACT_i | X_i] = X_i\beta + \mathbb{E}[u_i > g(X_i, Z_i)]$.

0. Hence we can rewrite equation (6a) as

$$dACT_i = dX_i\beta + d\epsilon_i \tag{8}$$

and estimate this using least squares. Intuitively, this approach avoids the need to approximate the selection correction term and instead differences it out of the test score model. This approach again yields a consistent estimator of β when the selection correction term is a sufficiently smooth function of the predicted probability of test-taking, so that $h_i \approx h_j$ when i and j have sufficiently similar predicted probabilities of ACT-taking. In practice, we sort the data by the predicted probability of test-taking and use a weight function that equals $1/\left(1 + |\hat{p_i} - \hat{p_j}|\right)$ for $0 < |i-j| < 5$ and zero otherwise. We then estimate the differenced equation using weighted least squares with weight $1/\sum_{i-j=-4}^{4} |\hat{p}_i - \hat{p}_j)|$. These weights mean that observations that have close matches on the predicted probability of ACT-taking influence the regression coefficients more than observations without close matches, as Ahn and Powell (1993) recommend. We obtain similar results (not reported in this draft) using a smaller number of matches in the differencing operation, taking an unweighted average in the differencing operation, and estimating the differenced equation without weights.[18]

Both the polynomial and differencing approaches to the ACT score model yield consistent estimators of β without making distributional assumptions on the unobserved determinants of test-taking or test scores, or functional form assumptions for the probability of test-taking or the selection correction term. However, this flexibility does have several costs. First, the identification proofs underlying both approaches assume that there is at least one excluded instrument: some observed variable Z_i affects the probability of test-taking but does not directly affect test scores. Intuitively, the coefficient vector β and the selection term in (6a) are separately identified only if there is additional information in the selection correction term (from an exclusion restriction) or by a nonlinear functional form of the selection correction term. The exclusion restriction is sufficient for identification of the slope coefficients in β but not the intercept, β_0. β_0 is identified when Z_i shifts the probability of test-taking from 0 to 1 as Z_i moves from its maximum to minimum value (or vice versa). This "identification at infinity"

[18]The asymptotic results in Ahn and Powell (1993) and Powell (1987) assume that this kernel function is continuously differentiable, which is not true of the weighted K-nearest neighbor kernels we consider. In simulations on a dataset with moments matched to our data the results are very robust to choices of different kernels. The asymptotic results also assume that the first stage model is undersmoothed, a topic that we address in Appendix C.

argument requires an unusually strong excluded instrument (Andrews and Schafgans, 1998; Chamberlain, 1986; Heckman, 1990). We exclude both driving distance from the student's home to the nearest ACT center and extreme weather events just before ACT testing dates from the outcome equation. Both driving distance and severe weather events are statistically significantly associated with lower ACT-taking. The negative relationships grow stronger as we control for student demographics, school- and district-level characteristics, and student scores on other tests. The probability of ACT-taking drops by 10-13 percentage points (depending on the covariate set) with a move from the 1^{st} to the 99^{th} percentile of the instruments. Over half of this shift is due to exposure to two severe weather events. This does not satisfy the identification at infinity argument, like most excluded instruments in the empirical literature, (Bulman, 2015; Card, 1995; Kane and Rouse, 1995). However, our main object of interest, the mean squared bias of the estimated slope coefficients, does not require identification at infinity.

Second, the semiparametric models yield consistent estimators only with appropriate choices of the tuning parameters: respectively the order of the polynomial and the weighting used in differencing. The parameter estimates may in principle be very sensitive to the choice of these parameters. In our application, results are robust to alternative polynomial orders and weighting functions. Third, some semiparametric and nonparametric sample selection correction models converge at slower rates than parametric models, particularly when the number of covariates is large. This means that the rate at which the estimators approach the true parameters as the sample size grows is slower, potentially generating estimates far from the truth with even moderate sample sizes. Ahn and Powell (1993) and Newey (2009) establish sufficient conditions for the estimators of the slope parameters in β to converge at parametric rates.

Both the semiparametric and parametric models assume that the unobserved determinants of test scores ϵ_i and test-taking u_i are homoskedastic conditional on the covariates. There exist parametric and semiparametric sample selection models that relax this assumption, which we do not evaluate (Donald, 1995; Chen and Khan, 2003).

C Robustness Checks and Extensions

In this section, we replicate our main findings under different conditions to assess their robustness and shed more light on the different performance of different estimators. In most cases, we show how the mean squared biases change under different conditions, presenting results in the same format as the main results in Figure II.

Coefficient estimates from main specifications: We first show the full set of estimated coefficients used to construct the mean squared bias estimates reported in the paper, included the race and free lunch coefficients already reported in the paper. This gives readers a more complete picture of our results. In column 1 of Appendix Tables 5 - 7, we show the parameter estimates from regressing post-reform ACT scores on each of the three vectors of covariates. In column 2, we show estimates from the same regression using the same data, reweighted to give it the same distribution of observed covariates as the pre-reform data. These are our preferred reference values that we use to estimate biases. In columns 3 to 10 we report the parameter estimates from regressing pre-reform ACT scores on each of the three vectors of covariates using our eight different selection correction models. At the bottom of each table, for each selection correction model, we report both the mean squared bias over all coefficients excluding the intercept and the share of coefficient estimates with different signs to the reference estimates. Like the mean squared biases, the latter statistic is not systematically lower for the semiparametric methods than OLS, and is highest for the bivariate normal model without instruments.

Mean squared bias with standardized covariates: In Section 4, we compare mean squared biases across covariate sets. This comparison is difficult to interpret, as the covariates in different sets have different scales. We show in Appendix Figure II the main mean squared bias results after standardizing all covariates to have a mean of zero and standard deviation of one. The basic covariate set, using only a few discrete covariates, continues to have much higher mean squared bias than the other two covariate sets, which include more covariates with less coarse distributions.

Mean squared bias with unweighted reference estimates: In our primary analyses, we use the coefficient estimates from the weighted post-reform models as the benchmark coefficient estimates. Some readers may prefer to see the results using the unweighted post-reform model as the benchmark. We show in Appendix Figures III, IV, and V, respectively the main mean

Appendix Table 5. The Relationship Between ACT Scores and Student Demographics

	Post-Reform (Uncensored)		Pre-Reform, by Correction Method							
	OLS				Bivariate Normal		Polynomial		Differencing	
	Unweighted	Weighted	OLS	Tobit	No IV	With IV	SP	NP	SP	NP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student Demographics										
Free Lunch	-2.551	-2.866	-1.841	-4.515	2.180	-0.116	-0.492	-1.216	-2.127	-1.185
	(0.033)	(0.040)	(0.046)	(0.067)	(1.117)	(0.209)	(0.192)	(0.091)	(0.398)	(0.123)
Male	-0.252	-0.298	0.130	-0.675	1.710	0.804	0.745	0.375	-0.025	0.288
	(0.028)	(0.034)	(0.033)	(0.046)	(0.432)	(0.088)	(0.072)	(0.042)	(0.169)	(0.056)
Black	-3.444	-3.414	-4.102	-5.036	-4.087	-4.085	-3.918	-4.010	-3.515	-4.314
	(0.034)	(0.046)	(0.044)	(0.080)	(0.098)	(0.058)	(0.060)	(0.045)	(0.195)	(0.075)
Hispanic	-2.113	-1.967	-1.818	-2.684	-0.443	-1.214	-1.294	-1.547	-1.450	-1.592
	(0.077)	(0.105)	(0.116)	(0.162)	(0.398)	(0.153)	(0.137)	(0.122)	(0.244)	(0.156)
Other Race	0.997	1.032	0.616	1.520	-1.295	-0.204	-0.511	0.390	0.204	-0.374
	(0.097)	(0.109)	(0.100)	(0.116)	(0.597)	(0.166)	(0.248)	(0.103)	(0.341)	(0.176)
Summary Measures										
P: Selection Correction Terms Jointly Zero					<0.001	<0.001	<0.001	<0.001		
% with Incorrect Signs			0.20	0.00	0.60	0.40	0.20	0.20	0.00	0.40
Mean Squared Bias			0.380	1.249	7.537	2.264	1.962	0.823	0.317	1.219
Sample Size	98,417	98,417	62,186	62,186	62,186	62,186	62,186	62,186	62,186	62,186

Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. Each column is from a separate regression. Standard errors estimated using 500 bootstrap replications reported in parentheses.

Appendix Table 6. The Relationship Between ACT Scores and Student Demographics, and School and District Characteristics

	Post-Reform (Uncensored)		Pre-Reform, by Correction Method							
	OLS				Bivariate Normal		Polynomial		Differencing	
	Unweighted	Weighted	OLS	Tobit	No IV	With IV	SP	NP	SP	NP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student Demographics										
Free Lunch	-1.703	-1.858	-1.078	-3.106	1.016	-0.819	-1.065	-1.148	-1.140	-1.120
	(0.033)	(0.039)	(0.048)	(0.071)	(0.282)	(0.154)	(0.076)	(0.056)	(0.100)	(0.070)
Male	-0.253	-0.288	0.058	-0.697	1.180	0.197	0.125	0.087	0.142	0.074
	(0.028)	(0.034)	(0.034)	(0.047)	(0.153)	(0.087)	(0.043)	(0.037)	(0.061)	(0.045)
Black	-2.870	-2.998	-3.370	-3.988	-3.592	-3.398	-3.324	-3.309	-3.197	-3.298
	(0.055)	(0.072)	(0.083)	(0.129)	(0.130)	(0.087)	(0.085)	(0.084)	(0.102)	(0.096)
Hispanic	-1.783	-1.781	-1.566	-2.185	-0.876	-1.480	-1.536	-1.521	-1.490	-1.678
	(0.074)	(0.094)	(0.110)	(0.145)	(0.180)	(0.125)	(0.111)	(0.111)	(0.139)	(0.130)
Other Race	0.506	0.505	0.157	0.675	-0.844	0.034	-0.049	0.044	-0.295	-0.020
	(0.092)	(0.106)	(0.101)	(0.112)	(0.221)	(0.126)	(0.114)	(0.101)	(0.142)	(0.107)
School Characteristics										
Pupil Teacher Ratio	0.012	0.001	-0.002	-0.010	0.002	-0.002	-0.002	-0.002	-0.001	-0.003
	(0.004)	(0.005)	(0.001)	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Fraction Free Lunch	0.888	0.636	-0.582	-0.797	-0.727	-0.607	-0.524	-0.572	-0.598	-0.520
	(0.199)	(0.247)	(0.178)	(0.269)	(0.275)	(0.179)	(0.189)	(0.178)	(0.242)	(0.206)
Fraction Black	1.808	1.712	1.017	1.454	-0.140	0.872	0.883	0.923	0.657	0.816
	(0.215)	(0.296)	(0.301)	(0.576)	(0.465)	(0.311)	(0.306)	(0.300)	(0.453)	(0.378)
Number of 11th Graders	0.000	0.000	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.002
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average 8th Grade Score	1.949	1.938	2.338	3.983	-0.187	2.020	1.909	2.033	1.947	1.979
	(0.088)	(0.102)	(0.092)	(0.155)	(0.360)	(0.201)	(0.118)	(0.099)	(0.159)	(0.123)
Average 11th Grade Score	2.592	2.741	1.224	2.445	-0.624	0.996	1.079	1.144	0.974	1.110
	(0.091)	(0.105)	(0.073)	(0.107)	(0.272)	(0.148)	(0.083)	(0.075)	(0.123)	(0.093)
District Characteristics										
Pupil Teacher Ratio	-0.027	-0.066	-0.020	-0.071	0.052	-0.010	0.002	-0.003	0.008	-0.008
	(0.007)	(0.010)	(0.009)	(0.013)	(0.017)	(0.010)	(0.010)	(0.009)	(0.014)	(0.012)
Fraction Free Lunch	-0.568	-0.554	0.300	0.053	0.906	0.383	0.271	0.346	0.694	0.388
	(0.188)	(0.229)	(0.206)	(0.287)	(0.327)	(0.211)	(0.232)	(0.207)	(0.301)	(0.250)
Fraction Black	0.900	1.510	0.864	2.199	-1.238	0.603	0.597	0.630	0.291	0.653
	(0.225)	(0.312)	(0.321)	(0.588)	(0.564)	(0.360)	(0.329)	(0.321)	(0.496)	(0.409)
Number of 11th Graders	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Suburb	-0.279	-0.169	-0.418	-0.306	-0.488	-0.429	-0.428	-0.419	-0.387	-0.314
	(0.043)	(0.058)	(0.053)	(0.071)	(0.079)	(0.054)	(0.055)	(0.053)	(0.088)	(0.069)
Town	-0.268	-0.177	0.023	0.251	-0.188	-0.007	0.088	0.070	0.077	0.134
	(0.064)	(0.076)	(0.074)	(0.098)	(0.114)	(0.077)	(0.076)	(0.074)	(0.107)	(0.093)
Rural	-0.251	-0.210	-0.201	0.103	-0.498	-0.239	-0.176	-0.160	-0.181	-0.049
	(0.055)	(0.069)	(0.065)	(0.087)	(0.104)	(0.071)	(0.068)	(0.065)	(0.097)	(0.081)
Pupil / Guidance Counselor Ratio	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Local Unemployment Rate	-0.020	-0.009	-0.032	-0.062	0.006	-0.028	-0.036	-0.032	-0.030	-0.028
	(0.007)	(0.009)	(0.010)	(0.014)	(0.016)	(0.010)	(0.010)	(0.010)	(0.014)	(0.012)
Summary Measures										
P: Selection Correction Terms Jointly Zero					<0.001	0.069	<0.001	<0.001		
% with Incorrect Signs			0.25	0.25	0.55	0.20	0.35	0.25	0.35	0.30
Mean Squared Bias			0.336	0.522	2.221	0.445	0.386	0.369	0.519	0.381
Sample Size	98,417	98,417	62,186	62,186	62,186	62,186	62,186	62,186	62,186	62,186

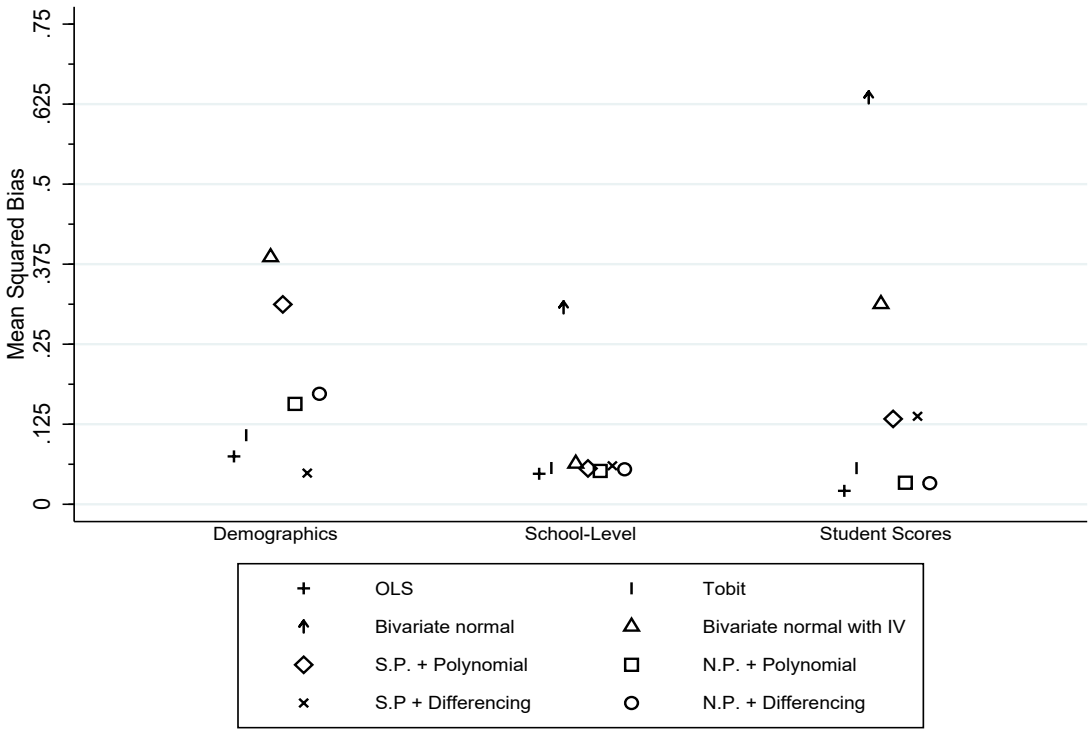
Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. Each column is from a separate regression. Standard errors estimated using 500 bootstrap replications reported in parentheses.

Appendix Table 7. The Relationship Between ACT Scores and Student Demographics and Test Scores and School and District Characteristics

	Post-Reform (Uncensored)		Pre-Reform, by Correction Method							
	OLS				Bivariate Normal		Polynomial		Differencing	
	Unweighted	Weighted	OLS	Tobit	No IV	With IV	SP	NP	SP	NP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student Demographics and Scores										
Free Lunch	-0.369	-0.383	-0.254	-1.239	1.444	0.860	0.283	-0.128	0.291	-0.131
	(0.019)	(0.022)	(0.033)	(0.054)	(0.083)	(0.081)	(0.050)	(0.036)	(0.057)	(0.043)
Male	-0.490	-0.505	-0.027	-0.473	1.091	0.709	0.402	0.094	0.426	0.084
	(0.017)	(0.020)	(0.024)	(0.032)	(0.062)	(0.059)	(0.038)	(0.025)	(0.039)	(0.026)
Black	-0.648	-0.696	-1.295	-1.267	-3.106	-2.485	-1.801	-1.276	-1.807	-1.261
	(0.032)	(0.042)	(0.060)	(0.094)	(0.136)	(0.115)	(0.081)	(0.061)	(0.092)	(0.063)
Hispanic	-0.607	-0.589	-0.727	-0.864	-0.753	-0.737	-0.741	-0.539	-0.773	-0.531
	(0.042)	(0.050)	(0.086)	(0.115)	(0.154)	(0.118)	(0.104)	(0.086)	(0.112)	(0.096)
Other Race	0.383	0.394	0.209	0.440	-1.384	-0.828	-0.285	0.079	-0.266	0.084
	(0.052)	(0.059)	(0.064)	(0.076)	(0.167)	(0.130)	(0.100)	(0.066)	(0.108)	(0.068)
Grade 8 Score	1.592	1.639	1.833	2.601	-0.135	0.537	1.069	1.683	1.062	1.680
	(0.021)	(0.029)	(0.021)	(0.028)	(0.067)	(0.083)	(0.044)	(0.022)	(0.048)	(0.025)
Grade 11 Score	3.036	3.048	2.616	3.608	0.109	0.965	1.692	2.419	1.688	2.436
	(0.015)	(0.019)	(0.018)	(0.026)	(0.09)	(0.107)	(0.054)	(0.022)	(0.054)	(0.022)
School Characteristics										
Pupil Teacher Ratio	0.005	-0.006	-0.003	-0.010	0.002	0.001	-0.001	-0.002	-0.001	-0.001
	(0.003)	(0.004)	(0.001)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
Fraction Free Lunch	-0.150	-0.536	-0.449	-1.073	-0.540	-0.561	-0.471	-0.385	-0.433	-0.342
	(0.147)	(0.187)	(0.122)	(0.195)	(0.279)	(0.203)	(0.170)	(0.123)	(0.190)	(0.144)
Fraction Black	-0.008	-0.253	-0.273	-0.809	-0.442	-0.406	-0.147	-0.474	-0.041	-0.371
	(0.142)	(0.187)	(0.250)	(0.397)	(0.467)	(0.357)	(0.305)	(0.244)	(0.332)	(0.278)
Number of 11th Graders	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.001	0.000	0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average 8th Grade Score	0.943	0.907	1.085	1.773	-1.248	-0.494	0.042	0.613	-0.005	0.564
	(0.052)	(0.065)	(0.074)	(0.108)	(0.218)	(0.175)	(0.108)	(0.075)	(0.111)	(0.076)
Average 11th Grade Score	-0.336	-0.231	-0.206	-0.164	-0.525	-0.401	-0.204	-0.265	-0.177	-0.229
	(0.055)	(0.066)	(0.057)	(0.077)	(0.137)	(0.098)	(0.076)	(0.057)	(0.083)	(0.061)
District Characteristics										
Pupil Teacher Ratio	-0.026	-0.044	-0.039	-0.063	0.061	0.031	0.022	-0.015	0.021	-0.011
	(0.005)	(0.007)	(0.007)	(0.010)	(0.015)	(0.011)	(0.011)	(0.007)	(0.011)	(0.008)
Fraction Free Lunch	-0.549	-0.272	-0.758	-0.681	0.611	0.197	0.094	-0.358	0.020	-0.392
	(0.127)	(0.160)	(0.146)	(0.210)	(0.328)	(0.246)	(0.213)	(0.147)	(0.235)	(0.160)
Fraction Black	0.724	1.150	1.260	2.183	-1.737	-0.713	-0.121	0.817	-0.286	0.724
	(0.146)	(0.191)	(0.263)	(0.403)	(0.515)	(0.403)	(0.331)	(0.256)	(0.369)	(0.304)
Number of 11th Graders	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Suburb	-0.180	-0.165	-0.356	-0.285	-0.394	-0.398	-0.367	-0.349	-0.366	-0.355
	(0.027)	(0.032)	(0.036)	(0.049)	(0.083)	(0.060)	(0.052)	(0.036)	(0.060)	(0.042)
Town	-0.155	-0.174	-0.072	0.067	-0.339	-0.278	-0.178	-0.084	-0.181	-0.067
	(0.037)	(0.043)	(0.052)	(0.069)	(0.116)	(0.085)	(0.071)	(0.052)	(0.081)	(0.060)
Rural	-0.093	-0.121	-0.224	0.038	-0.606	-0.483	-0.379	-0.206	-0.387	-0.211
	(0.032)	(0.039)	(0.046)	(0.063)	(0.103)	(0.077)	(0.064)	(0.046)	(0.072)	(0.052)
Pupil / Guidance Counselor Ratio	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Local Unemployment Rate	-0.014	-0.008	-0.039	-0.061	0.023	-0.002	-0.017	-0.029	-0.015	-0.030
	(0.004)	(0.005)	(0.007)	(0.009)	(0.015)	(0.011)	(0.010)	(0.007)	(0.010)	(0.008)
Summary Measures										
P: Selection Correction Terms Jointly Zero					<0.001	<0.001	<0.001	<0.001		
% with Incorrect Signs			0.00	0.09	0.45	0.36	0.27	0.05	0.32	0.05
Mean Squared Bias			0.099	0.307	1.754	0.850	0.365	0.103	0.380	0.074
Sample Size	98,417	98,417	62,186	62,186	62,186	62,186	62,186	62,186	62,186	62,186

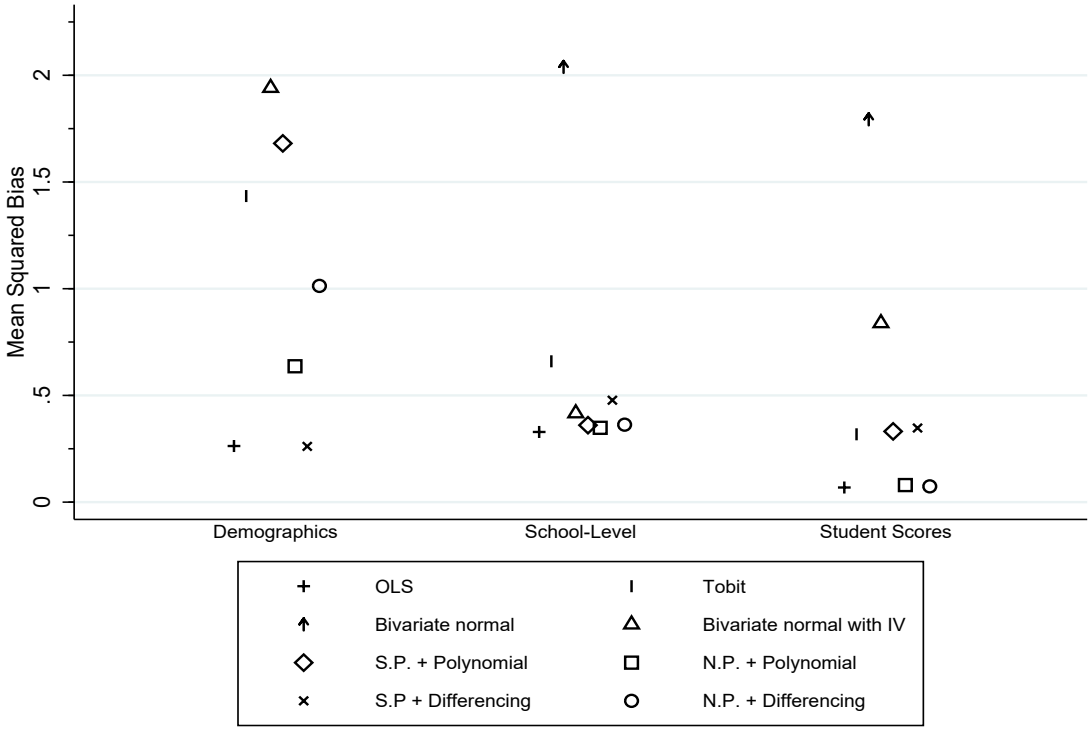
Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. Each column is from a separate regression. Standard errors estimated using 500 bootstrap replications reported in parentheses.

Appendix Figure II: Mean Squared Bias Using Standardized Covariates by Selection Correction and Covariate Set



Notes: Figure shows the mean squared bias for every selection correction and covariate set, where all covariates are standardized to mean of zero, standard deviation of one. We omit the bivariate normal correction without instruments for the model with the basic student demographics. Including this estimate, which has MSB of 1.19, compresses the other estimates and makes them difficult to read.

Appendix Figure III: Mean Squared Bias Using Unweighted Post-Reform Estimates by Selection Correction and Covariate Set



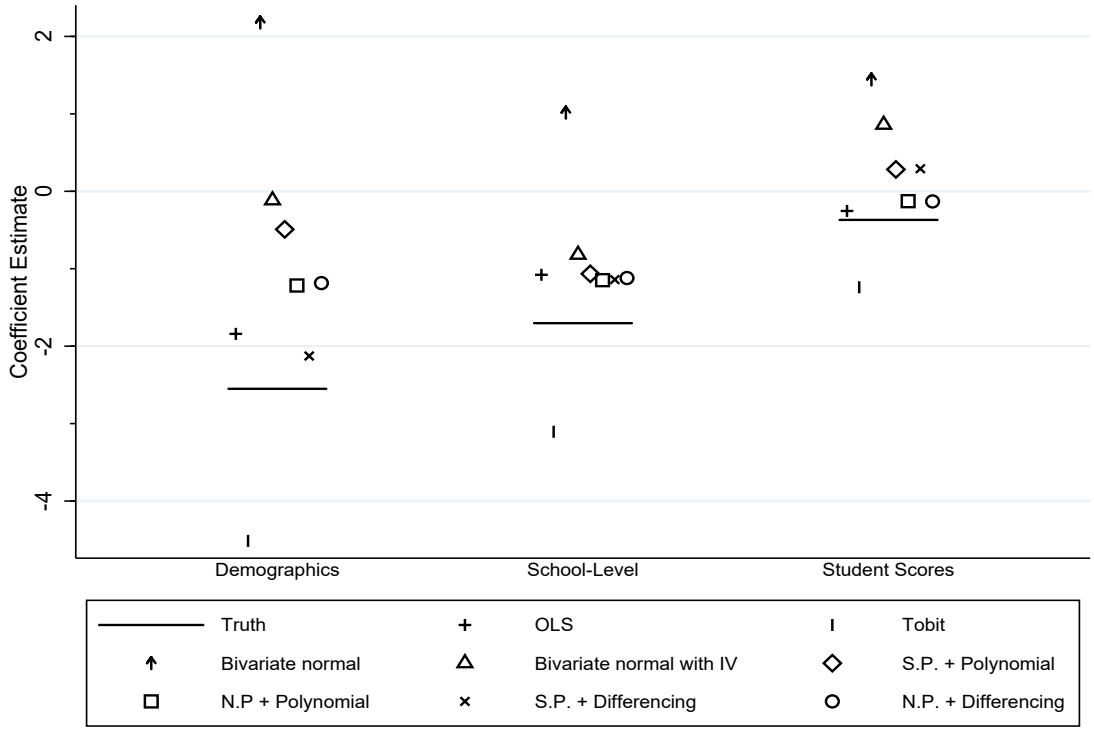
Notes: Figure shows the mean squared bias for every selection correction and covariate set, where the bias is calculated using coefficient estimates from the unweighted post-reform model. We omit the bivariate normal correction without instruments for the model with the basic student demographics. Including this estimate, which has MSB of 6.94, compresses the other estimates and makes them difficult to read.

squared bias results, coefficients on the indicator for free or reduced-price lunch receipt, and indicator for Black student race. There are no substantial differences between the results that uses the weighted and unweighted post-reform reference models.

Mean squared bias with different cohorts: We also verify that our finding are robust to comparing different pairs of pre- and post-reform cohorts. Our primary analysis compares the 2004/5 cohort to the 2007/8 cohort, as the mandatory ACT policy was piloted in some schools in 2006 and not implemented in all schools in 2007. We also compare the 2004/5 cohort to the 2006/7 cohort (Appendix Figure VI), the 2005/6 cohort to the 2006/7 cohort (Appendix Figure VII), and the 2005/6 cohort to the 2007/8 cohort (Appendix Figure VIII). Our main findings are still visible for all pairs of cohorts: no method systematically outperforms OLS, parametric estimators generally have higher biases than semiparametric estimators, and bias is generally higher when we use only a few discrete covariates.

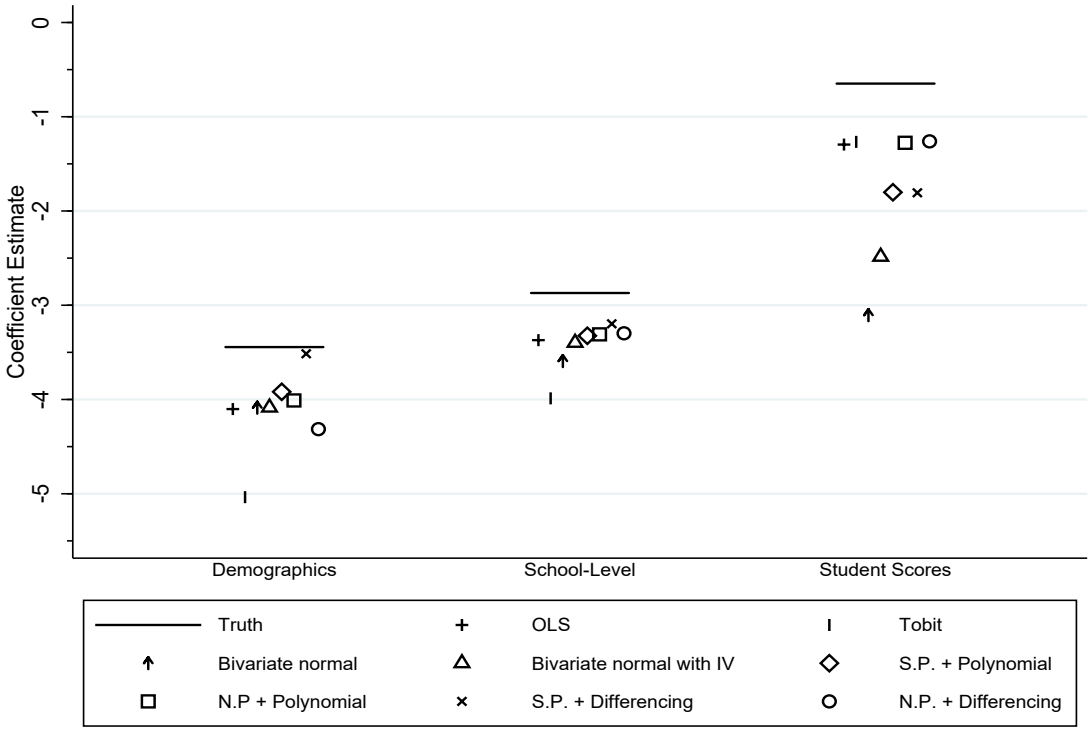
Mean squared bias with subsets of instruments: Our main analysis uses two sets of

Appendix Figure IV: Coefficient on Free Lunch Receipt Indicator Using Unweighted Post-Reform Estimates



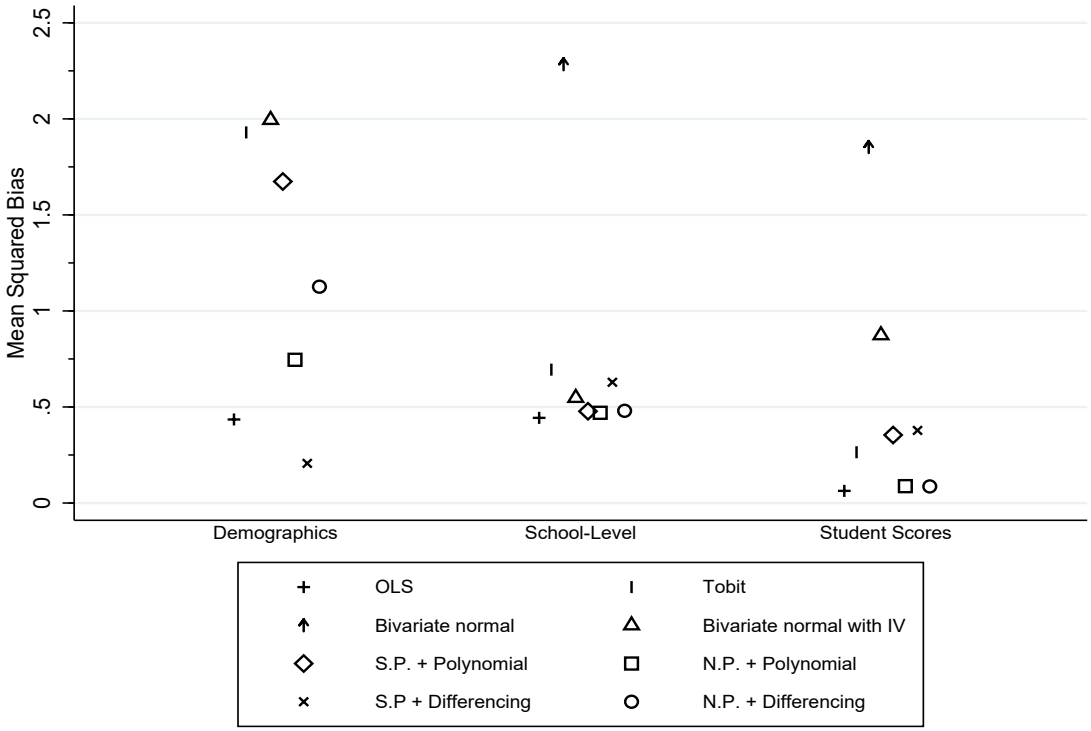
Notes: Figure shows the estimated coefficient on an indicator for free or reduced-price lunch receipt for every covariate set, for every selection correction and for the reference model that uses complete post-reform data without weights.

Appendix Figure V: Coefficient on Black Race Indicator Using Unweighted Post-Reform Estimates



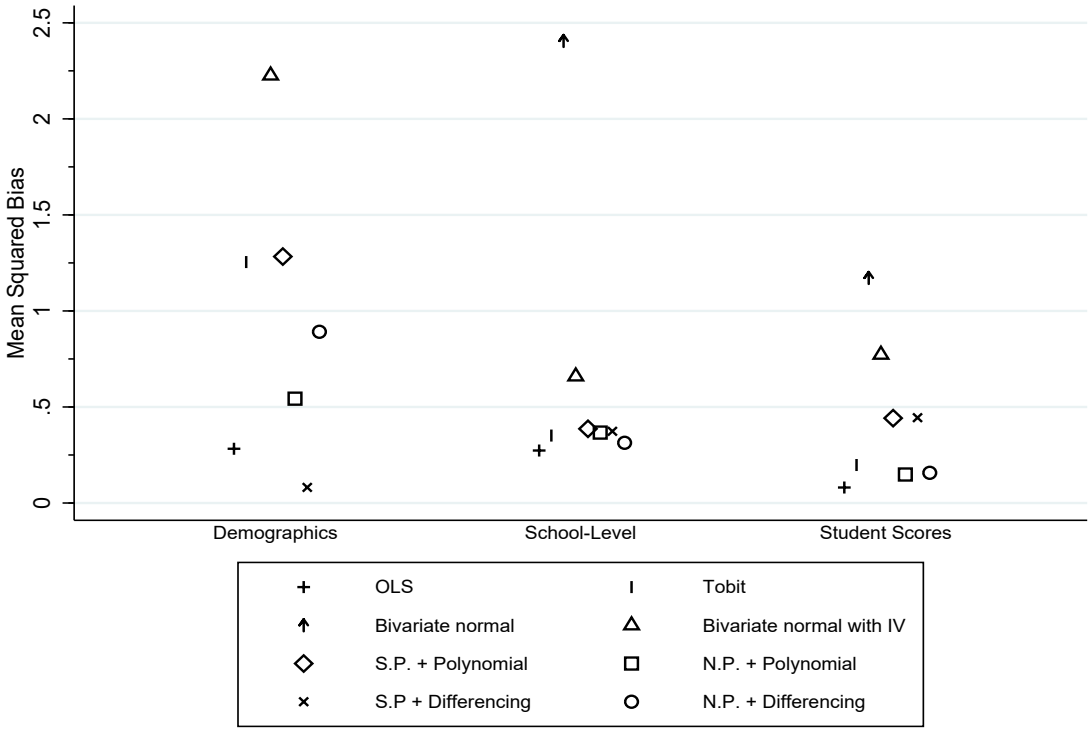
Notes: Figure shows the coefficient on an indicator for Black student race for every covariate set, for every selection correction and for the reference model that uses complete post-reform data without weights.

Appendix Figure VI: Mean Squared Bias Using 2004/5 and 2006/7 Cohorts by Selection Correction and Covariate Set



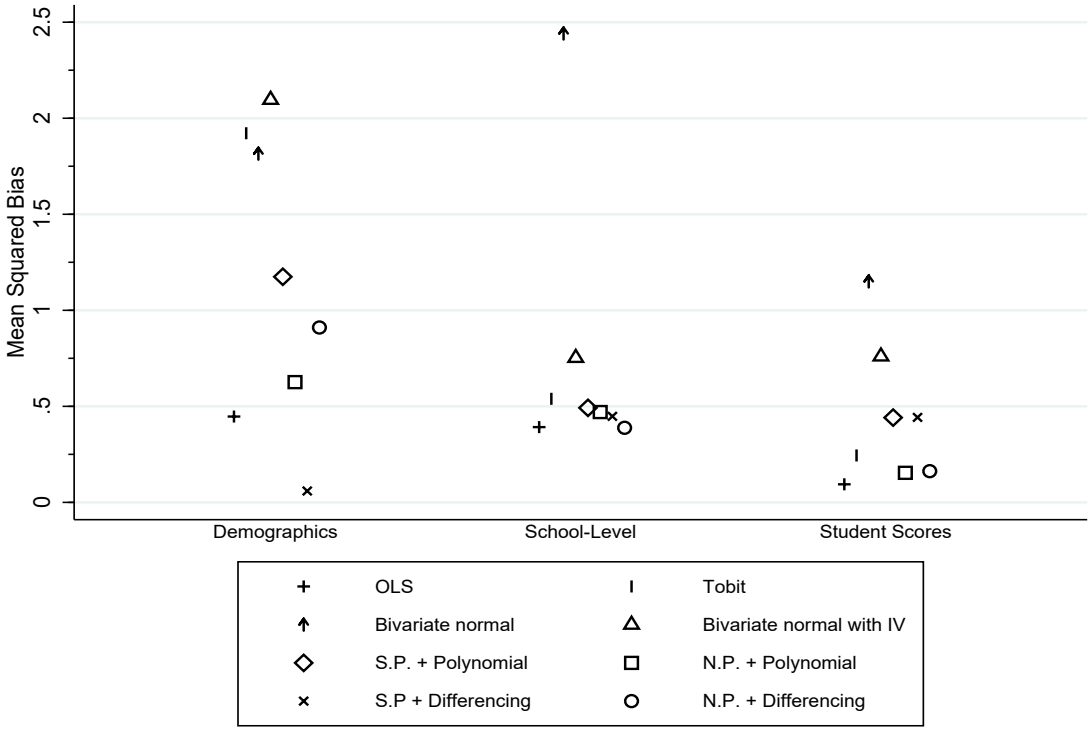
Notes: Figure shows the mean squared bias for every selection correction and covariate set, where the sample is students in the 11^{th} grade cohorts of 2004/5 and 2006/7. We omit the bivariate normal correction without instruments for the model with the basic student demographics. Including this estimate, which has MSB of 6.84, compresses the other estimates and makes them difficult to read.

Appendix Figure VII: Mean Squared Bias Using 2005/6 and 2007/8 Cohorts by Selection Correction and Covariate Set



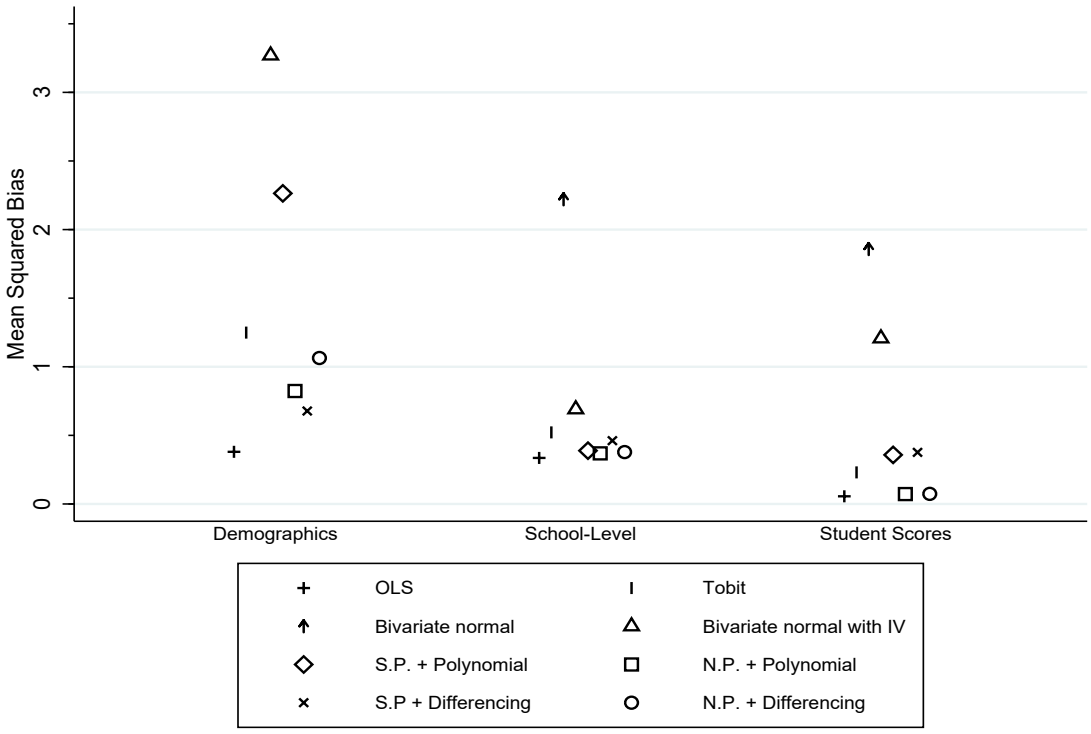
Notes: Figure shows the mean squared bias for every selection correction and covariate set, where the sample is students in the 11^{th} grade cohorts of 2005/6 and 2008. We omit the bivariate normal correction without instruments for the model with the basic student demographics. Including this estimate, which has MSB of 6.94, compresses the other estimates and makes them difficult to read.

Appendix Figure VIII: Mean Squared Bias Using 2005/6 and 2006/7 Cohorts by Selection Correction and Covariate Set



Notes: Figure shows the mean squared bias for every selection correction and covariate set, where the sample is students in the 11^{th} grade cohorts of 2006 and 2007.

Appendix Figure IX: Mean Squared Bias Using Only Distance Instruments by Selection Correction and Covariate Set

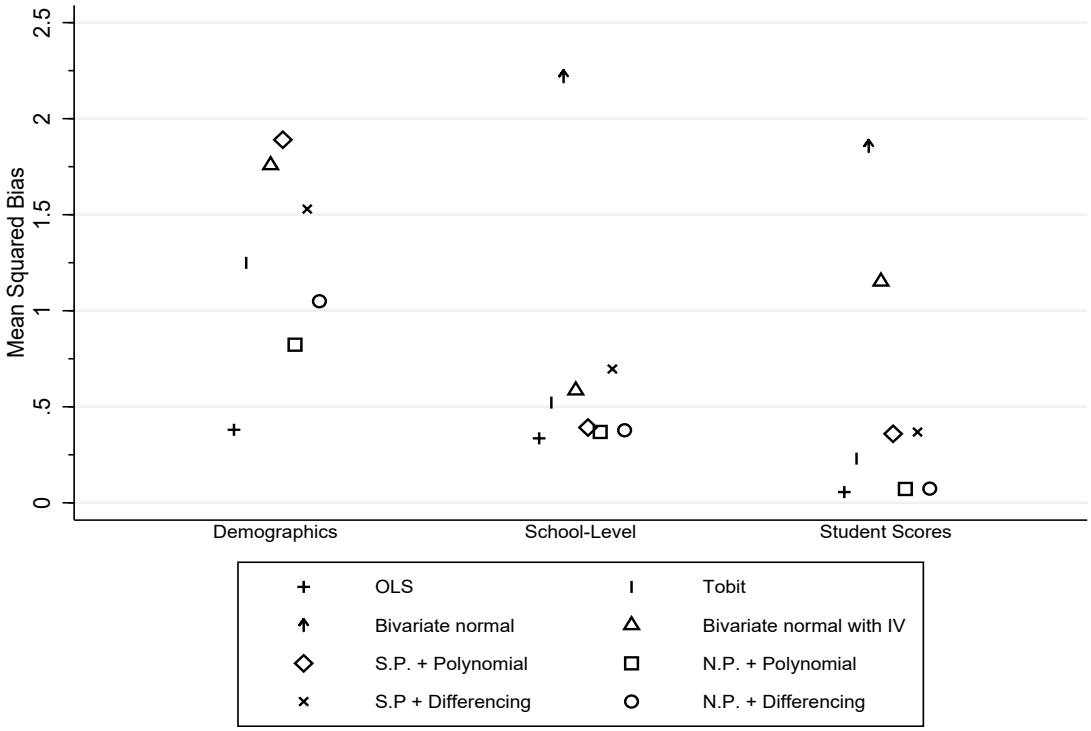


Notes: Figure shows the mean squared bias for every selection correction and covariate set, using only the driving distance instruments in the first stage models. We omit the bivariate normal correction without instruments for the model with the basic student demographics. Including this estimate, which has MSB of 7.54, compresses the other estimates and makes them difficult to read.

instruments related to distance to the ACT test center and exposure to severe weather near an ACT test center. We show in Appendix Figures IX and X that our main findings still hold when we use only the distance instruments or only the weather instruments.

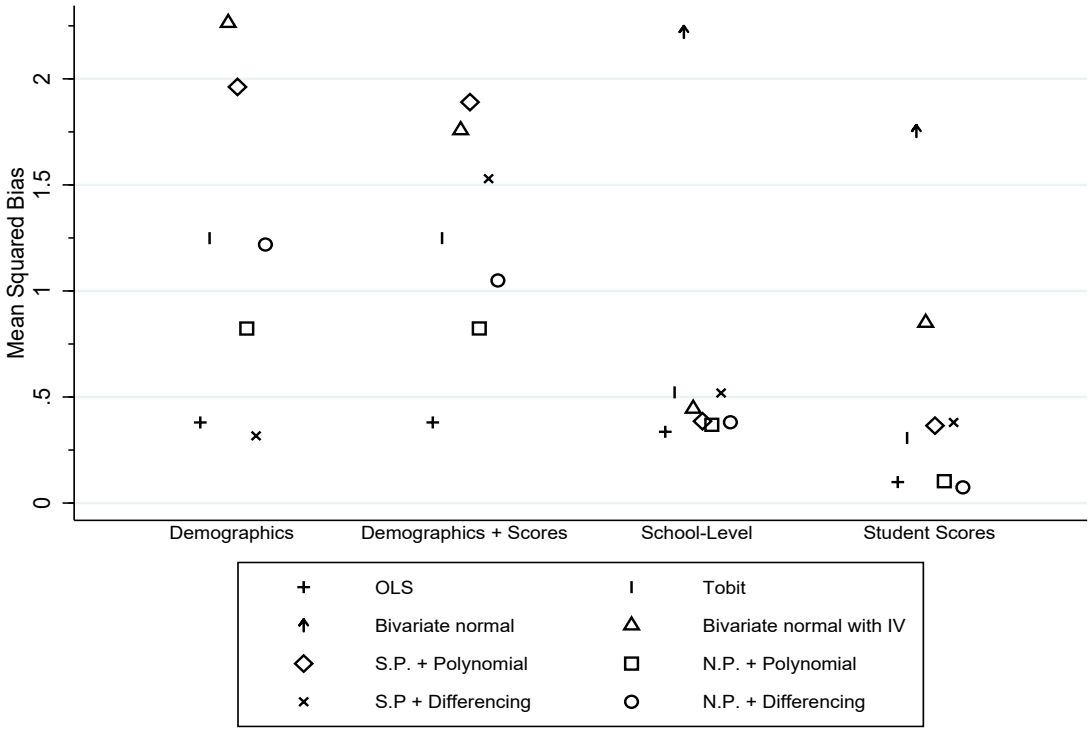
Mean squared bias with another covariate set: We further explore the importance of the choice of covariates with a fourth covariate set: student demographics and student test scores, without school- and district-level covariates. We might expect adding student test scores to substantially reduce mean squared biases relative to using only student demographics. The two test score measures explain 59% of the variation in ACT scores in a linear regression of selected scores, compared to 17% for all other covariates we observe. The pseudo-R^2 from regressing ACT-taking on these two measures is 0.19, compared to 0.06 for all other measures. In other education research, conditioning on lagged student test scores is particularly important for eliminating biases in value-added models (Angrist et al., 2013, 2017). However, mean squared bias from this fourth covariate set is only slightly lower than with only student demographics

Appendix Figure X: Mean Squared Bias Using Only Weather Instruments by Selection Correction and Covariate Set



Notes: Figure shows the mean squared bias for every selection correction and covariate set, using only the weather instruments in the first stage models. We omit the bivariate normal correction without instruments for the model with the basic student demographics. Including this estimate, which has MSB of 7.54, compresses the other estimates and makes them difficult to read.

Appendix Figure XI: Mean Squared Bias Using Extra Covariate Set With Student Demographics and Test Scores



Notes: Figure shows the mean squared bias for every selection correction and covariate set. The second column adds an extra covariate set consisting of basic student demograpics and prior test scores, but no school- and district-level characteristics. We omit the bivariate normal correction without instruments for the first two models, with the basic student demographics. Including these estimates, with MSBs of respectively 7.54 and 5.99, compresses the other estimates and makes them difficult to read.

(Appendix Figure XI), and this result persists when we use standardized covariates.

Mean squared bias with different series orders in the semiparametric first stage: Estimates of the coefficients of the ACT score models using semiparametric first stages may be sensitive to the series orders used in the first stages. This is particularly relevant for the differencing methods, as the consistency arguments in Ahn and Powell (1993) and Powell (1987) assume that the first stage model is undersmoothed, i.e. uses a higher series higher than might be chosen by cross-validation. We therefore estimate the mean squared biases of the polynomial and differencing methods using several different series orders as a sensitivity analysis.

The polynomial method's mean squared bias is generally robust to changes in the series order (Appendix Table XII). Undersmoothing with respect to the covariates slightly reduces bias only with the richest set of covariates. Perhaps surprisingly, oversmoothing with respect to the covariates slightly reduces bias with all but the basic (demographics only) set of covariates and sometimes leads to slightly lower bias than OLS. The series order in the instrument has

Figure XII: Mean Squared Bias of Semiparametric Polynomial Methods with Different First Stage Orders

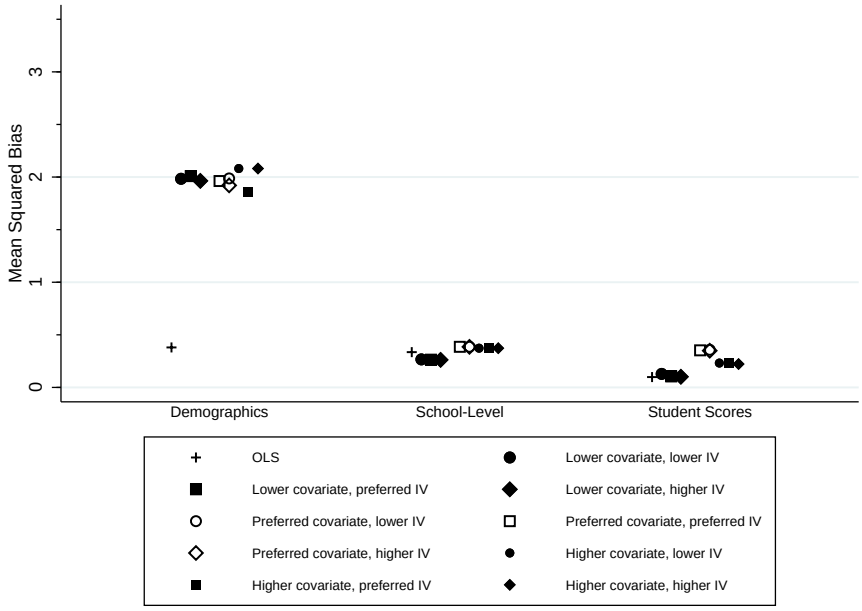


Figure shows the mean squared bias over all coefficients except the intercept from regressions using semiparametric (series logit) first stages and polynomial second stages, with different values of the series orders. The 'preferred" specifications use the series orders chosen by cross-validation. The "higher" and "lower covariate" models respectively increase and decrease the series order for the covariates by one. The "higher" and "lower instrument" models respectively increase and decrease the series order for the instruments by two.

a negligible effect on the bias. These patterns are similar for the differencing method (Appendix Table XIII). The most noticeable difference is that undersmoothing in the covariates substantially increases the bias when using the basic set of covariates.

Figure XIII: Mean Squared Bias of Semiparametric Differencing Methods with Different First Stage Orders

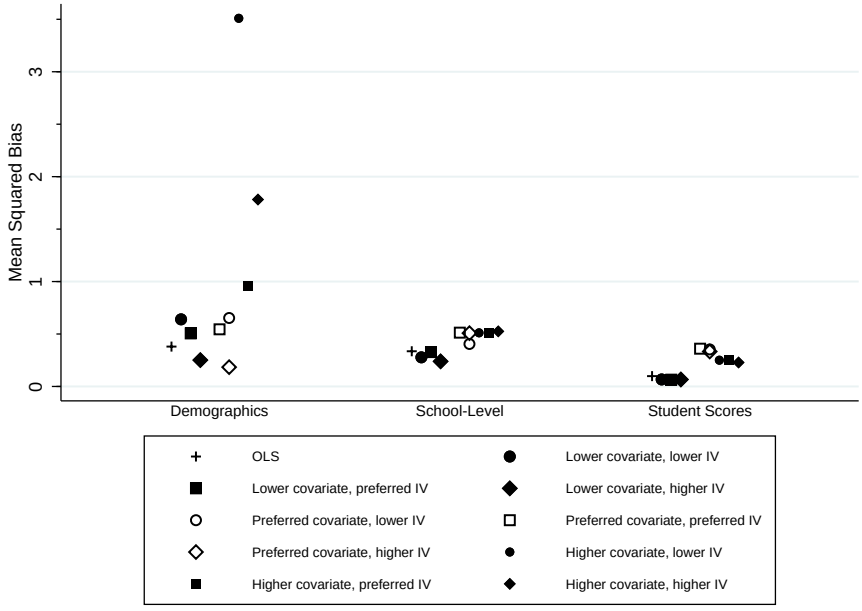


Figure shows the mean squared bias over all coefficients except the intercept from regressions using semiparametric (series logit) first stages and differencing second stages, with different values of the series orders. The 'preferred" specifications use the series orders chosen by cross-validation. The "higher" and "lower covariate" models respectively increase and decrease the series order for the covariates by one. The "higher" and "lower instrument" models respectively increase and decrease the series order for the instruments by two.

D Quantile Selection Correction Models

In the main paper we focus on the conditional mean function $\mathbb{E}[ACT_i^*|X_i]$. Researchers may also be interested in the conditional quantile functions $q(\tau, X) = F^{-1}_{ACT^*|X}(\tau)$ of the latent test score distribution for different values of τ. The conditional quantiles may behave in particularly interesting ways in the presence of sample selection. For example, if students with higher latent scores conditional on their covariates are more likely to select into test-taking, then sample selection will lead to larger biases at low than high quantiles.

Arellano and Bonhomme (2017), among others, propose methods for studying selection-corrected quantile estimation. Like the semiparametric correction methods for the mean, their approach assumes that there exist instruments Z that shift the probability of test-taking but are jointly independent of unobserved factors affecting both latent test scores and unobserved factors affecting test-taking. Their approach achieves nonparametric point identification if the instruments shift the probability of test-taking from zero to one (identification at infinity). If this condition fails, as it does in our data, their approach can deliver nonparametric bounds or point identification under parametric assumptions. Their approach assumes the two unobserved factors are strictly continuous and the distribution function of the latent outcomes conditional on the covariates is strictly increasing. This condition does not hold in our data as ACT scores, like many measures of educational achievement, take on a finite number of integer values.

The estimation approach involves three steps. First, we estimate the probability of test-taking given the covariates and instruments. Second, using the predicted probabilities, we estimate the parameters of the copula function that describes the joint distribution of the unobserved factors affecting latent test scores and test-taking. Third, using the copula function, we estimate the selection-corrected conditional quantile functions.

Following the empirical example in Arellano and Bonhomme (2017), and because we do not have identification at infinity, we impose parametric assumptions to implement all three steps. We use a probit model to estimate the predicted probabilities of test-taking, which is identical to the first stage of the bivariate normal selection method with instruments. We assume that the copula is Gaussian, so the second stage requires estimating only one parameter. We assume that the quantile functions are linear so that the third stage requires estimating only one parameter per covariate. We focus on quantiles 10, 20, 30, 40, 50, 60, 70, 80, and 90. We obtain confidence intervals using 50 replications of a nonparametric bootstrap that implements all three stages

Figure XIV: Mean Squared Bias of Quantile Regressions with Demographic Covariates Only

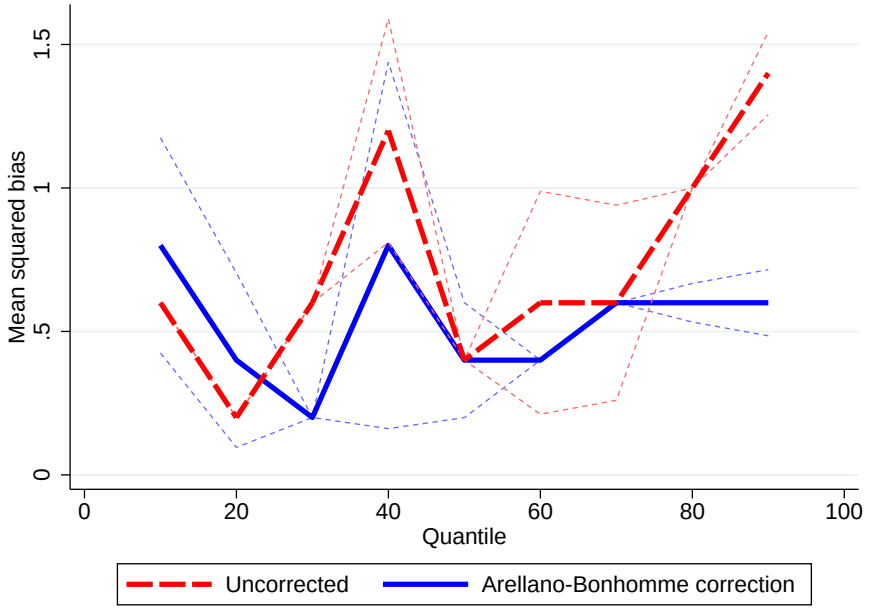


Figure shows the mean squared bias over all coefficients except the intercept from quantile regressions. The quantile regressions use only student demographics as covariates. The estimates labelled "uncorrected" run quantile regressions that ignore sample selection. The estimates labeled "Arellano-Bonhomme correction" apply the selection correction described in the text, with the driving distance and weather instruments. Bootstrap confidence intervals use 500 replications for the uncorrected estimates and 50 replications for the Arellano-Bonhomme-corrected estimates, running all steps of the estimation inside each replication.

of the estimation within each replication. We use only 50 replications because the second stage of the estimation relies on a computationally intensive grid search.

As a reference, we estimate the conditional quantiles in the complete post-reform data. We interpret the difference between the reference estimate for each covariate at each decile and the selection-corrected estimate for each covariate at each decile using pre-reform data as the bias. We calculate the mean squared bias at each decile over all the covariates except the intercept and display this in Appendix Figures XIV, XVI, and XVII for our usual three sets of covariates: respectively student demographics only, adding school- and district-level covariates, and adding student test scores. We also estimate uncorrected quantile regressions using pre-reform data that ignore the sample selection problem.

When we use only student demographics as covariates, the uncorrected estimates have mean squared biases of 0.6 to 1.4, with perhaps a slight trend toward higher bias at higher quantiles (Appendix Figure XIV). The corrected estimates are weakly lower than the uncorrected estimates for deciles 3-9, although the confidence intervals overlap for deciles 1-8.

When we include school- and district-level covariates, the uncorrected estimates are more

Figure XVI: Mean Squared Bias of Quantile Regressions with School-Level Covariates

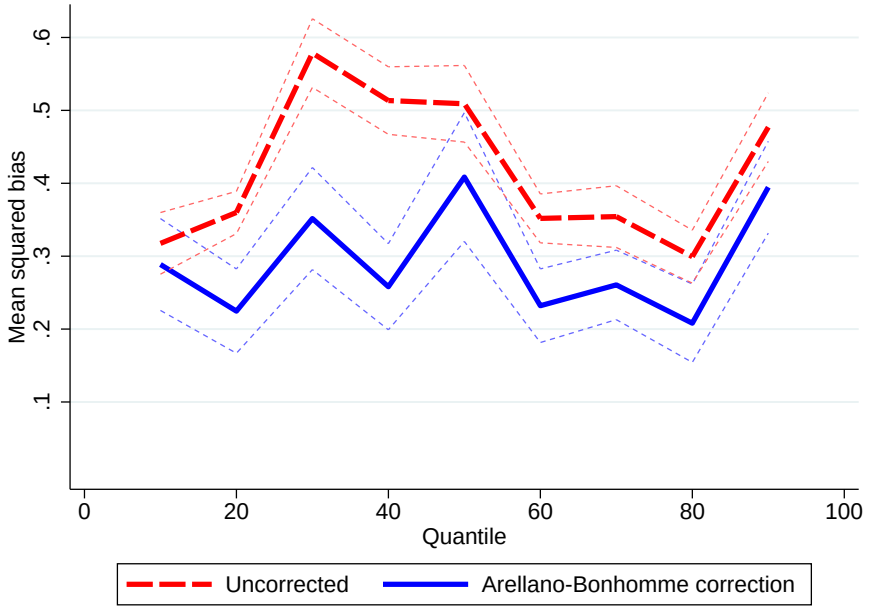


Figure shows the mean squared bias over all coefficients except the intercept from quantile regressions. The quantile regressions use student demographics and school- and district-level characteristics as covariates. The estimates labelled "uncorrected" run quantile regressions that ignore sample selection. The estimates labeled "Arellano-Bonhomme correction" apply the selection correction described in the text, with the driving distance and weather instruments. Bootstrap confidence intervals use 500 replications for the uncorrected estimates and 50 replications for the Arellano-Bonhomme-corrected estimates, running all steps of the estimation inside each replication. The confidence intervals for the uncorrected estimates have zero length at some deciles because there is no variation in the estimates over bootstrap replications, due to the discrete covariates and coarse outcome.

biased than the selection-corrected estimates at all deciles (Appendix Figure XVI). The pattern changes when we include student test scores as covariates: the uncorrected estimates are *less* biased at deciles 1-4 and 8-9, although the confidence intervals overlap at most quantiles and some of the differences are small (Appendix Figure XVII). None of the three figures show a clear trend toward lower bias at higher quantiles.

We conclude that there is suggestive evidence that the selection-corrected quantile method has lower bias than the uncorrected method. But this pattern is not robust across quantiles and covariate sets and the differences are seldom large relative to the confidence intervals. This pattern may also change with different parametric assumptions used to estimate the corrected estimates, or with instruments that more strongly shift the probability of test-taking. We view detailed investigation of the relative performance of selection-corrected and uncorrected methods as a topic for future work.

Figure XVII: Mean Squared Bias of Quantile Regressions with Student Test Score Covariates

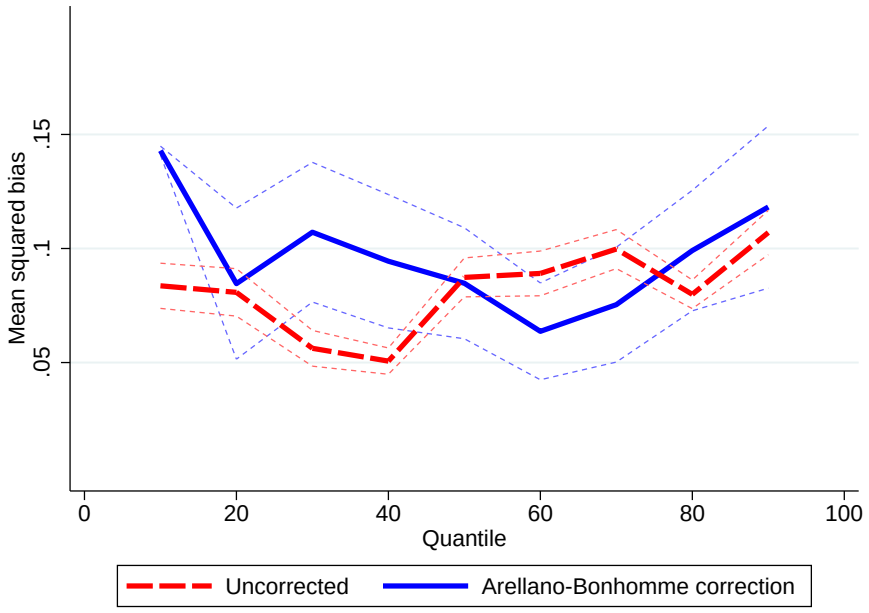


Figure shows the mean squared bias over all coefficients except the intercept from quantile regressions. The quantile regressions use student demographics, student test scores, and school- and district-level characteristics as covariates. The estimates labelled "uncorrected" run quantile regressions that ignore sample selection. The estimates labeled "Arellano-Bonhomme correction" apply the selection correction described in the text, with the driving distance and weather instruments. Bootstrap confidence intervals use 500 replications for the uncorrected estimates and 50 replications for the Arellano-Bonhomme-corrected estimates, running all steps of the estimation inside each replication.

Appendix References

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

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

Angrist, J., P. Hull, P. Pathak, and C. Walters (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics*, 132, 871–919.

Angrist, J., P. Pathak, and C. Walters (2013): "Explaining Charter School Effectiveness." *American Economic Journal: Applied Economics*, 5, 1–27.

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

Borra, S. and A. Di Ciaccio (2010): "Measuring the Prediction Error. A Comparison of Cross-validation, Bootstrap and Covariance Penalty Methods." *Computational Statistics and Data Analysis*, 54, 2976–2989.

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

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, Toronto: University of Toronto Press.

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

Chen, S. and S. Khan (2003): "Semiparametric Estimation of Heteroskedastic Sample Selection Models." *Econometric Theory*, 19, 1040–1064.

D'Haultfoueille, X. and A. Maurel (2013): "Another Look at Identification at Infinity of Sample Selection Models," *Econometric Theory*, 29, 213–224.

Donald, S. (1995): "Two Step Estimation of Heteroskedastic Sample Selection Models." *Journal of Econometrics*, 65, 347–380.

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

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

——— (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

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

Heckman, J. and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and S. Burton, Econometric Society Monograph Series.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, 71, 1161–1189.

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

Lee, F.-L. (1982): "Some Approaches to the Correction of Selectivity Bias." *Review of Economic Studies*, 49, 355–372.

——— (1983): "Generalized Econometric Models with Selectivity." *Econometrica*, 51, 507–512.

Lewbel, A. (2007): "Endogenous Selection or Treatment Model Estimation," *Journal of Econometrics*, 141, 777–806.

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

Olsen, R. (1980): "A Least Squares Correction for Selectivity Bias." *Econometrica*, 48, 1815–1820.

Pagan, A. and A. Ullah (1999): *Nonparametric Econometrics*, Cambridge: Cambridge University Press.

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, 53–68.

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