ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-37

Quasi-Experimental Evaluation of Alternative Sample Selection Corrections

Robert Garlick
Duke University

Joshua Hyman
University of Connecticut

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

VERSION: November 2018

Suggested citation: Garlick, R., & Hyman, J. (2019). Quasi-Experimental Evaluation of Alternative Sample Selection Corrections (EdWorkingPaper No.19-37). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-37

Quasi-Experimental Evaluation of Alternative Sample Selection Corrections*

Robert Garlick† and Joshua Hyman‡

November 25, 2018

Abstract

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

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

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

†Department of Economics, Duke University

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2 Sample Selection, Corrections, Evaluation Criteria

2.1 The Sample Selection Problem

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Selection Correction Methods

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

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

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

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

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

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

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

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

to construct two selection corrections for the ACT score model.

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

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

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

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

2.3 Evaluating Alternative Selection Correction Methods

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

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

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

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

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

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

3 Context, Data, and the Extent of Selection

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

3.1 Data

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

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

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

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

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

Table 1. Sample Means of Michigan 11th Grade Cohorts

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

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

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

3.2 Modeling ACT-Taking

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

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

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

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

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

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

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

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

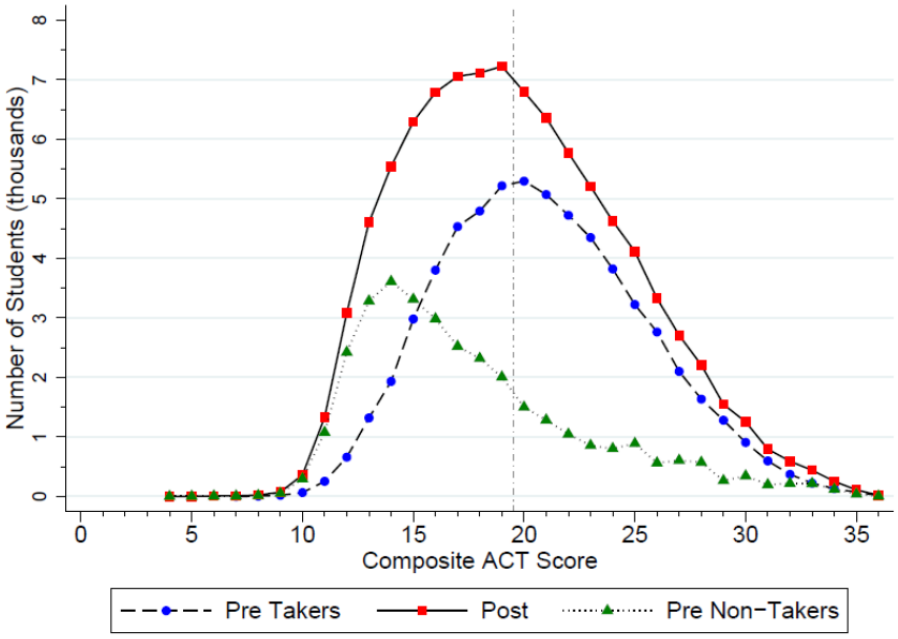
We estimate the latent ACT score distribution for non-takers by subtracting the number of test-takers with each ACT score in the pre-period from the number with each score in the post-period. We reweight the post-policy cohort to have the same number of students and distribution of observed characteristics.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

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

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

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

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



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

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

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

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

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

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

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

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

4 Results

4.1 Comparing Sample Selection Corrections

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

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

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

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

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

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

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

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

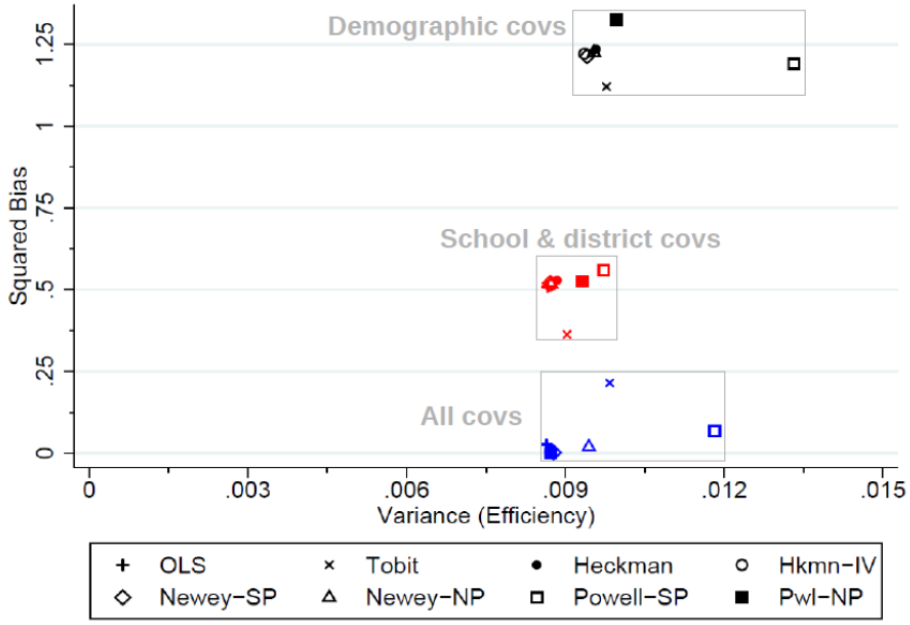
Our second selection correction is a Type 1 Tobit model, censoring at the 36th percentile of the post-policy ACT score distribution, as the test-taking rate in the pre-policy period is 64%. The predicted mean is similar to that from OLS. We next show results from the Heckman two-stage correction procedure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

4.2 Comparing Selection Corrections' Performance with Different Predictors

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

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

Figure II. MSE Comparison Across Correction Methods and Covariate Sets



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

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

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

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

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

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

4.3 Comparing Selection Corrections on Other Criteria

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

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

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

4.4 Comparing Selection Corrections for Different Subgroups

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

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

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

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

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

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

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

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

5 Explaining Similar Results across Different Corrections

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6 Conclusion

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Robert Garlick* and Joshua Hyman†

November 25, 2018

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 the MDE data with three other sources using a restricted access computer at the MDE. First, using student name, date of birth, sex, race, and 11th grade home zip code, 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-policy cohorts, we use students' first ACT score, which is typically from 11th grade, but in some cases is from 12th grade. For students taking the SAT but not the ACT pre-policy, 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. We geocode student home addresses during 11th grade and the addresses of these test centers to construct a student-level driving distance from 11th grade home to the nearest ACT test center. When a student has multiple addresses during 11th grade, we use the one with the shortest distance to a center. When 11th grade home address is missing, we use home address during the surrounding grades. The ≈2% of students with a missing address during every high school grade are dropped from the pre- and post-policy samples. Appendix Table 1 shows detailed summary statistics for driving distance.

*Department of Economics, Duke University

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

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-policy period, the distance is the 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.

Third, 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-policy cohorts, we measure students' ACT scores using their first attempt. This is typically from 11th grade, but in some cases is from 12th grade. For students taking the SAT but not the ACT pre-policy, 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 1 shows the distribution of observed pre- and post-policy test scores and the difference between these, interpreted as a measure of the latent scores of non-takers. Unlike Figure 1 in the main paper, this figure shows the distributions for subgroups based on race and free lunch (in)eligibility.

We construct student-level 8th and 11th grade test scores from state-wide assessments. For the 8th grade test score, we use the average of a student's standardized math and English scores. For 11th grade, we use standardized social studies scores because post-policy 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 predictors.

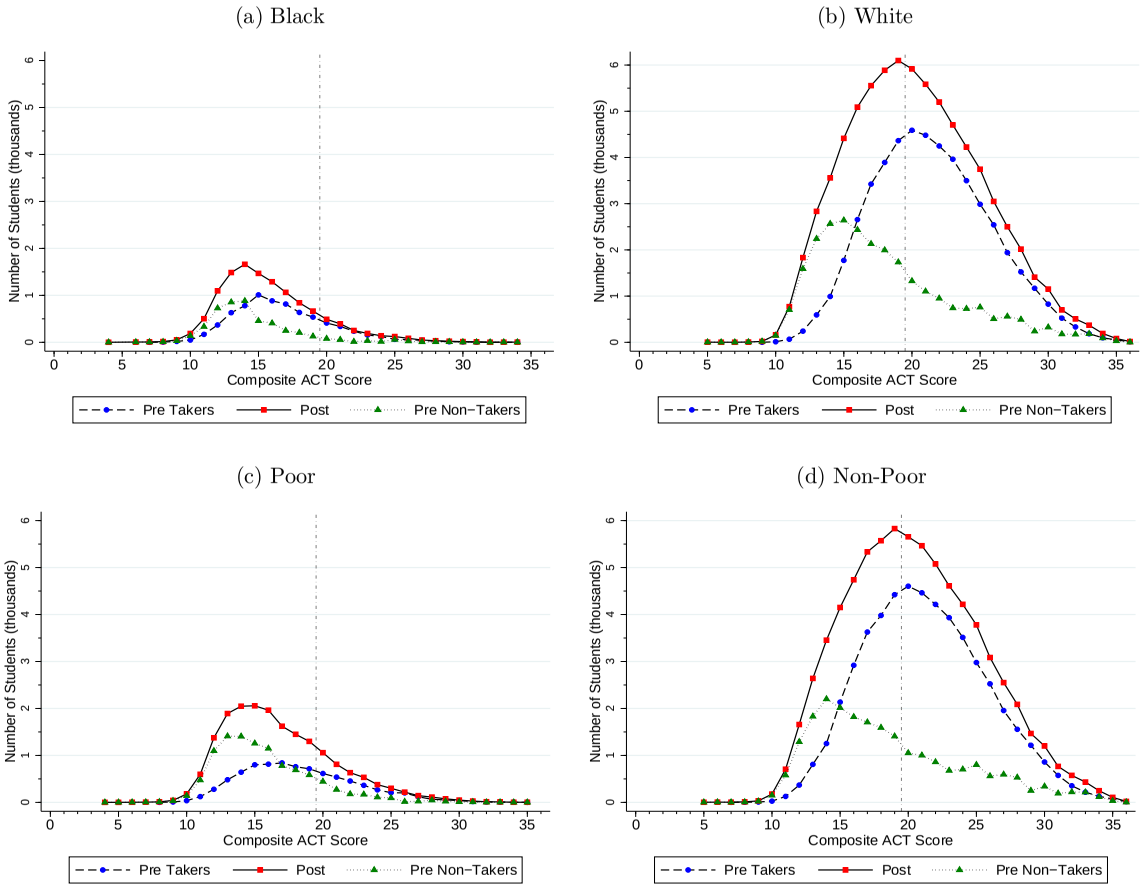
Sample restrictions: Our main analysis excludes the small number of students who do not complete high school and students who take the special education version of the state-wide 11th grade test. These students are not suited for our analysis because they are not required to take the ACT in either period. 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 D, we exclude these schools.

Additional statistics: Appendix Figure 2 graphically displays the test score gaps by race and free lunch (in)eligibility observed in the reference distribution and estimated from the selection-corrected pre-policy distributions. This displays the same information as Table 4 in a more compact form.

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 semiparametric models. We sum-

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



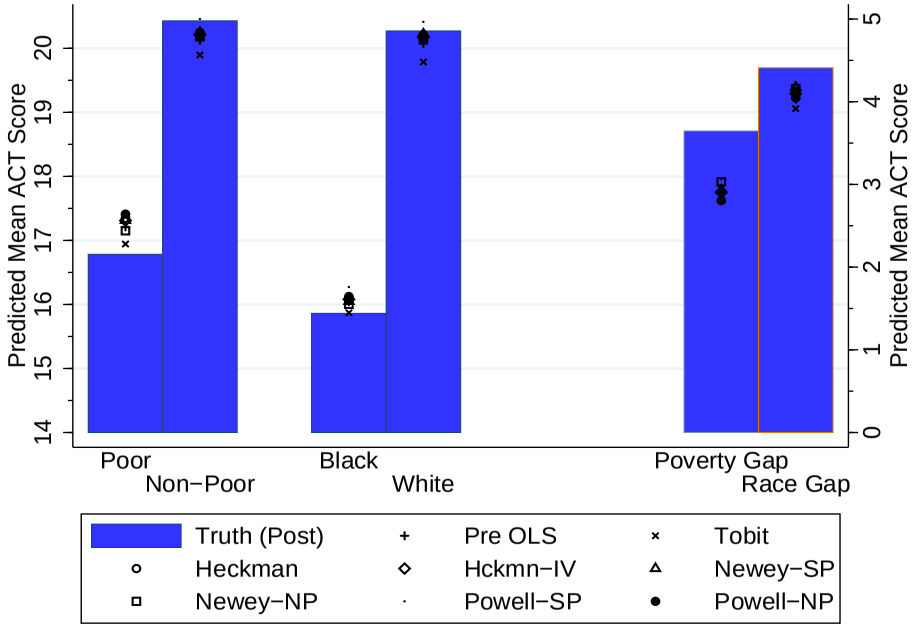
Notes: Figures show 1) the distribution of ACT scores pre-policy, 2) the distribution post-policy reweighted following DiNardo, Fortin, and Lemieux (1996) to resemble the pre-policy cohort, and 3) the difference between (1) and (2), which is the latent score distribution among non-takers in the pre-period.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

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

	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 is a Kolmorogov-Smirnov non-parametric test of the equality of the distributions.

Appendix Figure II: Predicted Mean ACT Score and Gaps in Mean ACT Score



Notes: The leftmost four bars and markers show the "true" and predicted mean latent ACT score across correction methods by poverty status and race. The rightmost two bars and markers show the "true" and predicted gaps in these measures across correction methods by poverty status and race. All specifications include basic student demographics, school- and district-level covariates, and student 8th and 11th grade test scores.

marize these models in Appendix Table 3 We do not evaluate imputation methods, bounding methods, or methods focused on identification at infinity without instruments.[1]

The variances for all models are estimated using a nonparametric bootstrap that resamples schools.[2] The bootstrap is not valid for the nonparametric first stage estimator we use (Abadie and Imbens, 2008). However, to the best of our knowledge, the econometric literature does not provide an analytical variance estimator for two-stage semiparametric selection correction models with clustered data. We follow most applied researchers in using the bootstrap 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 adjustment for sample selection bias using ordinary least squares. Specifically, we estimate the model

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

for the test-takers. This is a special case of system (1) where u_i and ϵ_i are independent and $Pr\left(TAKE_i = 1|X_i\right) > 0$ for all X_i. 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-policy period) and lower test scores (in the post-policy 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.

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

[1] Lewbel (2007) and D'Haultfoueille and Maurel (2013) propose methods that identify selection models without instruments or parametric assumptions. Intuitively, both approaches rely on identifying a subsample of students whose probability of taking the ACT is arbitrarily close to one. There is no missing data problem within this subsample, which facilitates identification of the parameters of the outcome equation. Both approaches make assumptions that are unlikely to hold in our setting.

[2] Analytical variance estimators have been developed for one-stage nonparametric estimators with clustered data (Hanson and Sunderam, 2012) or two-stage nonparametric estimators with independent data (Mammen, Rothe, and Schienle, 2016).

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	
Heckman	F(ε,u) is bivariate normal[1]	Unnecessary	Probit	Inverse Mills ratio	
Heckman with IV		Necessary			
Semiparametric Newey	No restriction on joint distribution	Necessary	Series logit	Polynomial approximation	
Nonparametric Newey			Nonparametric		
Semiparametric Powell			Series logit	Differenced out	
Nonparametric Powell			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.

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

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

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 34th percentile of the post-policy distribution of test scores, as the test-taking rate in the pre-policy period is 64%. Results reported in Section 4 are robust to substantial changes in this threshold.

B.2 Parametric Multiple-Equation Corrections for Sample Selection Bias ("Heckman" and "Heckman with IV")

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

$$ACT_i = X_i\beta + \sigma_u \rho_{\epsilon,u} \lambda\left(Z_i\gamma\right) + \epsilon_i \text{ if } TAKE_i^* \geq 0 \tag{3a}$$

$$TAKE_i^* = X_i\delta + Z_i\gamma + u_i \tag{3b}$$

$$TAKE_i = \begin{cases} 1 & \text{if} \quad TAKE_i^* \geq 0 \\ 0 & \text{if} \quad TAKE_i^* < 0 \end{cases} \tag{3c}$$

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 (3a) by ordinary least squares provides a consistent estimator of β. We estimate equation (3b) using only X_i as predictors ("Heckman") and also including a set of instruments Z_i that are excluded from equation (3a) and assumed not to affect test scores directly ("Heckman with IV"). 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). We report the coefficient estimates for equation (3b) in Appendix Table 4. In Appendix Figure 3 we show that the inverse Mills ratio is roughly linear when we use only demographic predictors but convex in $X_i\hat{\delta} + Z_i\hat{\gamma}$ when we use richer predictors.

This approach allows ACT-taking and ACT scores to depend jointly on both observed and unobserved characteristics. Unlike the Tobit model, the Heckman 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.[3]

B.3 Semiparametric Multiple-Equation Corrections for Sample Selection Bias ("Newey" and "Powell")

We now consider models of the form

$$ACT_i^* = X_i\beta + h\left(\hat{g}\left(X_i, Z_i\right)\right) + \epsilon_i \tag{4a}$$

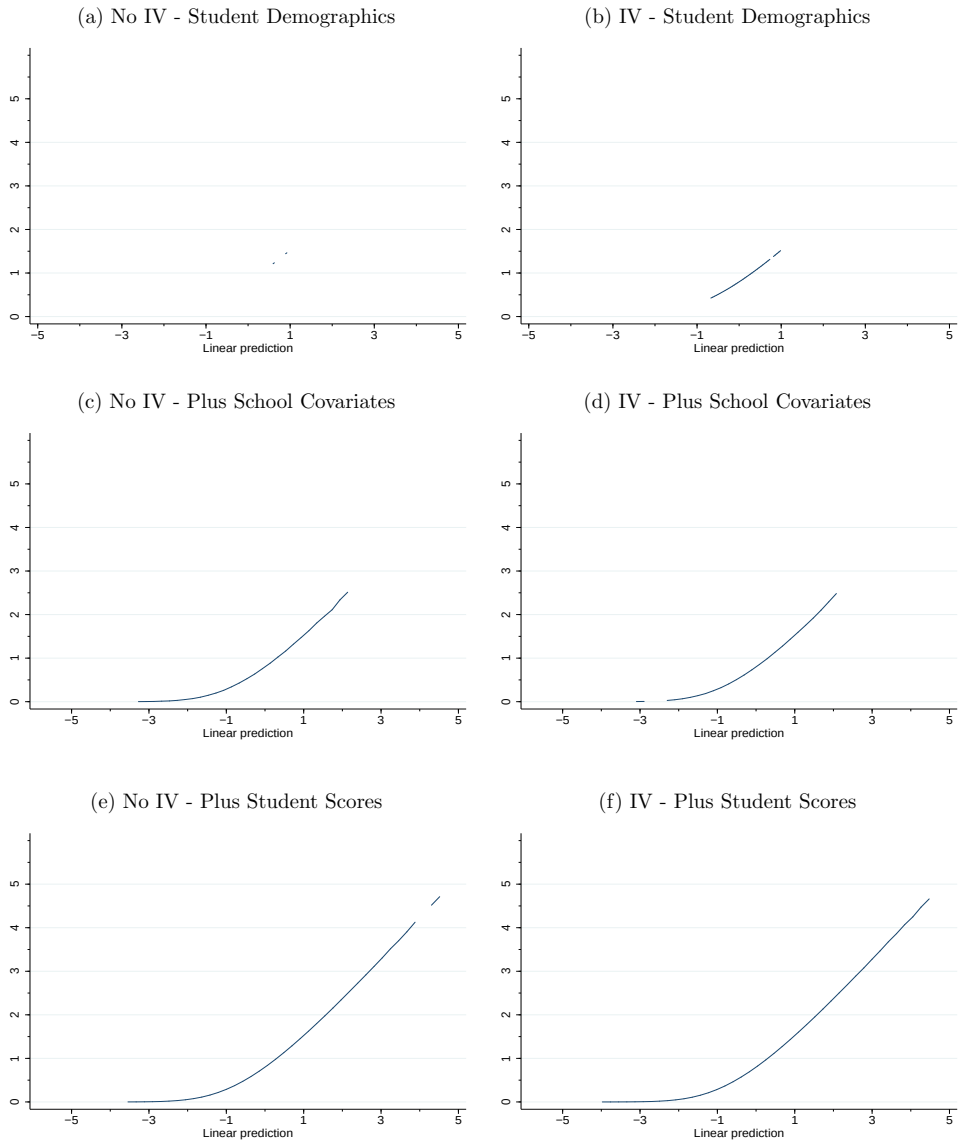
$$TAKE_i^* = g\left(X_i, Z_i\right) + u_i \tag{4b}$$

$$TAKE_i = \begin{cases} 1 & \text{if} \quad TAKE_i^* \geq 0 \\ 0 & \text{if} \quad TAKE_i^* < 0 \end{cases} \tag{4c}$$

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

[3]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 almost identical to those from the Heckman model.

Appendix Figure III: IMRs vs Linear Predictions From Probits



Notes: Figures plot the inverse Mills ratio against the linear prediction from the first stage Heckman corrections, with and without an IV and by predictor set. This demonstrates that the student test scores and school- and district-level predictors generate substantial nonlinearity in the inverse Mills ratio. This nonlinearity facilitates separate identification of the selection correction term and the predictors in the ACT score model.

Appendix Table 4: First Stage Results

	Coef.	Std. Err.
Student-Level		
Distance (Miles)	-0.007	0.001
Distance Squared (/ 10)	0.003	0.001
Free Lunch	-0.111	0.005
Female	0.067	0.003
Black	0.106	0.009
Hispanic	-0.004	0.012
Other Race	0.084	0.011
8th Grade Test Score	0.114	0.003
11th Grade Test Score	0.147	0.002
School-Level		
Average Class Size	0.000	0.000
Percent Free Lunch	0.001	0.034
Percent Black	-0.003	0.087
Grade 11 Enrollment	0.000	0.000
Average 8th Grade Score	0.127	0.020
Average 11th Grade Score	0.020	0.016
District-Level		
Suburb	0.006	0.011
Town	0.025	0.015
rural	0.034	0.013
Grade 11 Enrollment	0.000	0.000
Average Class Size	-0.005	0.002
Percent Free Lunch	-0.081	0.041
Percent Black	0.171	0.092
Student-Counselor Ratio	0.000	0.000
Local Unemployment Rate	-0.003	0.002

Notes: Table shows marginal effects from the first stage probit regression of a dummy for whether a student takes the ACT or SAT on student, school, and district demographics and test scores.

models (Pagan and Ullah, 1999), all of which 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, following Hirano, Imbens, and Ridder (2003). We assume that we can approximate $g(X_i, Z_i)$ using polynomial expansions in X_i and Z_i, inside a logistic link function:

$$Pr(TAKE_i = 1) = L\left(\sum_{p=1}^{P}\left(\sum_{k=1}^{K}\theta_k X_{i,k}\right)^p + \sum_{q=1}^{Q}\psi Z_i^q\right) \tag{5}$$

We observe multiple predictors $X_{i,1}, \ldots, X_{i,K}$, so we include polynomial terms in each element of X_i and interactions between the elements of X_i. We observe only a single instrument Z_i, so we include only polynomial terms of the instrument. 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.[4] 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 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. The sparse set of predictors includes only 1 continuous instrument and 6 binary predictors, so we do not need to consider values of P greater than 6. The richer sets of predictors include up to 24 binary and 14 continuous covariates. For these sets of predictors, we consider only $P \in \{1, 2, 3\}$. The fourth order expansion with all 38 covariates generates almost 80,000 predictors and estimation

[4]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, Imbens, and Ridder (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.

is infeasible without dimension reduction techniques.

This cross-validation algorithm selects a second-order polynomial in the predictors for the basic, school/district, and student test score sets of predictors. This polynomial contains linear terms in all predictors, quadratic terms in all continuous variables, and all pairwise interaction terms.[5] This yields 17, 585, and 731 terms when using the basic, school/district, and student test score sets of predictors. Some pairwise interaction terms are omitted because they are mutually exclusive (e.g. black and Hispanic). The cross-validation algorithm selects seventh-, eighth-, and seventh- order polynomials in the instrument when using respectively the basic, school/district, and student test score sets of predictors.

This semiparametric model therefore differs from the probit model used in the Heckman selection correction in three ways: the semiparametric model includes quadratic and interaction terms in the predictors, includes a seventh or eighth order polynomial in the instrument instead of a second order polynomial, and uses a logit instead of a probit link function. Nonetheless, we see in Appendix Table 5 that the predicted probabilities of ACT-taking are similar, with correlations of at least 0.93. The predicted probabilities are robust to all polynomial orders that we consider ($P \leq 3$ and $Q \leq 8$).

Our second ACT-taking model uses a K-nearest neighbor matching approach. 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 predictors/instruments and the covariances between predictors/instruments. We then identify the K nearest neighbors of each observation with respect 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.[6] This estimator directly constructs the conditional mean $\mathbb{E}[W_i = w]$ at each value w without making assump-

[5]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.

[6]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 $d_{i,k} = 0$.

Appendix Table 5. ACT-Hat Correlations, by Selection Correction

			Heckman		Newey		Powell	
	OLS	Tobit	No IV	With IV	Series Lgt	N.P.	Series Lgt	N.P.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: X = Student Demographics								
OLS	1.000							
Tobit	1.000	1.000						
Heckman (no IV)	0.999	0.999	1.000					
Heckman (with IV)	0.994	0.993	0.994	1.000				
Newey - Series Logit	0.989	0.989	0.992	0.994	1.000			
Newey - Nonparametric	0.997	0.996	0.997	0.994	0.993	1.000		
Powell - Series Logit	0.996	0.995	0.995	0.989	0.985	0.992	1.000	
Powell - Nonparametric	0.989	0.990	0.989	0.983	0.979	0.986	0.989	1.000
Panel B: X = …Plus School-Level Covs								
OLS	1.000							
Tobit	0.974	1.000						
Heckman (no IV)	0.996	0.963	1.000					
Heckman (with IV)	0.999	0.971	0.998	1.000				
Newey - Series Logit	0.997	0.971	0.997	0.998	1.000			
Newey - Nonparametric	0.997	0.972	0.996	0.997	0.998	1.000		
Powell - Series Logit	0.995	0.969	0.993	0.995	0.993	0.993	1.000	
Powell - Nonparametric	0.981	0.996	0.971	0.978	0.978	0.979	0.979	1.000
Panel C: X = …Plus Student Test Scores								
OLS	1.000							
Tobit	0.995	1.000						
Heckman (no IV)	0.985	0.980	1.000					
Heckman (with IV)	0.990	0.985	0.999	1.000				
Newey - Series Logit	0.984	0.980	0.995	0.995	1.000			
Newey - Nonparametric	0.997	0.992	0.989	0.993	0.990	1.000		
Powell - Series Logit	0.985	0.988	0.976	0.980	0.975	0.983	1.000	
Powell - Nonparametric	0.977	0.991	0.959	0.965	0.963	0.975	0.976	1.000

Notes: Table reports correlations of predicted ACT scores pre-policy by covariate set and selection correction model.

tions about the function $g(.)$. We report results in this paper using $K = 100$, but we find similar results with $K = 10$ and $K = 1000$. 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 (Newey, 2009).[7] We select the order of the series using leave-one-out cross-validation. We then estimate equation (4a) 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 thirteenth, fourth, and ninth order polynomials for the selection term when we use a semiparametric first stage with respectively basic, school/district, and student test score sets of predictors. The cross-validation algorithm selects third, sixth, 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 predictors. The main results are robust to choice of the polynomial orders between one and sixteen.

Second, we remove $h(.)$ from equation (4a) using a differencing approach (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, $dh_i = h_i - \frac{1}{N-1}\sum_{j \neq i} w(i,j) h_j \approx 0$. Hence we can rewrite equation (4a) as

$$dACT_i = dX_i \beta + d\epsilon_i \tag{6}$$

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

[7] 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 almost identical using predicted probabilities. Note that concerns about "forbidden regression" are not necessarily applicable here, as the series in Newey (2009) is simply an approximating function and not an exact replacement for the selection bias term $\mathbb{E}[ACT_i | X_i] = X_i\beta + \mathbb{E}[u_i > g(X_i, Z_i)]$.

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.[8]

Both the series ("Newey") and differencing ("Powell") approaches 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 exclusion restriction: 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 (4a) 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" argument requires an unusually strong excluded instrument (Andrews and Schafgans, 1998; Chamberlain, 1986; Heckman, 1990). We exclude driving distance from the student's home to the nearest ACT center from the outcome equation. The probability of ACT-taking falls by 4 percentage points with a move from the 5^{th} to the 95^{th} percentile of this variable. This does not satisfy the identification at infinity argument, like most excluded instruments in the empirical literature, (Card, 1995; Kane and Rouse, 1995; Bulman, 2015). This means we can identify the shape of ACT test score distribution around the mean, but not necessarily the mean. However, with the richer sets of predictors, we find that the semiparametric models almost perfectly predict the mean, suggesting this problem is not quantitatively important in practice.

[8]The consistency theorems 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.

Second, the semiparametric models yield consistent estimators only with appropriate choices of the tuning parameters: respectively the order of the series and the weighting function. The parameter estimates may in principle be very sensitive to the choice of these parameters. In our application, results are robust to alternative series 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 predictors 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. However, our object of interest is the ACT test score distribution, and it is not obvious that the empirical distribution of the predicted ACT scores converges at a parametric rate under Ahn and Powell's or Newey's assumptions.

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 predictors. There exist parametric and semiparametric sample selection models that relax this assumption but they have seldom been applied in practice (Donald, 1995; Chen and Khan, 2003).

C Alternative Evaluation Criteria

In the body of the paper we evaluate selection correction methods by running selection-corrected regressions of pre-policy ACT scores on a vector of predictors, predicting the mean ACT score, and comparing this to the mean ACT score in the reference distribution based on the complete post-policy ACT scores. In this appendix we consider three more evaluation criteria, all of which yield similar findings.

First, we evaluate the selection correction methods on how close the parameter estimates from the pre-policy selection-corrected regression of partly missing ACT scores on predictors are to the post-policy regression of complete ACT scores on parameters. Most theoretical papers on selection correction focus on this criterion. They try to correct the estimator of a specific parameter or vector of parameters for selection bias. Correction methods' performance may be very different with respect to prediction and parameter estimation.

In column 1 of Appendix Tables 6, 7, and 8 we show the parameter estimates from regressing

post-policy ACT scores on each of the three vectors of predictors (using inverse probability weights to equate the distribution of pre-policy predictors). In columns 2 to 9 we report the parameter estimates from regressing pre-policy ACT scores on each of the three vectors of predictors using our eight different selection correction models.[9] We evaluate the models' performance on parameter estimation against two criteria: the percentage of parameters whose signs are the same across the true and selection-corrected regressions, and the average squared difference between the parameters in the true and selection-corrected regressions (i.e. the squared bias of the estimates, averaged across the estimates). The general patterns are similar across the two criteria and are robust to weighting the squared biases by the variances of the corresponding predictors.

All methods perform better with richer predictors. The average squared bias is lowest for the rich set of predictors for seven out of eight models (all except the Heckman-IV model) and highest for the sparse set of predictors for all eight models. The squared bias averaged across all parameter estimates and across all eight models is 1.95 for the student demographic predictors, 0.67 when school- and district-level predictors are included, and 0.47 when student test scores are included. Similarly, adding richer predictors reduces the share of coefficient estimates with incorrect signs from 0.38 to 0.18. This pattern is entirely consistent with the pattern across predictions reported in Section 4. The only difference is that bias reduction from school- and district-level predictors is slightly larger for parameter estimation than for mean prediction.

The semiparametric models do not consistently outperform the more restrictive models. For the richest set of predictors, the squared bias is lowest for OLS (0.056), followed by the two semiparametric models with nonparametric first stages (0.075-0.082), Tobit (0.110), the two semiparametric models with series logit first stages (0.198-0.203), the Heckman-IV model (1.207), and the Heckman model (1.858). The pattern is similar for sign differences, though here Tobit and OLS both outperform any of the parametric or semiparametric two-stage models. There is a similar pattern with the two sparser sets of predictors. OLS always yields the lowest squared bias and fewest sign differences; the Heckman model without an instrument always yields the highest squared bias and the most sign differences. The semiparametric two-stage models generally outperform the parametric two-stage models but fail to outperform OLS and the Tobit model.

[9]We do not report parameter estimates for the missing data dummies. The general patterns are unaffected by including these in our analysis.

Appendix Table 6. The Relationship Between ACT Scores and Student Demographics

	Post-Policy	Pre-Policy, by Correction Method							
				Heckman		Newey		Powell	
	OLS	OLS	Tobit	No IV	With IV	Series Lgt	N.P.	Series Lgt	N.P.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student-Level									
Free Lunch	-2.866	-1.841	-2.361	2.180	0.449	-1.378	-1.367	-1.546	-1.247
	(0.105)	(0.104)	(0.141)	(1.825)	(0.573)	(0.588)	(0.162)	(0.680)	(0.172)
Female	0.298	-0.130	-0.213	-1.710	-1.025	-0.572	-0.331	-0.035	-0.292
	(0.036)	(0.034)	(0.043)	(0.702)	(0.232)	(0.162)	(0.050)	(0.247)	(0.058)
Black	-3.414	-4.102	-5.349	-4.087	-4.081	-3.836	-4.019	-3.330	-4.099
	(0.232)	(0.204)	(0.384)	(0.245)	(0.158)	(0.190)	(0.207)	(0.280)	(0.235)
Hispanic	-1.967	-1.818	-2.154	-0.443	-1.019	-1.495	-1.603	-1.212	-1.452
	(0.127)	(0.215)	(0.261)	(0.779)	(0.381)	(0.318)	(0.222)	(0.379)	(0.241)
Other	1.032	0.616	0.862	-1.295	-0.474	-0.355	0.412	-0.147	-0.155
	(0.307)	(0.290)	(0.319)	(0.978)	(0.342)	(0.364)	(0.264)	(0.451)	(0.268)
Inverse Mills Ratio				8.807	5.010				
				(4.025)	(1.256)				
Correction Term						1.629	-14.890		
						(1.709)	(6.973)		
Correction Term^2						-13.914	26.321		
						(8.024)	(12.913)		
Correction Term^3						-33.446	-13.058		
						(26.510)	(7.639)		
Correction Term^4						116.523			
						(70.223)			
Correction Term^5						183.034			
						(163.238)			
Correction Term^6						-434.349			
						(266.709)			
Correction Term^7						-360.897			
						(468.272)			
Correction Term^8						826.494			
						(495.019)			
Correction Term^9						204.032			
						(670.136)			
Correction Term^10						-744.410			
						(524.08)			
Correction Term^11						104.713			
						(379.809)			
Correction Term^12						234.721			
						(343.836)			
Correction Term^13						-83.986			
						(96.860)			
Summary Measures									
% with incorrect signs		0.2	0.2	0.6	0.6	0.4	0.2	0.4	0.4
Mean squared bias		0.380	0.865	7.537	3.270	1.059	0.705	0.764	1.023
Sample Size	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 3. The level of observation is the student. Each column is from a separate regression of ACT scores on the reported student-level demographics. Standard errors calculated using 500 bootstrap replications resampling schools.

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

	Post-Policy	Pre-Policy, by Correction Method							
				Heckman		Newey		Powell	
	OLS	OLS	Tobit	No IV	With IV	Series Lgt	N.P.	Series Lgt	N.P.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student-Level									
Free Lunch	-1.858	-1.078	-1.408	1.016	-0.405	-1.023	-1.137	-1.124	-1.136
	(0.072)	(0.073)	(0.100)	(0.581)	(0.377)	(0.118)	(0.100)	(0.118)	(0.090)
Female	0.288	-0.058	-0.124	-1.180	-0.419	-0.154	-0.089	-0.118	-0.055
	(0.036)	(0.033)	(0.042)	(0.318)	(0.207)	(0.057)	(0.042)	(0.069)	(0.048)
Black	-2.998	-3.370	-4.481	-3.592	-3.441	-3.324	-3.306	-3.299	-3.375
	(0.121)	(0.118)	(0.158)	(0.165)	(0.124)	(0.112)	(0.115)	(0.116)	(0.109)
Hispanic	-1.781	-1.566	-1.877	-0.876	-1.342	-1.524	-1.519	-1.532	-1.488
	(0.114)	(0.146)	(0.203)	(0.295)	(0.199)	(0.146)	(0.147)	(0.149)	(0.141)
Other	0.505	0.157	0.268	-0.844	-0.165	-0.084	0.041	-0.320	-0.104
	(0.197)	(0.193)	(0.209)	(0.337)	(0.244)	(0.180)	(0.187)	(0.167)	(0.139)
Inverse Mills Ratio				5.889	1.894				
				(1.661)	(1.069)				
Correction Term						-0.019	39.854		
						(0.157)	(30.019)		
Correction Term^2						0.041	-252.815		
						(0.092)	(189.779)		
Correction Term^3						0.003	752.899		
						(0.116)	(572.647)		
Correction Term^4						0.023	-1153.059		
						(0.053)	(890.813)		
Correction Term^5							871.799		
							(690.295)		
Correction Term^6							-255.603		
							(210.965)		
School-Level									
Pupil Teacher Ratio	0.001	-0.002	-0.005	0.002	-0.001	-0.002	-0.002	-0.002	-0.001
	(0.007)	(0.007)	(0.012)	(0.010)	(0.008)	(0.007)	(0.006)	(0.006)	(0.006)
Fraction Free Lunch	0.636	-0.582	-1.100	-0.727	-0.634	-0.486	-0.585	-0.365	-0.355
	(0.485)	(0.272)	(0.419)	(0.563)	(0.331)	(0.270)	(0.257)	(0.283)	(0.263)
Fraction Black	1.712	1.017	0.802	-0.140	0.644	0.814	0.892	0.619	0.835
	(0.445)	(0.771)	(1.236)	(1.645)	(1.007)	(0.657)	(0.670)	(0.577)	(0.570)
Number of 11th Graders	-0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average 8th Grade Score	1.938	2.338	2.904	-0.188	1.523	1.836	2.028	1.951	1.965
	(0.194)	(0.237)	(0.291)	(0.765)	(0.517)	(0.263)	(0.225)	(0.247)	(0.200)
Average 11th Grade Score	2.741	1.224	1.443	-0.624	0.628	1.066	1.141	1.004	1.126
	(0.185)	(0.197)	(0.237)	(0.506)	(0.356)	(0.193)	(0.186)	(0.169)	(0.145)
District-Level									
Pupil Teacher Ratio	-0.066	-0.020	-0.017	0.052	0.004	0.002	-0.002	0.012	-0.000
	(0.018)	(0.019)	(0.025)	(0.042)	(0.025)	(0.020)	(0.019)	(0.020)	(0.018)
Fraction Free Lunch	-0.554	0.300	0.980	0.906	0.499	0.236	0.370	0.182	0.057
	(0.457)	(0.346)	(0.537)	(0.767)	(0.440)	(0.347)	(0.333)	(0.371)	(0.338)
Fraction Black	1.510	0.864	1.428	-1.238	0.186	0.591	0.652	0.620	0.675
	(0.482)	(0.784)	(1.243)	(1.841)	(1.050)	(0.658)	(0.674)	(0.633)	(0.604)
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)
Suburb	-0.169	-0.418	-0.479	-0.488	-0.447	-0.430	-0.415	-0.401	-0.372
	(0.106)	(0.149)	(0.186)	(0.233)	(0.169)	(0.149)	(0.145)	(0.134)	(0.123)
Town	-0.177	0.023	0.038	-0.188	-0.052	0.079	0.080	0.078	0.166
	(0.125)	(0.168)	(0.206)	(0.289)	(0.201)	(0.169)	(0.168)	(0.161)	(0.145)
Rural	-0.210	-0.201	-0.172	-0.498	-0.303	-0.183	-0.157	-0.162	-0.102
	(0.114)	(0.156)	(0.194)	(0.247)	(0.180)	(0.155)	(0.150)	(0.150)	(0.132)
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)
Local Unemployment Rate	-0.009	-0.032	-0.051	0.006	-0.020	-0.030	-0.032	-0.025	-0.021
	(0.014)	(0.015)	(0.020)	(0.036)	(0.021)	(0.017)	(0.016)	(0.017)	(0.015)
Summary Measures									
% with incorrect signs		0.3	0.3	0.6	0.35	0.4	0.3	0.4	0.4
Mean squared bias		0.336	0.580	2.221	0.690	0.395	0.375	0.411	0.342
Sample Size	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 3. The level of observation is the student. Each column is from a separate regression of ACT scores on the reported student-, school- and district-level covariates. Missing value indicators also included but coefficients not reported. Standard errors calculated using 500 bootstrap replications resampling schools.

Appendix Table 8. The Relationship Between ACT Scores, Demographics, and Achievement

	Post-Policy	Pre-Policy, by Correction Method							
				Heckman		Newey		Powell	
	OLS	OLS	Tobit	No IV	With IV	Series Lgt	N.P.	Series Lgt	N.P.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student-Level									
Free Lunch	-0.383	-0.254	-0.317	1.444	1.086	0.141	-0.107	0.138	-0.102
	(0.027)	(0.045)	(0.062)	(0.109)	(0.146)	(0.070)	(0.064)	(0.068)	(0.067)
Female	0.505	0.027	0.076	-1.091	-0.856	-0.288	-0.106	-0.305	-0.117
	(0.023)	(0.025)	(0.032)	(0.078)	(0.098)	(0.044)	(0.031)	(0.046)	(0.031)
Black	-0.696	-1.295	-1.766	-3.106	-2.723	-1.569	-1.279	-1.581	-1.238
	(0.059)	(0.080)	(0.111)	(0.188)	(0.205)	(0.091)	(0.080)	(0.095)	(0.078)
Hispanic	-0.589	-0.727	-0.886	-0.753	-0.741	-0.745	-0.525	-0.744	-0.467
	(0.061)	(0.091)	(0.139)	(0.230)	(0.192)	(0.106)	(0.098)	(0.118)	(0.106)
Other	0.394	0.209	0.224	-1.384	-1.048	-0.127	0.081	-0.112	0.048
	(0.090)	(0.111)	(0.108)	(0.245)	(0.232)	(0.131)	(0.120)	(0.131)	(0.114)
8th Grade Score	1.639	1.833	2.155	-0.135	0.276	1.237	1.668	1.267	1.669
	(0.037)	(0.031)	(0.038)	(0.100)	(0.159)	(0.063)	(0.034)	(0.064)	(0.031)
11th Grade Score	3.048	2.616	3.238	0.109	0.634	1.940	2.402	1.952	2.397
	(0.024)	(0.035)	(0.044)	(0.132)	(0.203)	(0.076)	(0.045)	(0.075)	(0.042)
Inverse Mills Ratio				6.513	5.147				
				(0.333)	(0.521)				
Correction Term						0.312	-3.051		
						(0.098)	(6.903)		
Correction Term^2						0.324	12.537		
						(0.067)	(19.153)		
Correction Term^3						0.029	-23.072		
						(0.068)	(22.289)		
Correction Term^4						-0.012	15.245		
						(0.028)	(9.257)		
Correction Term^5						-0.025			
						(0.021)			
Correction Term^6						0.006			
						(0.005)			
Correction Term^7						0.002			
						(0.002)			
Correction Term^8						-0.001			
						(0.001)			
Correction Term^9						0.000			
						(0.000)			
School-Level									
Pupil Teacher Ratio	-0.006	-0.003	-0.008	0.002	0.001	-0.002	-0.002	-0.002	-0.002
	(0.007)	(0.005)	(0.010)	(0.010)	(0.009)	(0.005)	(0.005)	(0.005)	(0.004)
Fraction Free Lunch	-0.536	-0.449	-0.827	-0.540	-0.535	-0.367	-0.391	-0.503	-0.363
	(0.437)	(0.297)	(0.429)	(0.605)	(0.501)	(0.297)	(0.294)	(0.275)	(0.287)
Fraction Black	-0.253	-0.273	-0.644	-0.442	-0.413	-0.451	-0.489	-0.578	-0.369
	(0.474)	(0.578)	(0.916)	(1.617)	(1.348)	(0.504)	(0.505)	(0.491)	(0.463)
Number of 11th Graders	0.000	0.001	0.001	0.000	0.000	0.001	0.001	0.000	0.001
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average 8th Grade Score	0.907	1.085	1.198	-1.248	-0.771	0.165	0.595	0.137	0.578
	(0.192)	(0.181)	(0.214)	(0.363)	(0.340)	(0.178)	(0.173)	(0.171)	(0.166)
Average 11th Grade Score	-0.231	-0.206	-0.187	-0.525	-0.462	-0.131	-0.267	-0.094	-0.261
	(0.176)	(0.154)	(0.180)	(0.291)	(0.243)	(0.142)	(0.141)	(0.136)	(0.129)
District-Level									
Pupil Teacher Ratio	-0.044	-0.039	-0.040	0.061	0.044	-0.001	-0.015	-0.001	-0.012
	(0.017)	(0.017)	(0.021)	(0.037)	(0.032)	(0.019)	(0.017)	(0.018)	(0.016)
Fraction Free Lunch	-0.272	-0.758	-0.534	0.611	0.335	-0.325	-0.344	-0.281	-0.391
	(0.448)	(0.336)	(0.464)	(0.746)	(0.622)	(0.335)	(0.321)	(0.326)	(0.305)
Fraction Black	1.150	1.260	1.605	-1.737	-1.111	0.523	0.805	0.673	0.634
	(0.499)	(0.629)	(0.973)	(1.748)	(1.448)	(0.557)	(0.549)	(0.535)	(0.511)
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)
Suburb	-0.165	-0.356	-0.381	-0.394	-0.407	-0.350	-0.351	-0.333	-0.329
	(0.101)	(0.123)	(0.149)	(0.223)	(0.192)	(0.128)	(0.118)	(0.117)	(0.110)
Town	-0.174	-0.072	-0.098	-0.339	-0.310	-0.147	-0.090	-0.146	-0.064
	(0.125)	(0.142)	(0.176)	(0.268)	(0.226)	(0.144)	(0.133)	(0.131)	(0.120)
Rural	-0.121	-0.224	-0.196	-0.606	-0.550	-0.338	-0.205	-0.320	-0.202
	(0.112)	(0.134)	(0.164)	(0.239)	(0.202)	(0.140)	(0.128)	(0.130)	(0.115)
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)
Local Unemployment Rate	-0.008	-0.039	-0.058	0.023	0.009	-0.021	-0.028	-0.021	-0.025
	(0.015)	(0.014)	(0.018)	(0.036)	(0.030)	(0.016)	(0.014)	(0.015)	(0.014)
Summary Measures									
% with incorrect signs		0.045	0.045	0.455	0.409	0.182	0.091	0.136	0.091
Mean squared bias		0.056	0.110	1.858	1.207	0.203	0.075	0.198	0.082
Sample Size	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 3. The level of observation is the student. Each column is from a separate regression of ACT scores on the reported student-, school- and district-level covariates. Missing value indicators also included but coefficients not reported. Standard errors calculated using 500 bootstrap replications resampling schools.

Applied researchers are often interested in the full selection-corrected outcome distribution or in summary statistics other than the mean. Researchers working with test scores may be interested in the share of students who score above some threshold. We therefore use two additional evaluation criteria:

1. The squared difference between selection-corrected pre-policy ACT score distribution to the reference distribution, averaged over percentiles 1, 2, ..., 99.

2. The difference between the selection-corrected share of pre-policy students scoring above 19 (the ACT's "college readiness" threshold) and the share in the reference distribution.

To construct these evaluation criteria, we cannot simply use the predicted values $\hat{ACT}_i = X_i\hat{\beta}$ from the selection-corrected regression of ACT scores on predictors. The distribution of $\hat{ACT}_i$ is not comparable to the distribution of ACT_i or ACT_i^* because the former omits the variance of ϵ_i. We therefore predict the fitted residual $\hat{\epsilon}_j = ACT_j - \hat{ACT}_j$ for each student j who takes the ACT in the pre-policy period, and construct $\tilde{ACT}_i = \hat{ACT}_i + \hat{\epsilon}_{j\neq i}$, adding to each student's predicted ACT score one of the fitted residuals from another randomly chosen student. This generates a distribution of predicted ACT scores with variance comparable to the latent distribution. We repeat this residual-adding process 1000 times and average over these 1000 repetitions to obtain a predicted distribution $\hat{F}\left(\tilde{ACT}_i\right)$ and compare this to the reference distribution.

We estimate the variance of these two differences using a nonparametric cluster bootstrap, clustering at the school level to account for correlated unobserved school-level characteristics.[10] We use 500 bootstrap replications, each containing 100 iterations of the residual-adding process.

The selection correction methods we evaluate are not designed to predict the full outcome distribution, so this part of the evaluation should be interpreted with caution. To formalize this caution, note that the distribution of latent ACT scores $\tilde{F}_{ACT^*}(.)$ can be evaluated at any point a as $F_{ACT^*}(a) = \mathbb{E}_X\left[F_{\epsilon|X}\left(a - X\beta\right)\right]$, where the outer expectation is taken over the distribution of the predictors and the inner distribution is for the error distribution conditional on the predictors. Parameter-oriented selection corrections aim to identify only (elements of) β. Our approach entails identification of both β and $F_{\epsilon|X}(.)$. The residual-adding procedure

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

assumes that the error distribution does not vary with X or with ACT-taking: $F_{ACT*}(a) = \mathbb{E}_X \left[F_{\epsilon, D=1} \left(a - X\beta \right) \right]$. This is a strong assumption. In particular, the assumptions of the Tobit Type 1 and Heckman models imply that the error distribution should differ between ACT-takers and non-takers. The accurate predictions reported in Section 4 suggest that with sufficiently rich predictors, this assumption is innocuous.

We could instead adopt a parametric approach to identification of $F_{\epsilon|X}$. Specifically, the Tobit and Heckman models both assume that the errors have a homoskedastic normal distribution with zero mean. Both models recover estimates of the variance of this distribution, $\hat{\sigma}_\epsilon^2$. We could use this estimate to sample values of $\hat{\epsilon}_i$ from a $\mathcal{N}\left(0, \hat{\sigma}_\epsilon^2\right)$ distribution instead of sampling from the empirical distribution $\hat{F}_{\epsilon|TAKE=1}(.)$. This would introduce another difference between the parametric (Tobit and Heckman) and semiparametric (Newey and Powell) selection correction models.

Acknowledging this caveat, what do we learn from evaluating the selection correction methods on these two additional criteria? We show in Appendix Table 9 the difference between the selection-corrected pre-policy score distribution and the reference score distribution in the share of students scoring above 19 and averaged over the percentiles. This table is analogous to Table 3 in the main paper. We summarize the squared bias and variance of each comparison in Appendix Figure 6, which is analogous to Figure 2 in the main paper. We also display the observed pre- and post-policy scores and the selection-corrected scores in Appendix Figures 4 and 5.

The share of college-ready students in the reference distribution is 0.45. The share in the uncorrected pre-policy distribution is 0.59 percentage points higher. Using OLS with student demographics to predict the missing scores reduces prediction to 0.55. Using other selection corrections to predict the missing values predicts 0.54 to 0.56, which continues to overstate the share by 9 to 11 percentage points. Adding school- and district-level predictors reduces this overstatement to 6 to 7 percentage points and adding student-level test scores reduces this overstatement to 0 to 3 percentage points. As with the mean, richer predictors largely eliminate the difference between the selection-corrected and reference statistics; changing the selection correction method has little effect.

The mean squared difference between the percentiles of the raw pre-policy distribution and reference distribution is 1.69. Predicting missing scores using OLS and the basic student demographics reduces this to 1.32. Other selection correction methods yield differences between 1.27

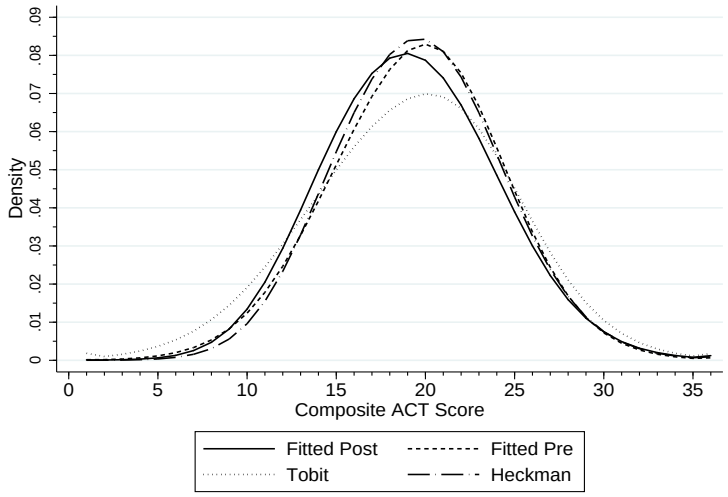
Appendix Table 9. Fraction College-Ready and Quantile Differences by Correction Method and Predictor Set

	Fraction ACT*>=20			Quantile Differences		
	Student Demographics	...Plus School-Level Covs.	...Plus Student Test Scores	Student Demographics	...Plus School-Level Covs.	...Plus Student Test Scores
	(1)	(2)	(3)	(4)	(5)	(6)
Post-Policy ("Truth")						
Raw	0.440	0.440	0.440	-	-	-
DFL	0.482	0.482	0.482	-	-	-
OLS	0.451	0.468	0.468	0.300	0.325	0.324
	(0.010)	(0.011)	(0.011)	(0.028)	(0.022)	(0.016)
Pre-Policy (Biased)						
Raw	0.588	0.588	0.588	1.687	1.687	1.687
OLS	0.554	0.532	0.469	1.323	1.058	0.623
	(0.008)	(0.008)	(0.008)	(0.148)	(0.128)	(0.033)
Tobit	0.559	0.536	0.460	1.276	1.053	1.382
	(0.008)	(0.007)	(0.007)	(0.151)	(0.130)	(0.108)
Heckman, No Instrument	0.554	0.533	0.463	1.334	1.078	0.444
	(0.008)	(0.008)	(0.008)	(0.148)	(0.129)	(0.031)
Heckman, With Instrument	0.541	0.532	0.463	1.302	1.062	0.453
	(0.008)	(0.008)	(0.008)	(0.149)	(0.128)	(0.031)
Newey, Series Logit	0.540	0.532	0.460	1.292	1.073	0.419
	(0.008)	(0.008)	(0.008)	(0.148)	(0.130)	(0.030)
Newey, Nonparametric	0.541	0.532	0.463	1.307	1.070	0.525
	(0.008)	(0.008)	(0.008)	(0.149)	(0.129)	(0.033)
Powell, Series Logit	0.546	0.535	0.497	1.265	1.127	1.084
	(0.009)	(0.008)	(0.010)	(0.184)	(0.132)	(0.081)
Powell, Nonparametric	0.554	0.533	0.479	1.400	1.097	0.721
	(0.008)	(0.008)	(0.009)	(0.152)	(0.129)	(0.040)

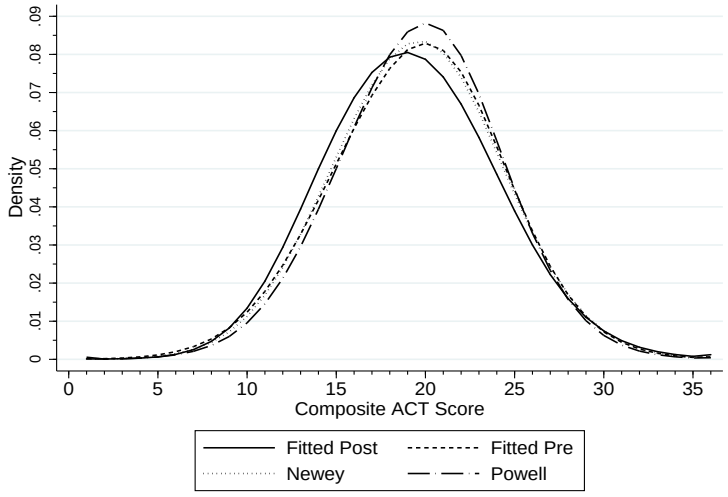
Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. For columns 1-3, the first and fourth rows report the raw fraction scoring greater than or equal to 20 post- and pre-policy, respectively. The second row reports that fraction from the DFL-weighted post-policy score distribution. All other rows report the predicted fraction scoring greater than or equal to 20 after implementing the regression or correction type noted in the row header. The predicted ACT score is calcuated for ACT-takers and non-takers. Columns 4-6 report quantile differences between the predicted score distribution from the regression or correction method noted in the row header and the post-policy DFL-corrected score distribution. Standard errors calculated using 500 bootstrap replications resampling schools.

Appendix Figure IV: Comparing the Performance of Sample Selection Corrections

(a) Parametric Corrections



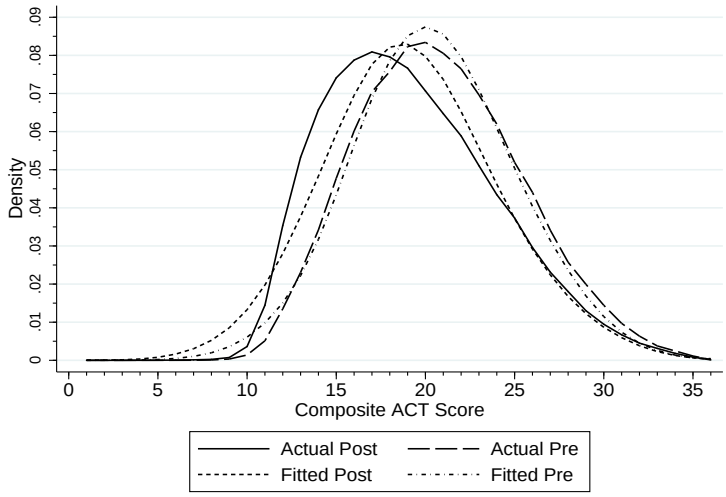
(b) Semiparametric Corrections



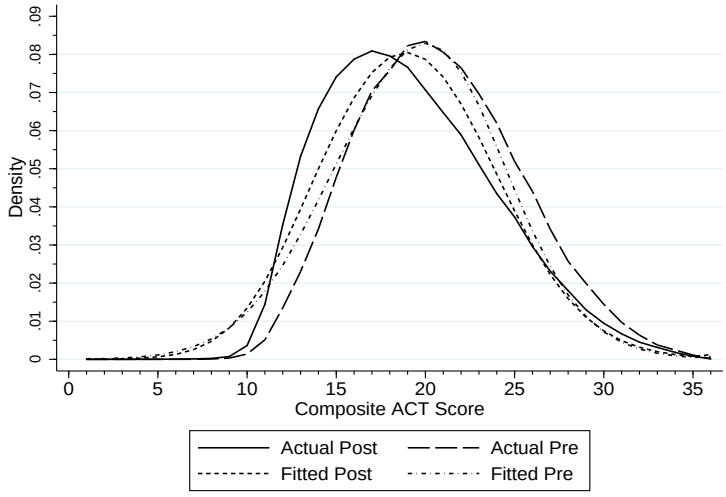
Notes: Figure shows pre- and post-policy fitted values from regressions of ACT scores on student-, school-, and district-level demographics, and 8th and 11th grade test scores. The post-policy regressions are DFL-weighted. The pre-policy fitted values are predicted out of sample to all students. Draws from the distribution of residuals are added to all fitted values. Tobit, Heckman, Newey, and Powell are several selection corrections estimated using the pre-policy sample. The semiparametric corrections use the nonparametric first stage. 95% confidence intervals are tiny and omitted for readability.

Appendix Figure V: Observed and Predicted ACT Scores Pre- and Post-Policy

(a) Predicting ACT Scores Using Basic Student and School Characteristics



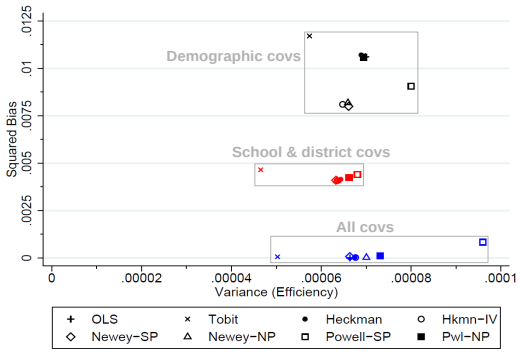
(b) Predicting ACT Scores Using Student 8th and 11th Grade Test Scores



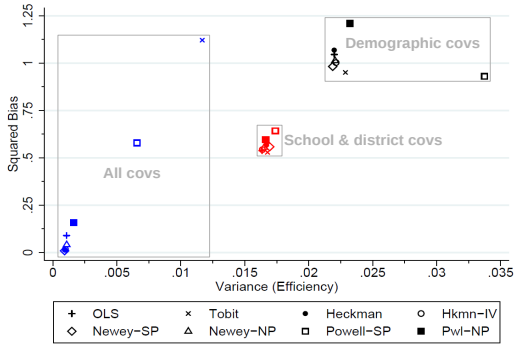
Notes: Figure (a) shows pre- and post-policy raw ACT scores and fitted values from regressions of ACT scores on student-level demographics and school- and district-level demographics and test scores. The post-policy regressions are DFL-weighted. The pre-policy fitted values are predicted out of sample to all students. Draws from the distribution of residuals are added to all fitted values. Figure (b) adds student-level 8th and 11th grade test scores to the prediction equations. 95% confidence intervals are tiny and omitted for readability.

Appendix Figure VI: MSE Comparison Across Correction Methods and Covariates

(a) Fraction College-Ready



(b) Test Score Distribution



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

and 1.40 using the basic student demographics as predictors. With the richer set of predictors, the mean squared difference falls to between 0.62 and 0.138. The Tobit model and Powell model with series logit first stage perform particularly poorly. All other methods deliver lower prediction error with richer covariates.

We conclude that for all four evaluation criteria, based both on prediction and parameter estimation, the gains from using less restrictive econometric methods are small relative to the gains from seeking richer or more disaggregated data. We find the same pattern when we repeat the subgroup analysis from Section 4.3 for these two new evaluation criteria (Appendix Table 10 and Appendix Figures 7, 8, and 9).

D Robustness Checks

In this section, we establish that our findings are robust to several changes in our methods: using a different reference distribution, changing the specification of the ACT regression model, and comparing different pre- and post-policy cohorts.

In the primary analyses, we use the post-policy ACT scores as the reference distribution, after adjusting for cross-cohort differences in the distribution of observed characteristics using inverse probability weights, and predicting scores for the 1.5% of post-policy students who do not take the ACT. We summarize the results using the unweighted post-policy distribution in Appendix Figure 10. This is analogous to Figure 2 and Appendix Figure 6, which use the weighted reference distribution. We display the subgroup means relative to the unweighted post-policy means in Appendix Figure 11. This is analogous to Appendix Figures 2 and 7, which use the weighted reference distribution. There are no substantial differences between the analysis that uses the weighted and unweighted reference distributions.

Our findings are robust to five changes in the ACT regression model. First, we estimate the model with a complete set of interactions between the predictors and squares of all continuous predictors in both the first and second stages (Appendix Table 11, panel 1).[11] The predictions are more accurate for most models with the rich set of predictors and essentially identical for all models with the two sparser sets of predictors. There remains no evidence that the more flexible methods outperform those with more restrictive assumptions.

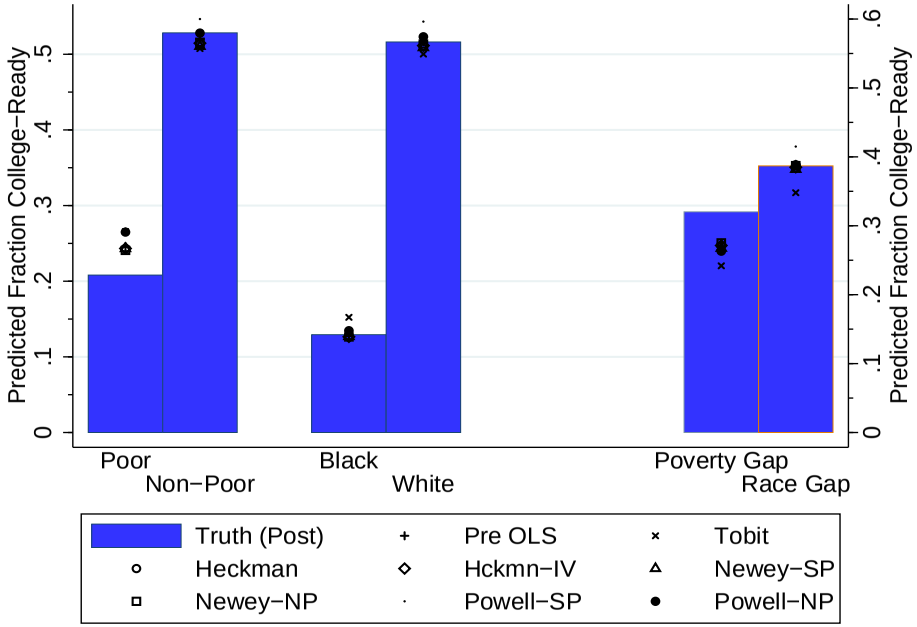
[11]The ACT-taking equations of the series logit model and nonparametric model already incorporate these interactions explicitly or implicitly. So in these cases we are simply establishing robustness to changes in the ACT score model.

Appendix Table 10. Race and Poverty Gaps in the Fraction College-Ready by Correction

	Black	White	Gap	Poor	Non-Poor	Gap
	(1)	(2)	(3)	(4)	(5)	(6)
Post-Policy						
Raw	0.124	0.506	0.383	0.224	0.522	0.298
DFL	0.156	0.532	0.376	0.232	0.545	0.313
OLS	0.129	0.516	0.387	0.208	0.528	0.320
	(0.024)	(0.009)	(0.025)	(0.007)	(0.010)	(0.010)
Pre-Policy						
Raw	0.201	0.647	0.446	0.350	0.628	0.278
OLS	0.127	0.516	0.389	0.246	0.520	0.274
	(0.017)	(0.007)	(0.017)	(0.008)	(0.008)	(0.009)
Tobit	0.152	0.500	0.348	0.266	0.508	0.242
	(0.017)	(0.006)	(0.017)	(0.008)	(0.007)	(0.009)
Heckman	0.127	0.511	0.385	0.243	0.515	0.271
(with IV)	(0.017)	(0.007)	(0.018)	(0.008)	(0.008)	(0.009)
Newey -	0.127	0.509	0.382	0.243	0.511	0.267
Series Logit	(0.017)	(0.007)	(0.018)	(0.009)	(0.008)	(0.010)
Newey -	0.126	0.514	0.387	0.241	0.516	0.275
Nonparametric	(0.017)	(0.007)	(0.017)	(0.008)	(0.008)	(0.009)
Powell -	0.128	0.543	0.415	0.269	0.547	0.277
Series Logit	(0.017)	(0.010)	(0.019)	(0.011)	(0.009)	(0.013)
Powell -	0.135	0.523	0.389	0.265	0.528	0.263
Nonparametric	(0.017)	(0.007)	(0.018)	(0.010)	(0.008)	(0.011)

Notes: The sample is as in Table 3. Table reports the fraction of the predicted ACT scores that are greater than or equal to 20 from regressions of ACT scores on the full set of covariates, including student-level 8th and 11th grade test scores. The predicted ACT score is calcuated for ACT-takers and non-takers. Poverty status is proxied for using free or reduced-price lunch receipt measured during 11th grade. Standard errors calculated using 500 bootstrap replications resampling schools.

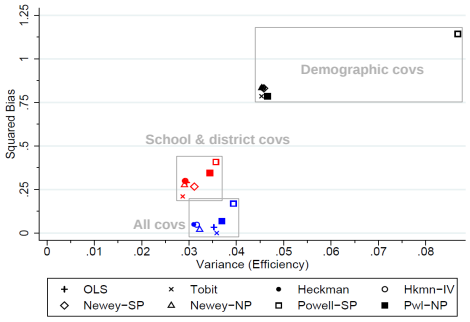
Appendix Figure VII: Predicted Fraction College Ready and Gaps



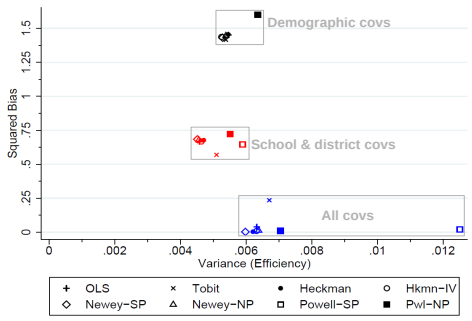
Notes: The leftmost four bars and markers show the "true" and predicted fraction scoring college ready across correction methods by poverty status and race. The rightmost two bars and markers show the "true" and predicted gaps in these measures across correction methods by poverty status and race. All specifications include basic student demographics, school- and district-level covariates, and student 8th and 11th grade test scores.

Appendix Figure VIII: MSE Comparison by Race

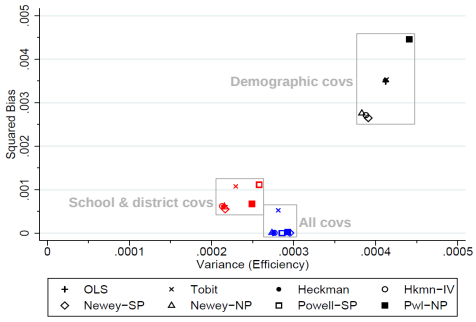
(a) Mean ACT Score - Black



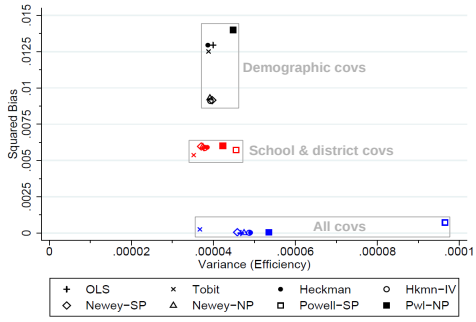
(b) Mean ACT Score - White



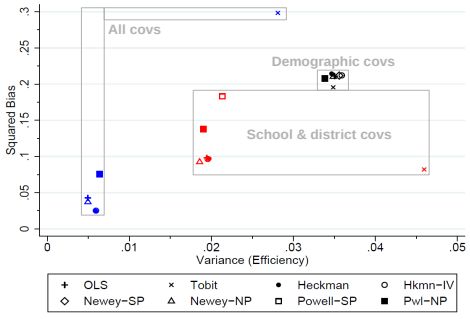
(c) Fraction College-Ready - Black



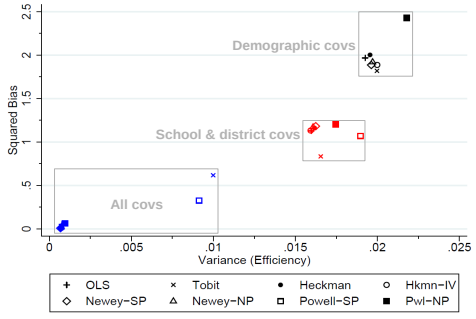
(d) Fraction College-Ready - White



(e) Test Score Distribution - Black



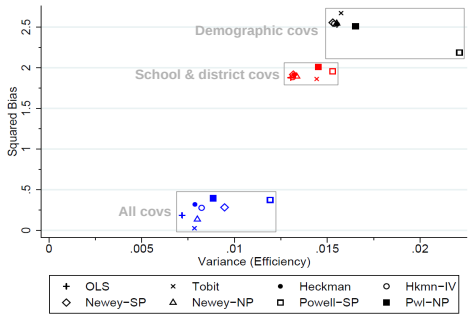
(f) Test Score Distribution - White



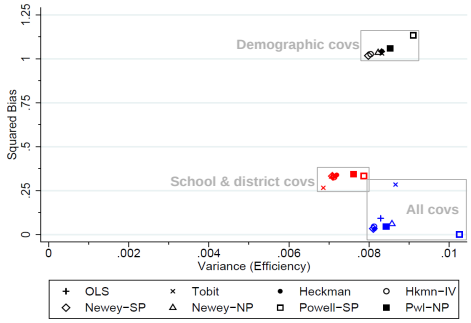
Notes: Figure shows the mean squared error of each combination of correction method and covariate set by race. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the difference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

Appendix Figure IX: MSE Comparison by Poverty Status

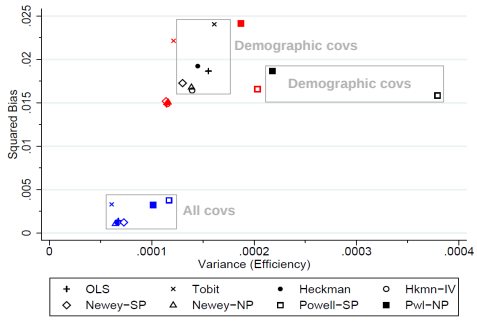
(a) Mean ACT Score - Poor



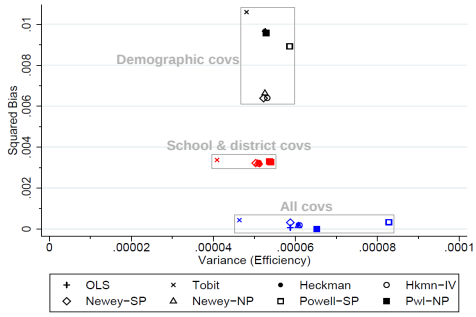
(b) Mean ACT Score - Non-Poor



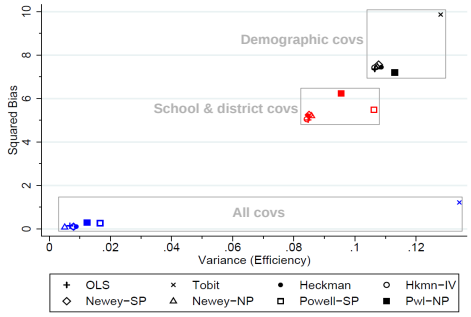
(c) Fraction College-Ready - Poor



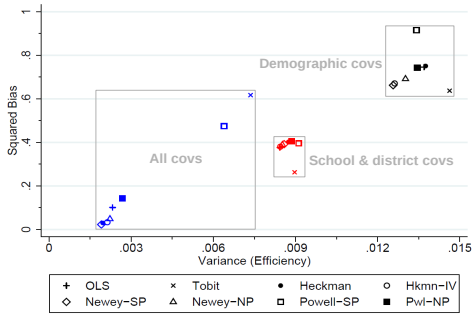
(d) Fraction College-Ready - Non-Poor



(e) Test Score Distribution - Poor



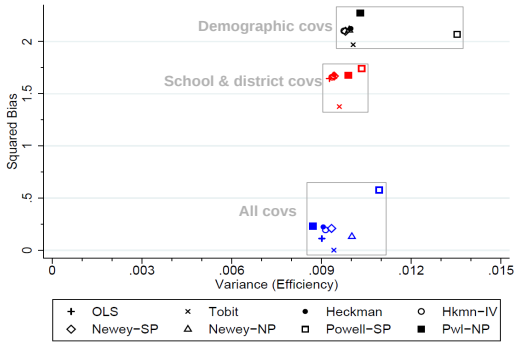
(f) Test Score Distribution - Non-Poor



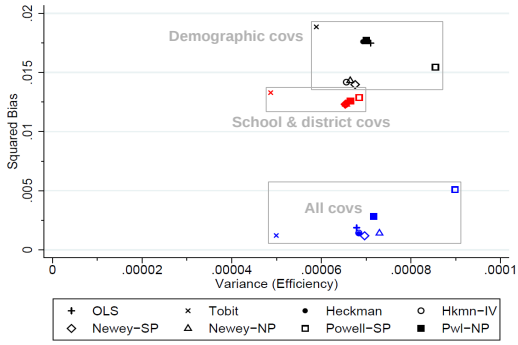
Notes: Figure shows the mean squared error of each combination of correction method and covariate set by poverty status. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the difference between the statistic predicted by a) the correction method applied to the pre-policy data and b) the post-policy, DFL-weighted, fitted distribution. Markers with very large variance or squared bias excluded for readability.

Appendix Figure X: MSE Comparison Using Post-Policy Distribution W/Out DFL Weights

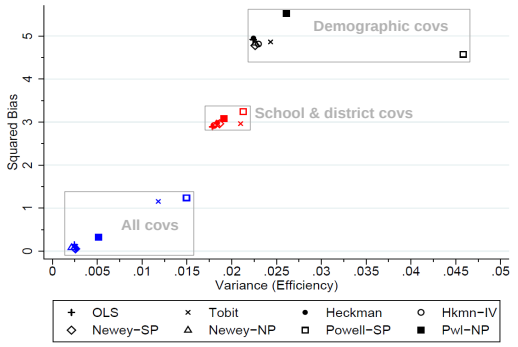
(a) Mean ACT Score



(b) Fraction College-Ready



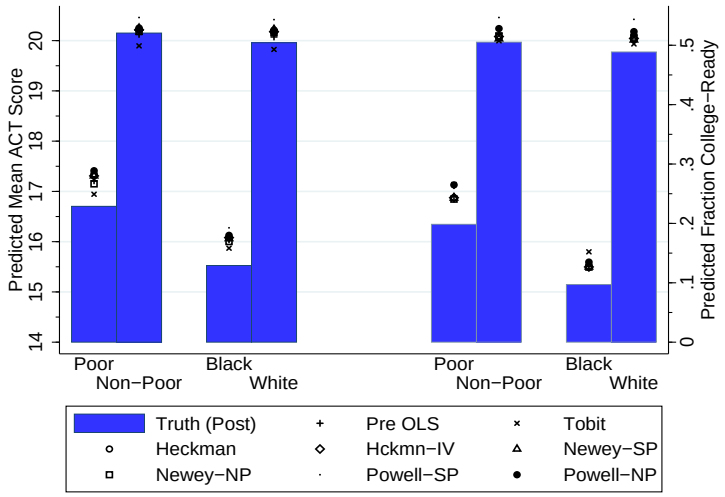
(c) Test Score Distribution



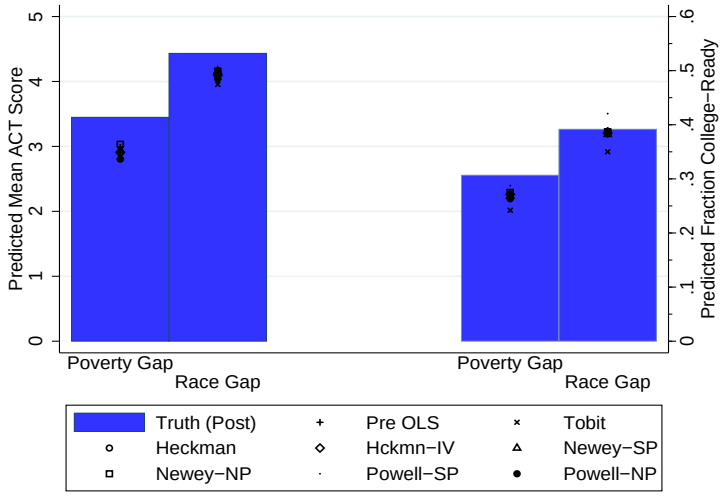
Notes: Figure shows the mean squared error of each combination of correction method and covariate set estimated without DFL weights. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the difference between the statistic predicted by a) the correction method applied to the pre-policy data and b) the post-policy fitted distribution without DFL weights.

Appendix Figure XI: Score Gaps Compared to Post-Policy Distribution W/Out DFL Weights

(a) Predicted Mean ACT Score and Fraction College-Ready



(b) Predicted Poverty and Race Gaps



Notes: Figure (a) shows the "true" (using the fitted post-policy distribution without the DFL weights) and predicted mean latent ACT score and fraction college-ready across correction methods by poverty status and race. Figure (b) shows the "true" (using the fitted post-policy distribution without the DFL weights) and predicted gaps in these measures across correction methods by poverty status and race. All specifications include basic student demographics, school- and district-level covariates, and student 8th and 11th grade test scores.

Appendix Table 11. Specification Checks for Individual-Level Corrections

	Post-Policy ("Truth")			Pre-Policy (Biased)		Pre-Policy, by Correction Method						
							Heckman		Newey		Powell	
	Raw	DFL	OLS	Raw	OLS	Tobit	No IV	With IV	Series Lgt	N.P.	Series Lgt	N.P.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Spec Check 1: Include Interactions and Squared Terms												
X = Student Demographics												
E[ACT*]	19.25	19.73	19.56	20.86	20.67	20.63	20.67	20.66	20.66	20.67	20.65	20.71
Fraction ACT*>=20	0.440	0.482	0.450	0.588	0.553	0.560	0.553	0.540	0.540	0.542	0.544	0.554
Quantile Differences	0.000	0.000	0.295	1.687	1.310	1.284	1.310	1.292	1.291	1.303	1.249	1.362
X = …Plus School-Level Covs												
E[ACT*]	19.25	19.73	19.76	20.86	20.48	20.38	20.49	20.48	20.49	20.49	20.51	20.49
Fraction ACT*>=20	0.440	0.482	0.467	0.588	0.532	0.537	0.532	0.532	0.532	0.532	0.534	0.533
Quantile Differences	0.000	0.000	0.314	1.687	1.053	1.055	1.072	1.056	1.070	1.064	1.114	1.095
X = …Plus Student Test Scores												
E[ACT*]	19.25	19.73	19.68	20.86	19.64	19.35	19.61	19.63	19.64	19.60	19.94	19.83
Fraction ACT*>=20	0.440	0.482	0.463	0.588	0.457	0.459	0.458	0.458	0.452	0.452	0.476	0.471
Quantile Differences	0.000	0.000	0.251	1.687	0.428	0.894	0.438	0.430	0.383	0.406	0.921	0.735
Spec Check 2: Only Eighth Grade Student Test Scores												
E[ACT*]	19.25	19.73	19.73	20.86	19.96	19.69	19.94	19.95	19.98	19.69	20.31	20.54
Fraction ACT*>=20	0.440	0.482	0.472	0.588	0.488	0.489	0.488	0.488	0.486	0.464	0.520	0.532
Quantile Differences	0.000	0.000	0.424	1.687	0.694	0.957	0.627	0.619	0.574	0.466	1.243	1.317
Spec Check 3: No School- & District-Level Predictors												
E[ACT*]	19.25	19.73	19.62	20.86	19.55	19.27	19.70	19.66	19.63	19.55	19.66	19.75
Fraction ACT*>=20	0.440	0.482	0.463	0.588	0.472	0.464	0.465	0.466	0.466	0.463	0.480	0.483
Quantile Differences	0.000	0.000	0.342	1.687	0.634	1.331	0.439	0.446	0.462	0.521	0.636	0.766
Spec Check 4: No DFL Weights For Post-Policy Distribution (Quantile Differences)												
X = Student Demographics	0.000	-	0.246	2.939	2.460	2.452	2.468	2.441	2.427	2.444	2.375	2.571
X = …Plus School-Level Covs	0.000	-	0.325	2.939	2.029	2.047	2.057	2.034	2.046	2.042	2.124	2.080
X = …Plus Student Test Scores	0.000	-	0.325	2.939	0.710	1.399	0.577	0.573	0.544	0.607	1.435	0.894

Notes: Table presents estimated parameters as in Table 3, but with slightly altered specifications. Standard errors are nearly identical to Table 3 and omitted for readability. Specification check 1 includes interactions between the predictors as well as squares of any continuous variables. Specification check 2 mimics the "rich" specification including student test scores, but only includes eighth grade scores and excludes eleventh grade scores. Specification check 3 includes student demographics and student eighth and eleventh grade test scores, but excludes all school- and district-level predictors. Specification check 4 excludes the DFL-weights from the post-policy fitted distribution.

Second, we omit 11th grade social studies test scores from the "rich" set of predictors and use only 8th grade test scores, student demographics and school- and district-level predictors (Appendix Table 11, panel 2). The predictions are slightly less accurate for every model and every summary statistic, particularly for the mean squared difference between the predicted and reference distributions. But the predictions are still substantially more accurate than without using any student test scores and there remains no clear winner amongst the selection correction models.

Third, we estimate models with a different combination of predictors: student demographics and student test scores, but without school- and district-level predictors (Appendix Table 11, panel 3). The predictions are generally slightly less accurate than for the models including all predictors, but are always substantially more accurate than for the models that do not use any student test scores as predictors. Once again, the two-stage semiparametric models fail to outperform two-stage or one-stage parametric models.

Fourth, we calculate the mean squared quantile differences between the selection-corrected distributions and the reweighted and predicted reference distribution (Appendix Table 11, panel 4). The general pattern of results is unchanged, though here the parametric two-stage selection models slightly outperform the semiparametric two-stage selection models. Readers who wish to compare the mean ACT score and fraction college-ready generated by the correction models to the reference distribution in columns 1 or 3 can do so by directly comparing across columns in the first four panels.

Fifth, we implement a test of the assumption that the predictors and selection correction term are additively separable in the ACT score model. We regress ACT scores on the set of predictors and the inverse Mills ratio (for all three sets of predictors, with and without an instrument), generate the residuals from this regression, regress the residuals on a full set of interactions between the predictors and the inverse Mills ratio, and test the joint significance of all the interactions. We fail to reject the hypothesis that they are jointly zero ($F < 0.12$ for all tests). Additivity is a standard assumption in most of the literature on selection models and this assumption seems at least plausible in our setting.[12]

We also verify that our finding are robust to comparing different pairs of pre- and post-policy cohorts. Our primary analysis compares the 2005 cohort to the 2008 cohort, as the mandatory ACT policy was piloted in some schools in 2006 and not implemented in all schools in 2007. We

[12]See Arellano and Bonhomme (2017), Altonji, Ichimura, and Otsu (2012) and Manski (1990) for exceptions.

also compare the 2005 cohort to the 2007 cohort (Appendix Figure 12), the 2006 cohort to the 2007 cohort (Appendix Figure 13), and the 2006 cohort to the 2008 cohort (Appendix Figure 14). The main findings are unchanged across choices of cohorts: predictive accuracy is higher with richer predictors and does not vary substantially across selection correction methods.

E Group-level Correction Methods

Many researchers using test scores as a dependent variable observe only students who take the test and so cannot estimate individual probabilities of test-taking (Card and Payne, 2002; Rothstein, 2006). The individual-level corrections discussed thus far are infeasible in this case. We also evaluate the performance of selection correction models that use only group-level data. These methods are useful when researchers observe only the mean non-missing outcome and share non-missing outcomes for each group. For example, labor economists might observe regional employment rates and mean wages conditional on employment, while education economists might observe school-level test-taking rates and mean test scores conditional on taking. Building on Gronau (1974), Card and Payne (2002) adapt equation system (1) for use with aggregate data:

$$ACT^*_{ig} = X_{ig}\beta + \epsilon_{ig} \tag{7a}$$

$$TAKE^*_{ig} = W_g\mu + u_{ig} \tag{7b}$$

$$TAKE_{ig} = \begin{cases} 1 & \text{if} \quad TAKE^*_{ig} \geq 0 \\ 0 & \text{if} \quad TAKE^*_{ig} < 0 \end{cases} \tag{7c}$$

$$ACT_{ig} = \begin{cases} ACT^*_{ig} & \text{if} \quad TAKE^*_{ig} \geq 0 \\ . & \text{if} \quad TAKE^*_{ig} < 0 \end{cases} \tag{7d}$$

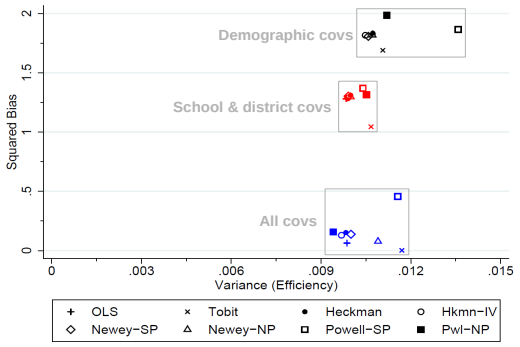
The key difference between systems (1) and (7) is the ACT-taking model. In this model we assume ACT-taking depends on a vector of group-level characteristics Wg and an individual error term u_{ig} that may be correlated with ϵ_{ig}. Card and Payne (2002) evaluate the observed test score equation at group means, yielding an estimating equation:

$$ACT_g = X_g\beta + h\left(\overline{TAKE}\right) + \overline{\epsilon}_g \tag{8}$$

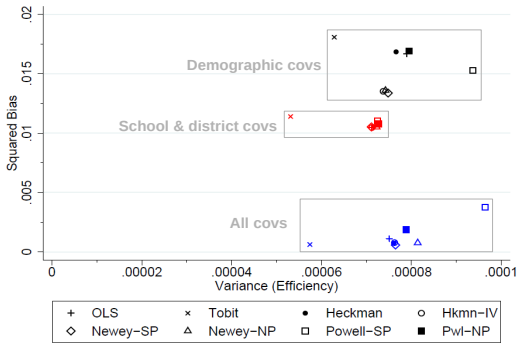
The selection correction term uses only the observed ACT-taking rate in each group, so we do not require that the predictors of ACT-taking, W_g, are observed.

Appendix Figure XII: MSE Comparison Using 2005 and 2007 Student Cohorts

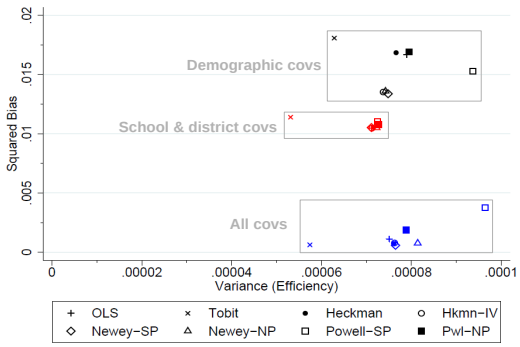
(a) Mean ACT Score



(b) Fraction College-Ready



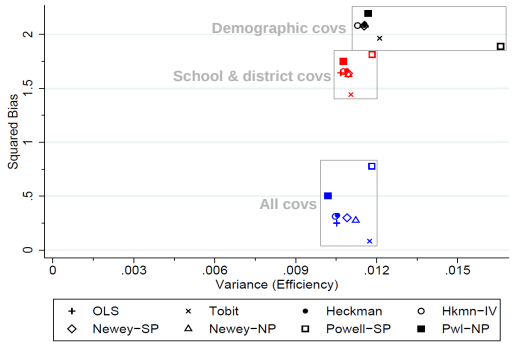
(c) Test Score Distribution



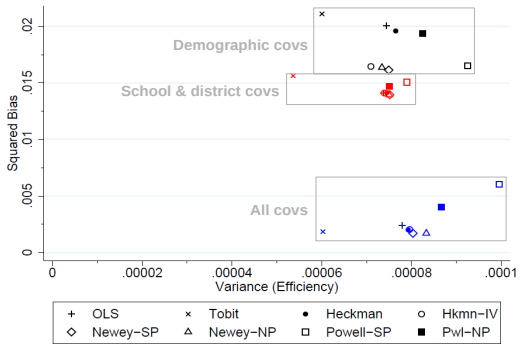
Notes: Figure shows the mean squared error of each combination of correction method and covariate set estimated using the 2005 and 2007 student cohorts, instead of the 2005 and 2008 cohorts. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the difference between the statistic predicted by a) the correction method applied to the pre-policy data and b) the post-policy, DFL-weighted, fitted distribution.

Appendix Figure XIII: MSE Comparison Using 2006 and 2007 Student Cohorts

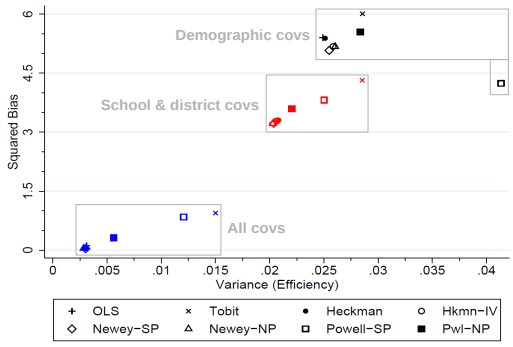
(a) Mean ACT Score



(b) Fraction College-Ready



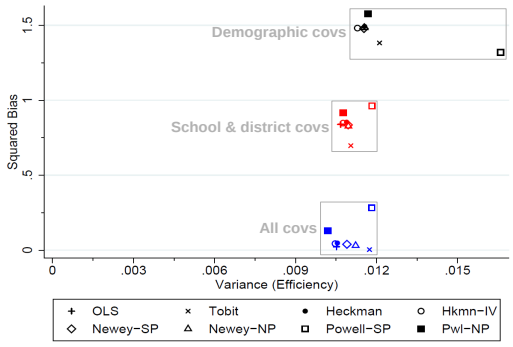
(c) Test Score Distribution



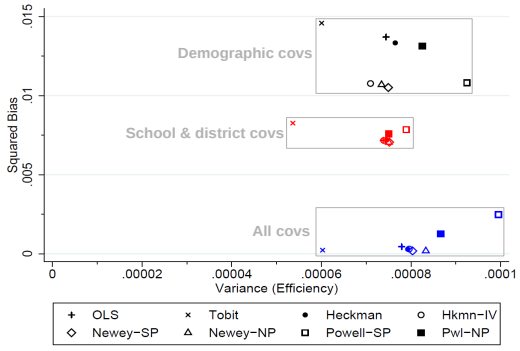
Notes: Figure shows the mean squared error of each combination of correction method and covariate set estimated using the 2006 and 2007 student cohorts, instead of the 2005 and 2008 cohorts. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the difference between the statistic predicted by a) the correction method applied to the pre-policy data and b) the post-policy, DFL-weighted, fitted distribution.

Appendix Figure XIV: MSE Comparison Using 2006 and 2008 Student Cohorts

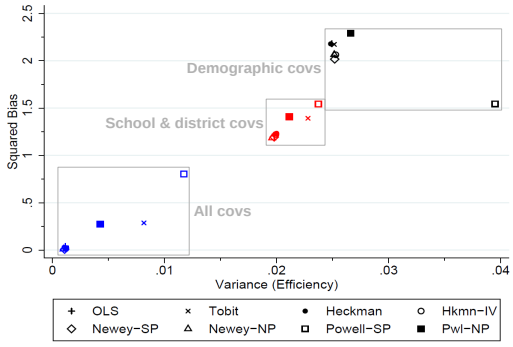
(a) Mean ACT Score



(b) Fraction College-Ready



(c) Test Score Distribution



Notes: Figure shows the mean squared error of each combination of correction method and covariate set estimated using the 2006 and 2008 student cohorts, instead of the 2005 and 2008 cohorts. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the difference between the statistic predicted by a) the correction method applied to the pre-policy data and b) the post-policy, DFL-weighted, fitted distribution.

This estimating equation is corrected for within-group selection but not for between-group selection, conditional on the observed ACT score predictors X_{ig}. Within-group selection occurs if individual ACT-taking covaries with individual deviations from mean latent ACT scores of the group, $\text{cov}\left(\epsilon_{ig} - \overline{\epsilon}_g, u_{ig} - \overline{u}_g\right) \neq 0$. Between-group selection occurs if the group ACT-taking rate covaries with the group mean latent ACT score, $\text{cov}\left(\overline{\epsilon}_{ig}, \overline{TAKE}_g\right) \neq 0$. As an example, assume groups are schools. The group-level model (8) is corrected for within-school selection, which could occur if individual students with higher latent scores are more likely to take the ACT than students in the same school with lower latent scores. But model (8) is not corrected for between-school selection, which could occur if "good" schools have high mean latent scores and high ACT-taking rates. This means that the level of aggregation is important. With larger groups, more of the selection is within-group and is addressed by the selection correction.[13] However, the group mean predictors X_g are less informative in larger groups. So using larger, more aggregated groups relies more on the correction model and less on the data.

The functional form of the selection correction term depends on the assumed distribution of the unobserved factors influencing ACT scores and ACT-taking, ϵ_{ig} and u_{ig}. If these are jointly normally distributed, then the selection correction term equals the inverse Mills ratio evaluated at the group mean ACT-taking rate (Card and Payne, 2002). We estimate equation (8) using a variety of functional forms for the selection correction term, including a polynomial in $\overline{TAKE}_g$, following the strategy in Newey (2009).[14]

Clark, Rothstein, and Whitmore Schanzenbach (2009) use this approach to study selection into ACT-taking in Illinois. They observe no data on non-takers (neither ACT scores nor lagged test scores and demographic characteristics). They therefore use only group-level methods and consider only parametric correction models based on joint normality assumptions. The study uses the shift from voluntary to mandatory ACT-taking in Illinois in 2002 as an instrument in these models. They conclude that this correction allows a reasonable approximation to the latent distribution of ACT scores.

We estimate group-level selection models of the form of equation (8) using pre-policy data, generate the predicted distribution of group mean ACT scores, and compare this to the dis-

[13] As the group size approaches one, the correction term approaches a constant.

[14] We estimate equation (8) using weighted least squares, where the weights equal the number of students in each group. We construct the predicted distribution of school mean ACT scores using 1000 replications of the same residual-adding process described in Section 2.3. We construct the standard errors using 500 replications of a nonparametric bootstrap, each containing 1000 residual-adding iterations.

tribution of group mean ACT scores in the post-policy period. We also estimate models that use the group-level fraction of ACT-taking students who score at or above the ACT's college-readiness threshold score. The vector of predictors, $\overline{X}_g$, includes the group-level fraction black, fraction on free lunch, teacher-pupil ratio, average 11th grade social studies score (standardized across individuals at the grade-year level), and average 8th grade math and English scores. We drop groups where there is not at least one ACT-taking student in the pre-policy and the post-policy periods, losing approximately 2% of the students in the sample.

We vary two features of the comparison. First, we vary the form of the control function, $h(.)$, while defining groups as schools. We use no control function, a linear function, a cubic function, a log function, and the inverse Mills ratio. The inverse Mills ratio is the appropriate functional form if the individual ACT score and ACT-taking errors are jointly normally distributed. The other functional forms can be interpreted as approximations to an unknown form of $h()$. The logarithmic form is used by Card and Payne (2002) and the linear and cubic forms follow from ideas in Heckman and Robb (1985) and Newey (2009).

We report the predicted mean ACT score and predicted fraction scoring college-ready in panel A of Appendix Table 12. The mean ACT score from the post-policy reference distribution is 19.26 and pre-policy is 20.63, again using inverse probability weighting to adjust for time differences in student demographics and school characteristics. The observed fractions college-ready are 0.443 and 0.569. Using the pre-policy data and omitting any selection correction generates predictions almost identical to the raw numbers (20.62 and 0.565). The control functions improve slightly on the uncorrected OLS regression but are nearly identical to one another and remain far from the benchmark value.[15] We also account for the possibility that the within-school selection process may differ between schools, by interacting the control function with the fraction of students who qualify for free lunch and the mean 11th grade test score. This allows the selection correction term, and hence the underlying distribution of individual errors, to vary by school type. However, this does not change the predicted outcomes. The estimates are robust over all our choices of the control function, echoing Card and Payne (2002) and Rothstein (2006). However, our results suggest that the estimates may simply be robustly incorrect.

Second, we vary the group definition, using demographic and academic subgroups within schools instead of schools. With these less aggregated groups, the predictor vector $\overline{X}_g$ contains

[15]We omit estimates from the cubic correction model, which are identical to those from the linear model.

Appendix Table 12. Group-Level Mean Latent ACT Score and Fraction College-Ready by Control Function and Aggregation Level

	Post-Policy (Truth)			Pre-Policy (Biased)		Pre-Policy, By Control Function Term					
	Raw	DFL	OLS	Raw	OLS	p	ln(p)	IMR(p)	p*Lunch	IMR(p)* lunch	IMR(p)* Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Panel A: School											
E[ACT*]	19.28	19.26	19.26	20.63	20.62	20.61	20.60	20.62	20.61	20.61	20.61
			(0.12)		(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Fraction ACT*>=20	0.443	0.443	0.440	0.569	0.565	0.564	0.562	0.564	0.563	0.564	0.565
			(0.010)		(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Panel B: Schl-Free Lunch-Minority											
E[ACT*]	19.28	19.20	19.09	20.59	20.40	20.43	20.41	20.44	20.44	20.43	20.39
			(0.12)		(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Fraction ACT*>=20	0.443	0.437	0.424	0.566	0.541	0.544	0.541	0.543	0.543	0.544	0.540
			(0.010)		(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Panel C: Schl-Free Lunch-Minority-Test Score Quartile											
E[ACT*]	19.28	19.25	19.11	19.96	19.49	19.52	19.61	19.51	19.54	19.54	19.59
			(0.11)		(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Fraction ACT*>=20	0.443	0.442	0.430	0.498	0.449	0.449	0.459	0.448	0.452	0.452	0.457
			(0.010)		(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)

Notes: The sample is as in Table 3 but excludes the 2% of the sample who enroll in high schools that do not appear in both 2005 and 2008 with at least one ACT-taker. Cells report the mean and fraction scoring greater than or equal to 20 for the predicted ACT score from group-level regressions of average ACT score on group-level covariates. IMR=inverse Mills ratio. Standard errors calculated using 1,000 bootstrap replications resampling schools.

more information, which facilitates better prediction. However, the group-level selection correction models correct only for within-group selection. Using less aggregated groups increases the scope for between-group selection and hence worse prediction. Using less aggregated groups thus emphasizes the role of the predictors relative to the corrections.

We begin by creating cells at the school-by-free lunch status-by-minority status level and report the results in panel B of Appendix Table 12. Disaggregating cells to this level leaves the raw post-policy mean and fraction college-ready unchanged, though the summary statistics for the post-policy reweighted and predicted distributions are slightly lower. The pre-policy predicted parameters are slightly closer to the truth than in panel A, closing approximately 0.2 points of the 1.4 point gap for the mean, and 2 of the 13 percentage point gap for the fraction college-ready. Again, the predictions do not differ with the functional form of the correction.

We next group the data at the school-by-free lunch status-by-minority status-by-11th grade test score quartile level and report the results in panel C of Table 12. Variants of this strategy are feasible when researchers observe prior academic performance for demographic subgroups of students, which are available in many NCLB-mandated school reports. The raw mean score and fraction college-ready are lower in the pre-period for this sample, while they are unchanged in the post-period.[16] The predictions are substantially better with this less refined data and some fall almost within the 95% confidence intervals of the parameters of the reference distribution (column 3). The functional form of the correction is again almost irrelevant; the uncorrected predictions are as accurate as any of the selection-corrected predictions.

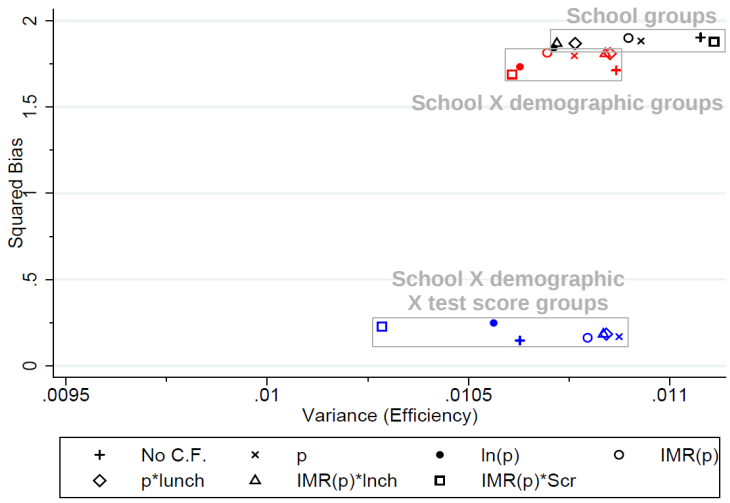
We display these estimates in Appendix Figure 15, showing the variance and squared bias for each combination of control functions and data aggregation levels. The finer aggregation levels clearly generate less biased estimates of the mean and fraction college-ready, particularly for the finest aggregation level; the estimates for the mean are also lower variance than those based on coarser aggregation levels. There is little variation across control functions in squared bias. There is some variation in variance, though no clearly dominant control function. We repeat this exercise using as a reference distribution the post-policy score distribution without reweighting and show the results in Appendix Figure 16. The results are unchanged.

We conclude that none of the functional form choices for the selection correction term robustly

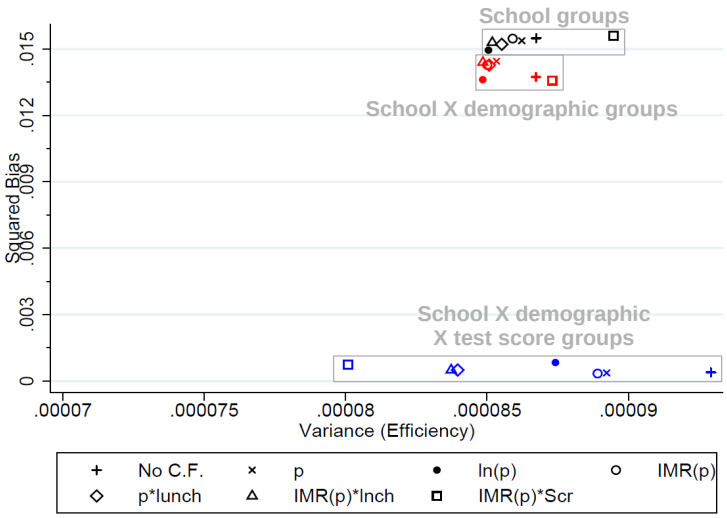
[16]The change in these statistics occurs for two reasons. Students with missing 11th grade scores are now dropped, as they do not fall into a test score quartile. There are also some school-by-poverty-by-test score quartile cells that contain no ACT takers. Students in these cells are assigned zero weight in this disaggregated analysis but received positive weight in the previous, more aggregated, analysis.

Appendix Figure XV: MSE Comparison Across Control Functions and Aggregation Levels

(a) Mean ACT Score



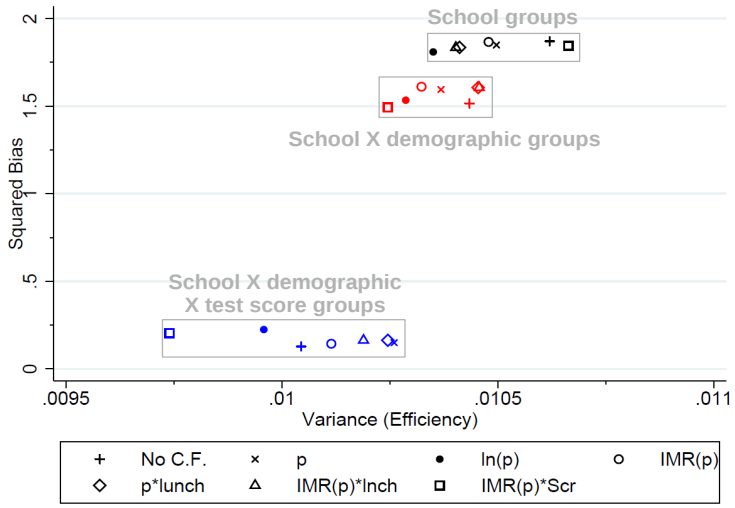
(b) Fraction College-Ready



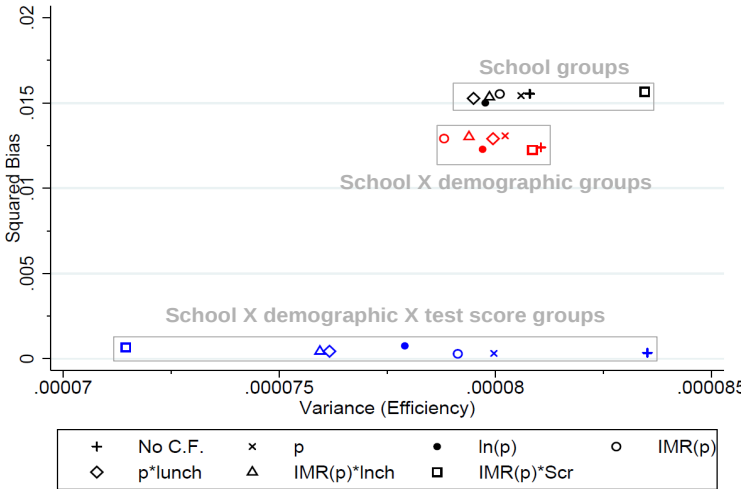
Notes: Figure shows the mean squared error of each combination of control function and data aggregation level for the group-level selection corrections from Table 5. Black (top of each figure): school-level; Red (middle): school*free lunch*minority-level; Blue (bottom): school*free lunch*minority*test score quartile-level. Bias is the difference between the statistic predicted by 1) the correction method applied to the pre-policy data and 2) the post-policy, DFL-weighted, fitted distribution.

Appendix Figure XVI: Group-Level MSE Comparison W/Out DFL Weights

(a) Mean ACT Score



(b) Fraction College-Ready



Notes: Figure shows the mean squared error of each combination of control function and data aggregation level for the group-level selection corrections from Table 5, fitting the post-policy distribution without DFL weights. Black (top of each figure): school-level; Red (middle): school*free lunch*minority-level; Blue (bottom): school*free lunch*minority*test score quartile-level. Bias is the difference between the statistic predicted by 1) the correction method applied to the pre-policy data and 2) the post-policy fitted (non-DFL weighted) distribution.

outperforms the others. However, the less aggregated data yields substantially more accurate predictions. This emphasizes the importance of the predictors, relative to the correction model, for prediction. Research based on highly aggregated data, such as state-level reports, should be interpreted with caution.

References

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

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

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

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

Arellano, M., and S. Bonhomme (2017): "Quantile Selection Models with an Application to Understanding Changes in Wage Inequality," *Econometrica*, 85(1), 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(12), 2976–2989.

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

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

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

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

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

D'Haultfoueille, X., and A. Maurel (2013): "Another Look at Identification at Infinity of Sample Selection Models," *Econometric Theory*, 29(1), 213–224.

Donald, S. (1995): "Two Step Estimation of Heteroskedastic Sample Selection Models.," *Journal of Econometrics*, 65(2), 347–380.

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

Hanson, S., and A. Sunderam (2012): "Another Look at Identification at Infinity of Sample Selection Models," *The Variance of Nonparametric Treatment Effect Estimators in the Presence of Clustering*, 94(4), 1197–1201.

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42(4), 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(4), 475–492.

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

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

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

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score.," *Econometrica*, 71(4), 1161–1189.

Kane, T., and C. Rouse (1995): "Labor Market Returns to Two-Year and Four-Year Colleges," *American Economic Review*, 85(3), 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(2), 507–512.

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

Mammen, E., C. Rothe, and M. Schienle (2016): "Another Look at Identification at Infinity of Sample Selection Models," *Econometric Theory*, 32(5), 1140–1177.

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

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(7), 1815–1820.

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

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

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

Rothstein, J. (2006): "Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions.," *American Economic Review*, 96(4), 1333–1350.

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