Data vs. Methods: Quasi-Experimental Evaluation of Alternative Sample Selection Corrections for Missing College Entrance Exam Score Data

Robert Garlick Joshua Hyman

Duke University Duke University

June 7, 2016

ERID Working Paper Number 221

Data vs Methods: Quasi-Experimental Evaluation of Alternative Sample Selection Corrections for Missing College Entrance Exam Score Data [1]

Robert Garlick[2] and Joshua Hyman[3]

June 7, 2016

Abstract

In 2007, Michigan began requiring all high school students to take the ACT college entrance exam. This natural experiment allows us to evaluate the performance of several parametric and semiparametric sample selection correction models. We apply each model to the censored, pre-policy test score data and compare the predicted values to the uncensored, post-policy distribution. We vary the set of model predictors to imitate the varying levels of data detail to which a researcher may have access. We find that predictive performance is sensitive to predictor choice but not correction model choice. All models perform poorly using student demographics and school- and district-level characteristics as predictors. However, all models perform well when including students' prior and contemporaneous scores on other tests. Similarly, correction models using group-level data perform better with more finely disaggregated groups, but produce similar predictions under different functional form assumptions. Our findings are not explained by an absence of selection, the assumptions of the parametric models holding, or the data lacking sufficient variation to permit useful semiparametric estimation. We conclude that "data beat methods" in this setting: gains from using less restrictive econometric methods are small relative to gains from seeking richer or more disaggregated data.

JEL codes: J01, I20, C10

---

[1] We thank Susan Dynarski, John Bound, Brian Jacob, and Jeff 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, Michigan, NBER Economics of Education, and SOLE. 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.

[2] Department of Economics, Duke University

[3] Corresponding author. Department of Public Policy, University of Connecticut. Address: 1800 Asylum Ave., 4th Floor, West Hartford, CT 06117; Email: joshua.hyman@uconn.edu; Telephone: (860) 570-9038; Fax: (860) 570-9114

I. Introduction

Economists and other social scientists routinely use datasets where the outcome of interest is unobserved for some cases. This situation is typically known as a "sample selection problem" when the latent outcomes are systematically different for the cases with observed and unobserved outcomes. Many canonical research areas in economics face this challenge: wages are unobserved for people who are unemployed or out of the labor force, test scores are unobserved for non-takers, output is unobserved for inactive firms, and outcomes are unobserved for participants who attrit from longitudinal studies or social experiments. Econometricians and statisticians have proposed many sample selection correction methods to recover the latent outcomes for those unobserved cases. However, without observing the complete distribution of latent outcomes as a measure of the truth, it is difficult to evaluate the performance of such methods at attaining this distribution.

We exploit a natural experiment that allows us to evaluate how well different sample selection correction methods recover the true distribution of latent outcomes. Between 2006 and 2007, Michigan implemented a policy requiring all high school students to take the ACT college entrance exam, increasing the percentage of students taking a college entrance exam from 64% to 99%. This sudden increase allows us to observe the complete post-policy distribution of scores, which we interpret as a measure of the "true" distribution. We use this benchmark to evaluate the relative performance of different sample selection correction methods applied to the pre-policy change distribution. We use student-level enrollment, demographic, ACT and SAT score, and state assessment data for two recent cohorts of eleventh grade students in Michigan straddling the implementation of the reform.

We compare the performance of a wide range of sample selection correction methods: linear regression, a one-stage parametric censored regression model (Tobin, 1958), a two-stage parametric selection model (Heckman, 1976; 1979), and several two-stage semiparametric selection models (Ahn & Powell, 1993; Newey, 2009; Powell, 1987). These models make successively weaker assumptions about the economic or statistical model generating the latent outcomes and the probability that the outcomes are missing. We vary the set of predictors used in each model in order to mimic the varying levels of data detail to which a researcher may have access. We examine whether the accuracy of these correction methods varies by student race or poverty status and whether the corrections can accurately predict latent race and income gaps in achievement. Finally, given that in some cases researchers only observe aggregate data, we evaluate the performance of group-level correction methods based on Gronau (1974). We vary both the functional form and data aggregation level used in these group-level correction methods.

We find that predicted test scores are very similar across sample selection correction methods, but vary substantially across sets of predictors. Models with relatively weak distributional and functional form assumptions produce predicted test scores that are almost identical to those produced from more restrictive models. Including an instrument, the driving distance from a student's home to the nearest test-taking center, does little to improve the predictions in the two-stage models. Using only student demographics and school- and district-level characteristics as predictors yields inaccurate predictions across all methods. However, all methods including simple linear regression yield accurate model predictions when students' prior and contemporaneous scores on other tests are included as predictors. These predictions are less accurate for black and low-income students, leading to incorrect predictions of latent achievement gaps. This result is consistent with findings by Avery & Hoxby (2013) and Hyman

(forthcoming), who find that academic performance more strongly predicts college-going for white and high-income than black and low-income students.

The results from the group-level analyses reveal the same pattern as the individual-level analyses: predicted test scores are similar across different functional form choices but vary substantially with the level of aggregation. We conclude that in this setting the gains from using less restrictive econometric methods are small relative to the gains from seeking better predictors: data beat methods.

Some authors focus on estimating the parameters of the outcome model (in our case, the ACT score model), rather than prediction of the outcome distribution.[1] We extend our evaluation of correction models to consider parameter estimation in appendix II. We compare parameter estimates from a selection-corrected regression of ACT scores using pre-policy data to parameter estimates from a regression of ACT scores using post-policy data. We find no robust evidence that less restrictive econometric models yield parameter estimates closer to their "true" values.

We consider several candidate explanations for the similarity of predicted test scores across models. There is a high student-level correlation across models in both predicted ACT-taking and predicted ACT scores. This does not appear to be explained by an absence of selection, the assumptions of the parametric models holding, or the data lacking sufficient variation to permit useful semiparametric estimation. Instead, we conclude that the violations of the parametric models' assumptions are simply not quantitatively important in this setting.

We believe this is the first paper to evaluate the performance of different selection correction models against a quasi-experimental benchmark. Other papers comparing estimates

[1] When data is missing due to non-response, selection correction models are arguably more commonly used to identify parameters of outcome models. When data is missing due to non-random selection into treatment, selection correction models are arguably more commonly used to identify distributions of outcomes (Heckman & Robb, 1986; Lee 1978; Willis and Rosen, 1979).

across selection correction models lack a quasi-experimental or experimental benchmark against which to evaluate their estimates (Mroz, 1987; Newey *et al.*, 1990; Melenberg & Van Soest, 1996; Clark *et al.*, 2009). Our approach is similar to a series of papers comparing the performance of different treatment effects estimators against experimental benchmarks (LaLonde, 1986; Heckman *et al.*, 1998; Dehejia & Wahba, 1999; Smith & Todd, 2005).[2]

Our findings are relevant to three distinct audiences. The first audience consists of researchers using sample selection corrections. We provide evidence about the relative performance of different selection correction models, which can inform their choices in future work. Sample selection is a pervasive feature of economic datasets and many applied researchers establish that their results are robust across different selection correction models (Krueger & Whitmore, 2001; Card & Payne, 2002; Angrist *et al.*, 2006; Rothstein, 2006). Our findings suggest that results may be robust across different modeling choices without being correct.

Second, our findings are relevant to econometricians developing and evaluating sample selection corrections. Econometricians have developed selection correction models that yield consistent parameter estimates under successively weaker assumptions (Tobin, 1958; Gronau, 1974; Heckman, 1976; 1979; Gallant & Nychka, 1987; Ahn & Powell, 1993; Das *et al.*, 2003; Newey, 2009). Several papers also compare the performance of different estimators using real data (Mroz, 1987; Newey *et al.*, 1990; Melenberg & Van Soest, 1996) or simulations (Goldberger, 1983; Heckman *et al.*, 2003; Paarsch, 1984; Vella, 1998). These papers provide valuable information about the performance of sample selection correction methods in different settings. However, the papers using real data do not observe a reference or true distribution and

[2] LaLonde (1986) compares experimental treatment effects estimates to estimates from several non-experimental methods, including the parametric normal selection model (Heckman, 1976 1979).We compare our findings to LaLonde's in more detail in section V.

so cannot distinguish between robustly correct and robustly incorrect corrections. The papers using simulated data do not necessarily reflect real-world selection processes.

Third, our findings are relevant to researchers and policymakers who want to use college entrance exam scores to assess and compare achievement levels across groups of students, schools, districts, and states. We find that college entrance exam scores can be used to predict latent levels of college-readiness in the population provided states have access to other measures of student test score performance. This finding echoes Clark et al. (2009), who study the extent of selection into ACT-taking in Illinois and argue that parametric selection correction models using group-level data can approximate the latent distribution of ACT scores. We further show that policymakers should exercise caution, however, when interpreting predicted race and income gaps in college entrance exam scores.

We introduce the sample selection problem in section IIa, discuss selection correction models in IIb, and explain our criteria for evaluating alternative corrections in IIc. We extend the discussion to include corrections based on aggregate, not individual, data in IId. This discussion is fairly brief and informal; we lay out the theory and discuss implementation thoroughly in appendix I. In section III, we introduce the data, describe the Michigan setting and policy experiment in more detail, and compare the pre- and post-policy ACT score distributions to document the extent of sample selection. We report the main findings in section IV and a series of robustness checks in appendix III. In section V we discuss and interpret our findings, exploring possible reasons for greater sensitivity of predictions to the choice of predictors than the choice of correction methods. We conclude in Section VI, offering suggestions for future empirical practice, and discussing the extent to which findings might generalize.

II. Sample Selection, Selection Correction Models, and Evaluation Criteria

IIa. The Sample Selection Problem

Early research into the sample selection problem focused on understanding female labor supply, which is complicated by the fact that wages are unobserved for women who are unemployed or not active in the labor market (Gronau, 1974; Heckman, 1974). We discuss sample selection in the context of studying the predictors of student achievement, using students' scores on a college entrance examination such as the ACT as a proxy for achievement. We observe this proxy measure for only a subset of the population of students and consider three types of selection.[3]

First, there is no sample selection problem if selection into test-taking is uncorrelated with all observed and unobserved predictors of student achievement. We can simply analyze the students with observed test scores and then safely generalize these findings to the entire population. Our results may not reflect causal relationships between student achievement and observed predictors but they do reflect a predictive relationship that is valid for the population. Second, there is a simple type of selection problem if selection into test-taking depends on observed and unobserved characteristics but the unobserved characteristics do not also influence test scores. For example, students might select into test-taking if and only if their results on previous tests exceed some threshold score or their family income exceeds some threshold level. In this case, we can analyze the students with observed test scores and obtain an internally valid description of the predictors of their achievement. We cannot necessarily generalize these finding to the entire population because some non-takers may have no observationally equivalent test-takers.

[3] These three types of selection correspond to the notions of missing-completely-at-random data (Rubin 1976), selection on observed characteristics, and selection on unobserved characteristics (Heckman & Robb 1985).

The third and most widely studied case of sample selection arises when selection into test-taking is determined by some unobserved characteristics, which may be correlated with predictors of student achievement. The distribution of both observed and unobserved predictors of student achievement may differ between test-takers and non-takers. We cannot even recover an internally valid relationship between test scores and observed predictors for a subset of the population that is well-defined in terms of observed characteristics. Most sample selection correction methods focus on this third case. This reflects both economists' concern with selection on unobserved characteristics and the fact that it nests the first and second cases. We now outline a formal model of the third form of selection. We note the special cases of the model that correspond to the first and second cases of sample selection.

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

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

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

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

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

where ACT_i^* is the latent ACT score of student *i* with observed score ACT_i. $TAKE_i^*$ is a latent variable with associated indicator $TAKE_i$ reflecting whether a student takes the ACT. Equation (1a) is the object of primary interest. We wish to recover β to describe the relationship between ACT scores and a vector of observed predictors X_i. We assume that the functional form of the relationship between ACT_i^* and X_i is known.[4] We restrict attention to models where ε_i is a mean

[4] We abstract away from uncertainty about which variables belong in the vector X_i and about the functional form of the relationship between ACT_i^* and X_i. We compare parametric, semiparametric, and nonparametric methods for the

zero unobserved scalar variable and where X_i and ε_i are additively separable.[5] 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 an instrumental variable Z_i that is independent of ε_i conditional on X_i, influences the probability of taking the ACT, and does not directly influence ACT scores conditional on taking the ACT. 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.

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

$$E[ACT_i|X_i] = X_i\beta + E[\varepsilon_i|g(X_i, Z_i) + u_i > 0, X_i]. \quad (2)$$

If u_i and ε_i are not independent, the compound error term is correlated with X_i, creating an endogeneity problem.[6] We draw a distinction between the sample selection problem, induced by missing values of ACT_i, and the more general identification problem, induced by dependence between X_i and ε_i in equation (1a). Even without sample selection, β may not be identified due to some other omitted variable problem or measurement error in X_i. We abstract away from this

ACT-taking model in (1b) only. We confirm in appendix 3 that our results are robust to some alternative specifications of the relationship between ACT_i^* and X_i.

[5] The additive separability assumption is common in the empirical and theoretical literature on sample selection. See Arellano and Bonhomme (2015) and Altonji *et al*. (2012) for exceptions. We informally test and fail to reject this assumption using our data, as discussed in appendix III. Additive separability is often not assumed in work focusing on partial identification or bounding of the parameters in equation (1a), following Manski (1989, 1990). We focus on correction methods designed to achieve point identification of β. "No-assumption" Manski-style bounds are very wide in our application. The bounds for the mean ACT score cover 50% of the distribution of observed ACT scores in the post-policy period.

[6] The first case of sample selection, missing-completely-at-random data, arises when $TAKE_i^*$ does not depend on X_i or Z_i (i.e. $g(X_i, Z_i) + u_i = u_i$) and u_i is uncorrelated with both ε_i and X_i. Selection is determined entirely by the unobserved characteristic u_i, which is in turn uncorrelated with any of the objects of interest. Hence, the conditional expectation $E[ACT_i|X_i]$ simply equals $X_i\beta$. The second case of sample selection, selection on observed characteristics, arises when u_i is uncorrelated with both ε_i and X_i. Selection is determined by the unobserved characteristic u_i and the observed characteristics (X_i, Z_i), but the former is uncorrelated with the objects of interest. Hence, the conditional expectation $E[ACT_i|X_i]$ equals $X_i\beta$ for a subset of the population defined in terms of (X_i, Z_i).

problem by assuming that the object of interest is the population linear projection of ACT_i^* on X_i. This means that the ordinary least squares estimator of β will be unbiased and consistent in the absence of sample selection. We therefore frame the discussion in terms of "predictors" of test scores rather than "determinants" or "causes."

IIb. Selection Correction Models for Individual Data

We estimate and evaluate eight selection correction models that use individual-level data. All are discussed in detail in appendix I. First, we estimate $ACT_i = X_i\beta + \varepsilon_i$ using ordinary least squares and the sample of ACT-takers. This approach provides a consistent estimator of β if selection into ACT-taking depends only on X_i and not on any unobserved student characteristics. Second, we estimate $ACT_i = X_i\beta + \varepsilon_i$ using a Tobit-style maximum likelihood estimator and the sample of ACT-takers (Tobin, 1958). This approach provides a consistent estimator of β if $\varepsilon_i = u_i$ is normally distributed. This allows ACT-taking and ACT scores to be jointly determined by the same unobserved student characteristic. If students with high latent ACT scores do not take the ACT (or vice versa), this correction is not appropriate. Both the OLS and Tobit models only require estimation of the ACT score model in equation (1a), not the ACT-taking model in equation (1b).

Third, we jointly estimate the ACT score and ACT-taking model using the Heckman selection correction framework (Heckman 1974, 1976, 1979) and assuming that $g(X_i, Z_i) = X_i\delta + Z_i\gamma$. This approach provides a consistent estimator of β if (ε_i, u_i) are jointly normally distributed. Under the joint normality assumption, the selection bias term in equation (2) can be calculated exactly. We can thus estimate the ACT-taking model in equation (1b), use the results from that model to construct an estimate of the selection bias term, and then estimate the ACT

model in equation (1a) controlling for the estimated selection correction. This does not impose the Tobit model's restrictive assumption that student selection into ACT-taking is based on their latent scores. But this approach relies on specific distributional assumptions and may perform poorly if there is no instrument Z_i that predicts ACT-taking but does not directly predict ACT scores (Puhani, 2002).[7] As our fourth model, we therefore estimate a Heckman selection correction model including an instrument. We use the driving distance from each student's home to the nearest ACT test center as an instrument (following Card, 1995, among others) and justify this decision in section IIIb.

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 of the semiparametric models is a combination of one of two ACT-taking models, estimated for all students, and one of two selection-corrected ACT score models, estimated for only ACT-takers and including some correction for sample selection. We first estimate the ACT-taking model using 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 (see, for example, Hirano *et al.*, 2003). We also estimate the ACT-taking model using a nonparametric matching estimator. Both models generate predicted values of the probability of taking the ACT and the series logit generates predicted values of the latent index $TAKE_i^*$ that determines ACT-taking decisions. We then use these predicted values to construct two selection corrections for ACT score model. The first selection correction approximates the bias term in equation (2) with a polynomial in $\widehat{TAKE}_i^*$, following Heckman and Robb (1985) and Newey (2009). The second selection correction removes the bias

[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 alternative parametric assumptions on the joint distribution that are also sufficient.

term by comparing groups of students with very similar values of $\widehat{TAKE}_i$, following Ahn and Powell (1993) and Powell (1987). These approaches are less restrictive than the Heckman models that rely on joint normality. But they do impose some restrictions on the joint distribution of (ε_i, u_i) and on the function $g(.,.)$ and may have poor statistical performance in even moderately large samples. We discuss the assumptions and implementation of these semiparametric models in detail in appendix I.

We refer to these eight models through the paper as respectively the OLS, Tobit, Heckman, Heckman with IV, semiparametric Newey, nonparametric Newey, semiparametric Powell, and nonparametric Powell models. We assume throughout that functional form of the latent ACT score model is known to be $X_i\beta$; we only vary the form of ACT-taking model and the form of the selection correction term.

IIc. Evaluating Alternative Selection Corrections for Individual Data

We evaluate each of the eight selection correction models by how well the model predictions match the uncensored distribution of ACT scores, which we observe in the post-policy period. Prediction is not the same problem as parameter estimation. We discuss this distinction in appendix II and show that our findings are unchanged when we evaluate the models on parameter estimation.

For each of the selection correction models, we use pre-policy data to estimate $\hat{\beta}$ and, for all students, $\hat{ACT}_i = X_i\hat{\beta}$.[8] 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 construct $\tilde{ACT}_i = \hat{ACT}_i + \hat{\varepsilon}_j$,

[8] When the selection model includes a selection correction term (Heckman, Newey), we include this term in the prediction. We show in appendix III that including these terms slightly increases the accuracy of the predictions. Including the correction terms in prediction also improves predictive accuracy in Monte Carlo simulations that we run.

where $\hat{\varepsilon}_j$ is the predicted residual from a randomly chosen student who took the ACT in the pre-policy period. This generates a distribution of predicted ACT scores with variance comparable to the latent distribution.[9] We finally calculate selected summary statistics for the empirical distribution $\hat{F}(A\tilde{C}T_i)$, averaging over 1000 iterations of the residual-adding process. We then evaluate the estimated test score distribution on three criteria:

1. The bias and variance of the predicted mean ACT score. The bias is measured by the difference between the predicted mean and the mean of the true distribution. Note that this statistic is unaffected by the residual-adding process, as the predicted residuals have zero mean.
2. The bias and variance of the predicted proportion of the students scoring above 19, the college-readiness threshold score recommended by the ACT (ACT, 2002). The bias is measured by the difference between the predicted proportion of scores above 19 and the proportion of scores above 19 in the true distribution.
3. The mean squared difference between $\hat{F}(A\tilde{C}T_i)$ and the true distribution, evaluated at percentiles 1, 2, …, 99. This measures the predictive fit of the full distribution and is informative about the predictive accuracy of other potential summary statistics.[10]

We estimate the variance for each of the three parameters (mean, fraction college-ready, squared difference between distributions) using the bootstrap.[11] We use a nonparametric cluster

[9] This method assumes that the distribution of the residuals conditional on Xi is symmetric, homoscedastic, and identical for ACT-takers and non-takers. Simulations show that our results are unaffected by allowing simple forms of conditional heteroscedasticity. We return to this issue in appendix II.

[12] This is similar to the Cramer-Von Mises criterion, except that we do not weight by the true density of ACT scores.

[11] The validity of the bootstrap does not appear to have been established for two-stage semiparametric selection correction models applied to clustered data. However, the bootstrap is typically used in empirical applications of these models. We believe that analytical variance estimators have been developed only for one-stage nonparametric estimators with clustered data (Hanson & Sunderam, 2012) or two-stage nonparametric estimators with independent data (Mammen *et al.*, forthcoming).

bootstrap, clustering at the school level to account for correlated unobserved school-level characteristics. We use 500 bootstrap replications, each containing 100 iterations of the residual-adding process.[12]

We construct the "true distribution" by adjusting the post-policy ACT test score distribution in two ways. First, we use inverse probability weights to adjust for differences in cohort demographics and other characteristics (DiNardo *et al.*, 1996). Second, we estimate equation (1a) for students in the post-policy period, with the inverse probability weights, and construct $\hat{F}(A\tilde{C}T_i)$ using the same procedure described above. This adjustment generates predicted test scores for the 1.5% of students who remain non-takers in the post-treatment period. Using predicted pre-policy scores for the entire pre-policy population and observed scores for only 98.5% of the post-policy population would be an inappropriate comparison.[13] However, all findings are robust to using the raw post-policy ACT distribution and to using the predicted post-policy distribution estimated without the inverse probability weights.

IId. Selection Correction Models for Group Data

We also evaluate the performance of selection correction models that use only group-level data. Researchers in many applications observe only group mean ACT scores and ACT-

[12] We use 100 iterations instead of the 1000 iterations used to create the statistics due to processing speed constraints.

[13] Assume that the population can be divided into three latent strata. The first stratum consists of students who take the ACT whether or not it is mandatory. The second stratum consists of students who take the ACT if and only if it is mandatory. The third stratum consists of students who do not take the ACT even if it is mandatory, including some special education students. In the pre-policy period, we observe test scores for the first stratum and wish to predict scores for the second and third strata. In the post-policy period, we observe test scores for the first and second strata. We should thus compare the predicted scores for the pre-policy cohort to the predicted scores for the full post-policy cohort, because we cannot distinguish between students in the second and third strata in the pre-policy cohort. In practice, the third stratum is sufficiently small that this distinction is irrelevant.

taking rates or, more generally, group mean censored outcomes and censoring rates. Card and Payne (2002) adapt equation system (1) for use with data aggregated to the group level:

$$ACT_{ig}^{*} = X_{ig}\beta + \varepsilon_{ig} \tag{3a}$$

$$TAKE_{ig}^{*} = W_{g}\mu + u_{ig} \tag{3b}$$

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

$$ACT_{ig} = \begin{cases} ACT_{ig}^{*} & \text{if } TAKE_{ig}^{*} \geq 0 \\ . & \text{if } TAKE_{ig}^{*} < 0 \end{cases} \tag{3d}$$

The key difference between systems (1) and (3) is the ACT-taking model. In this model we assume ACT-taking depends on a vector of group-level characteristics W_g and an individual error term, which may be correlated with ε_{ig}. We evaluate the observed test score equation at group means, yielding an estimating equation

$$\overline{ACT}_{g} = \bar{X}_{g}\beta + h\left(\overline{TAKE}_{g}\right) + \bar{\varepsilon}_{g} \tag{4}$$

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

This estimating equation is corrected for within-group selection but not for between-group selection, conditional on the observed ACT score predictors. Between-group selection occurs if group level ACT-taking rate covaries with the group mean latent ACT score. Within-group selection occurs if individual deviations from the group ACT-taking rate covary with individual deviations from group mean latent ACT scores. If groups are schools, for example, the correction model addresses selection from individual students within schools selecting non-randomly into ACT-taking but not selection from variation in ACT-taking across schools. This means that the level of aggregation is important for the credibility of this estimator. With larger

groups, more of the selection is within-group and is addressed by the selection correction.[14] However, the group mean predictors 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. If the individual errors in equations (3a) and (3b) are drawn from a joint normal distribution that does not vary across groups, then the selection correction term equals the inverse Mills ratio evaluated at the group mean ACT-taking rate (Card & Payne 2002; Clark *et al.*; 2009). We estimate equation (4) using a variety of functional forms for the selection correction term, including a polynomial in $\overline{TAKE}_g$, following the strategy in Newey (2009).[15]

Clark *et al.* (2009) use this approach to study the extent of selection in ACT- and SAT-taking in Illinois. They observe no data on non-takers (neither ACT scores nor lagged test scores and demographic characteristics). They can therefore use only these group-level methods and they focus specifically on 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.

[14] As the group size or number of groups shrinks to one, the selection correction term approaches a constant.
[15] We estimate equation (7) 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 IIc. We construct the standard errors using 500 replications of a nonparametric bootstrap, each containing 1000 residual-adding iterations.

III. Context, Data, and the Extent of Selection

We use a student level data set containing two recent cohorts (2004-05 and 2007-08) of all first-time 11th graders attending Michigan public high schools.[16] We drop students who do not go on to complete high school and students who take the special education version of the 11th grade test, as these students are not required to take the ACT post-policy. Using the last pre-policy cohort (2006) and first post-policy cohort (2007) would minimize demographic differences between the samples. However, several thousand students in the 2006 cohort were required to take the ACT as part of a pilot program the year before the statewide launch. Non-compliance was also higher in the first post-policy cohort, as districts struggled to adapt to the reform. We thus present results using the 2005 and 2008 cohorts, which have the largest difference in test-taking rates. Our results are robust to alternative cohort combinations.[17]

IIIa. Data

We use student-level administrative data from the Michigan Department of Education (MDE) that include first-time 11th grade students in Michigan public schools. The data contain time-invariant demographics such as sex, race, and date of birth, as well as time-varying characteristics such as free and reduced-price lunch status, limited-English-proficiency status (LEP), special education status (SPED), and student home addresses.[18] The data also contain 8th

[16] For the remainder for the paper, we refer to academic years using the spring year. For example, we refer to 2007-08 as 2008.
[17] We use three alterative combinations of cohorts: (a) the 2008 data as a benchmark to judge prediction in 2006, (b) the 2007 data as a benchmark to judge prediction in 2006, and (c) the 2007 data as a benchmark to judge prediction in 2005. When using 2006 data, we drop the students who took the ACT as part of the mandatory ACT pilot.
[18] Student home addresses were accessed on a restricted-access computer at MDE. The address data never left that computer and were not included in our main analysis data file.

and 11th grade state-assessment results in multiple subjects.[19] For the cohorts of students subject to the mandatory ACT exam, the 11th grade results include ACT scores.

We have acquired and merged on several other key pieces of information using a restricted access computer at the Michigan Department of Education. 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. We also 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.[20]

Table 1 shows sample means for the combined sample (column 1) and separately for the two cohorts of interest (columns 2 and 5). Figure 1 shows sample means for a smaller number of characteristics for six 11th grade cohorts straddling the policy change. The demographic composition of the Michigan grade 11 class changed between 2005 and 2008. The percentage of 11th graders who are black increased from 13% to 16%, while the percentage eligible for free lunch rose from 20% to 28%. The local unemployment rate rose from 7.3% to 7.7% during the sample period.[21] Comparison of the 2006 and 2007 cohorts show that these changes occurred smoothly over the four years, rather than jumping sharply. The fraction of students taking the

[19] All test scores are standardized within cohort and grade to have mean of zero and standard deviation of one. For 8th grade, we use the average of a student's math and English scores, where both are standardized before taking the average. For 11th grade, we use 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.

[20] 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. 2% of the sample has no non-missing address during any grade in high school and are dropped from the analysis.

[21] Unemployment rates at the city (when available) or county level are from the Bureau of Labor Statistics.

ACT, however, rose discontinuously when the policy was introduced between 2006 and 2007.[22] The ACT-taking rate rose from 64.1% in 2005 to 98.5% in 2008. While 95% of students in each school are required to take the 11th grade assessment for NCLB purposes, sitting the exam is not technically a graduation requirement, hence this small remaining gap. In 2005, 7.6% of high school graduates in Michigan took the SAT, compared to 3.9% in 2008.[23]

ACT-taking rates increased more for those groups of students who had lower rates prior to the policy. This is particularly pronounced among students eligible for free or reduced-price lunch, whose rate of ACT-taking more than doubled from 43% to 97%. These same groups tend to experience larger drops in their mean scores. Black students' ACT-taking rose by slightly more than that of white students, but black students' mean score decreased by slightly less than that of white students.

IIIb. Instruments for ACT-Taking

Several of the selection correction models we implement require an instrumental variable. The instrument should affect the probability of ACT-taking, not affect ACT scores conditional on ACT-taking, and be uncorrelated with the unobserved factors affecting ACT scores. The Heckman correction is identified without an instrument but its performance is typically poor without one (Puhani 2002). The Newey and Powell models are not identified without an instrument.

[22] The driving distance to the nearest ACT test center also dropped sharply between the 2006 and 2007 cohorts. 50.1% of students pre-policy had a center in their high school. All high schools post-policy were considered a test center, so the post-policy distance was simply the distance to the student's high school.
[23] For all of our analyses, we convert SAT scores to the ACT metric using the standard conversion table. For students who took the SAT or ACT multiple times, we use their first score.

We use the student-level driving distance from a student's home to the nearest ACT test-center as an instrument for ACT-taking. This is essentially a cost-shifter instrument. We assume that students with easier access to a test center have a higher probability of taking the test but do not have systematically different latent ACT scores, conditional on the other test score predictors.[24] The mean distance to a test center in the pre-policy period is 4.9 miles (Table 1, row 8), and the median is 3.1 miles. This distance measure varies substantially by urban/rural status: urban and rural means are 2.3 and 8.5 miles respectively. Appendix table 1 shows percentiles of the distance distribution by period and by urban/rural status. This instrument follows closely from prior research on education participation (Card, 1995; Kane & Rouse, 1995). We do not claim that the instrument is perfect, but rather that it is consistent with common empirical practice in the economics of education. This is the appropriate benchmark if we aim to inform empirical researchers' choice of selection correction models, conditional on the type of instruments typically available and in use.

We test if this distance variable is robustly associated with higher ACT-taking rates. Using pre-policy data, we run a probit regression of the ACT-taking indicator on a quadratic in distance.[25] The quadratic allows the marginal cost of 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 becomes stronger as we control for more student- and school-level characteristics. This likely reflects low test-taking by disadvantaged students living in urban areas who live close to a test center. The distance terms

[24] The first part of this assumption is supported by recent research: Bulman (2015) shows that SAT-taking rises by 3 percentage points when a high school becomes an SAT center, demonstrating a causal relationship between test-center proximity and test-taking.
[25] Standard errors are clustered at the school level and calculated without the bootstrap.

are individually and jointly significant when we control for basic student demographics ($\chi^2 = 22.38, p < 0.001$). Adding controls for school- and district-level characteristics and for student scores in other tests yields similar point estimates on the distance terms but reduces their estimated variance (so $\chi^2 = 25.15, p < 0.001$). This exceeds standard critical values for instrument strength, though these critical values are not developed for selection correction models (Stock & Yogo 2005). The probability of ACT-taking is robustly falling with distance, as theory suggests. Moving from the 5th to the 95th percentile of driving distance to the nearest ACT test center (14.1 miles) reduces the probability of taking the ACT by 12 percentage points, conditional on the full set of predictors. We return to the interpretation of the instrument in section V, including a discussion of identification at infinity.

We also use a placebo test to assess whether distance is associated with test performance conditional on test-taking. We regress the average of the student's 11th grade math and English state test scores on the quadratic in distance and report the results in columns 5-8 of table 2. Without controlling for any other predictors, students living further away from an ACT-test center tend to have higher scores. This again likely reflects low test scores and short distances to test centers in urban areas. However, the relationship disappears once covariates are included in the regression ($\chi^2 = 1.30, p = 0.480$). This shows that the distance measure is associated with ACT-taking but not latent academic performance, providing reassurance about the validity of the exclusion restriction.[26]

[26] We also attempt to test the validity of the exclusion restriction by regressing observed ACT scores in the post-policy period on the driving distance to the nearest ACT test center, when we would expect there to be no relationship. However, all schools became ACT test centers in the post-policy period, removing much of the variation in the driving distance measure. There is no statistically significant relationship between ACT test scores and driving distance in this period but the lack of variation makes this a very weak test.

IIIc. Describing Sample Selection by Comparing Pre- and Post-Policy ACT Score Distributions

In this subsection, we combine data from the pre- and post-policy periods to descriptively estimate the pre-policy distribution of ACT scores for non-takers. This distribution informs us about the nature of sample selection: positive/negative selection occurs if non-takers' scores are systematically higher/lower than ACT-takers' scores. This distribution also informs us about the magnitude of sample selection. Previous researchers using selected test scores often assumed that all non-takers would score below some percentile in the observed distribution (Angrist et al. 2006) or below all takers (Krueger & Whitmore 2001). We are able to assess whether these assumptions are plausible in our setting.

We estimate the latent ACT score distribution by subtracting the number of test-takers scoring at each ACT score in the pre-period from the number scoring at each score in the post-period. We reweight the post-policy cohort to have the same distribution of observed characteristics as the pre-policy cohort (DiNardo *et al.*, 1996). We estimate

$$\Pr(PRE_{is} = 1 | X_i, S_{st}) = L(\alpha + X_i\beta + S_{st}\delta) \qquad (5)$$

where $PRE_{is}$is an indicator for student *i* in school *s* being in the pre-period. *X* is a vector of individual level characteristics, *S* is a vector of time-varying school-level characteristics, and *L(.)* is the logistic function.[27] We use estimates from the logit regression to predict $\hat{PRE}_{is}$ and construct weights $PRE_{is} + (1 - PRE_{is})\left(\frac{\hat{PRE}_{is}}{1-\hat{PRE}_{is}}\right)$. We censor the top and bottom percentiles of the weight distribution and normalize so that the pre- and post-policy cohorts are of equal size. If the reweighting estimator accounts for all latent test score predictors that differ between the pre-

[27] *X* includes all interactions of LEP, SPED, free lunch status, race dummies, and a gender dummy. *S* includes fraction on free lunch, fraction black, number of 11th grade students, pupil-teacher ratio, student-guidance counselor ratio, and dummies for urban-rural status. The pseudo-R^2 from the regression is 0.149. We do not include student test scores in this reweighting model because they are standardized by year.

and post-policy periods, then the difference in the number of students at each ACT score equals the number of non-takers with that latent score.[28]

Figure IIa plots the frequency distribution of ACT scores pre-policy (blue circles), the reweighted post-policy distribution of scores (red squares), and the difference, or the latent scores of non-takers pre-policy (green triangles).[29] The censored test score distribution is approximately normal, reflecting the test's design. The latent test score distribution is shifted to the left but there is a long tail of students with reasonably high latent scores. The non-takers have a lower mean, higher variance, and greater skew. Almost 60% of takers score at a college-ready level, while less than 30% of the non-takers would do so. Finally, Kolmogorov-Smirnov tests of equality between the pre-policy non-taker and taker distributions, and between the observed pre- and post-policy distributions are strongly rejected. We plot the densities of observed and censored ACT scores in the pre-policy period in figure IIb, using vertical lines to show selected percentiles of the observed ACT score distribution. We also note the percentage of latent test scores that exceed each of these selected percentiles. For example, 68% and 24% of the latent scores exceed the 10th and 50th percentiles of the observed test score distribution.

The figure and table provide clear evidence of selection into ACT-taking. But the extent of selection is less extreme than that assumed in prior studies. Angrist *et al.* (2006) use Tobit analyses, censoring Colombian college entrance exam scores at the 1st and 10th (among other) percentiles. The authors suggest that while the assumption that non-takers would all score below the 1st percentile of observed scores is unlikely, it might be reasonable that they score below the 10th percentile. Neither assumption holds in our data, though we study a population in a different

[28] Hyman (forthcoming) 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 exam. That paper also examines the effect of the mandatory ACT policy on postsecondary outcomes.
[29] Appendix Table 2 reports moments and percentiles of the three distributions.

country. Both Angrist *et al*. (2006) and Krueger and Whitmore (2001) construct bounds on their parameters of interest assuming that all non-takers would score below specific quantiles of the observed distribution. The display in figure IIb shows that this type of assumption would hold only at very high quantiles, generating uninformative bounds. We conclude that correction methods relying on strong assumptions about negative selection are not justifiable in a setting such as this one.

IV. Results

IVa. Comparing Individual-Level Selection Bias Corrections

In this section we evaluate the performance of multiple sample selection correction methods. We estimate the selection-corrected distributions from the observed pre-policy ACT score distribution using the methods described in sections IIb, section IIc and appendix I. We construct the reference or "true" distribution from the post-policy ACT score distribution using the methods described in section IIc. We report all results in table 3 and summarize these results in figure V.

In table 3, we report the mean and fraction college-ready 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, discussed in section IIc, of the "true" uncensored 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.[30] We report the mean and fraction college-ready from the censored 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 "truth" in columns 1-3. We also report the mean-squared difference between each selection-corrected pre-policy distribution and the reweighted post-policy distribution with adjustments for missing ACT scores.[31]

Our first selection correction model uses a simple linear regression adjustment: we regress observed test scores on a vector of student demographics and use the coefficients to predict test scores, adding fitted residuals in line with the procedure from section IIb. The mean of the predicted values using OLS is 20.67 (standard error 0.10). The raw mean of observed scores in the pre-policy period is 20.86, so OLS adjustment closes only 11% of the gap between the observed and benchmark mean. Similarly, the fraction with latent scores at a college-ready level is 0.554 (standard error 0.008), only slightly closer than the raw pre-policy fraction (0.588) to the post-policy fraction (0.451). The mean squared difference between the predicted and observed distributions is 1.323 (standard error 0.148). The poor predictive fit for the entire distribution from OLS is perhaps unsurprising, as OLS only approximates the conditional mean

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

[31] This post-policy distribution is our preferred measure of the "truth" because it adjusts for changing covariates over time. However, the poverty measure used in reweighting may include time-varying measurement error: it is a binary proxy and marginally poor students who cross the poverty threshold post-policy are likely less disadvantaged than the average poor student pre-policy. Thus, we present the mean-squared differences and other results using the non-weighted, predicted post-policy distribution as the reference distribution in appendix figures III, IV, and V and appendix table 6. The pattern of results is similar, and the ideal reference distribution may be somewhere in between the weighted and non-weighted results.

function. There is likely to be substantial heterogeneity within each conditional mean cell (e.g., within race groups), which OLS will do a poor job of capturing.[32]

Our second selection correction model is a 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%. All three summary statistics from the Tobit model (table 3, column 6) are very similar to the OLS model (table 3, column 5). The predicted mean and mean squared difference fall slightly and the predicted fraction college-ready rises, but neither change is substantial or statistically significant.

We next test the performance of the Heckman two-stage correction procedure (table 3, column 7). We estimate a probit specification for the test-taking model and then estimate the selection-corrected test score model using OLS. When the test-taking model does not include an instrument, the mean predicted score (20.67), fraction college-ready (0.554), and mean squared difference (1.334) are essentially identical to those predicted by OLS and Tobit analysis. Performance of the Heckman correction changes only slightly when we use driving distance from students' home to the nearest ACT test center as an instrument for test-taking (table 3, column 8). The predicted mean ACT score does not change but the fraction college-ready and the estimated distribution move closer to the reference distribution.

Finally, we implement several two-stage semiparametric sample selection corrections: the Newey model that approximates the selection correction term in equation (2) with a polynomial in the predicted probability of test-taking, and the Powell model that removes the selection correction term by differencing the ACT scores and predictors with respect to observations with

[32] Appendix figure I replicates figure IIa for black/white and low-income/higher-income students. The latent test score distributions for all subsamples span a similar range to the full sample and each subsample distribution remains quite skewed.

similar values of the predicted probability of test-taking. We estimate each model using both the series logit and the nonparametric averaging to generate the predicted probabilities of test-taking, including the driving distance instrument in all cases. See appendix I for details on how we implement these estimators, including the data-driven choice of predictors in the series logit and order of the polynomial correction term. We report the results using the Newey correction in table 3 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 appears marginally more biased with the nonparametric than the semiparametric first stage.

IVb: 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 8^{th} and 11^{th} grade test scores) can do a better job at correction for selection in ACT scores. We report these results in the second panel of table 3. Adding these controls moves the predicted mean and fraction college-ready closer to the references values for all models and lowers the squared deviation between the predicted and reference distributions. However, the predicted means and fractions college-ready still exceed the reference values by at least 0.6 ACT points (0.27 standard deviations) and 6.4 percentage points respectively. 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 score in the prediction model. These data generally exist and are available to state education administrators, though researchers seldom have such data matched to a student's college entrance exam scores. We report these results in the third panel of table 3. The performance of all of the corrections is much better using the student-level scores in the prediction. This is perhaps unsurprising as we would expect students' past and contemporaneous achievement to be an excellent predictor of their 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. The predicted fractions college-ready are now within a percentage point of the reference fraction, with the exception of the semiparametric Powell correction. The full predicted distributions also move closer to the reference distribution, reflecting the extent to which prior test scores explain the variation in ACT scores. Although the skewness of the latent test score distribution visible in figure 2 caused serious problems for models with few predictors, this problem is much less serious with richer predictors.

The (a) predicted mean, (b) predicted fraction college-ready and (c) predicted distribution are closest to the reference values for, respectively, the (a) nonparametric Powell model, the (b) OLS model, and (c) the Newey model with a series logit first stage. There is no clear evidence that the semiparametric models robustly outperform the parametric models, single-equation models, or even simple OLS.[33] There is no clear pattern across the models in the variances of these summary statistics.

[33] We include several robustness checks where we further vary the set of predictors. These checks are described in appendix III with results presented in appendix table 6. Our main findings are robust to including interaction and polynomial terms 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.

We also present these results in two sets of figures. We plot kernel densities of the observed pre-policy and post-policy ACT scores in figure IIIa, along with the predicted scores from OLS estimation using both student demographics and school- and district-level covariates.[34] 95% confidence intervals of the predicted values are tiny and omitted for readability. We plot the corresponding densities in figure IIIb adding student-level test scores for prediction. As was apparent in table 3, the fit of the latter predictions is substantially better, with the fitted pre-policy distribution shifting left toward the fitted post-policy distribution. We plot predicted scores from several different selection correction models in figure IV. Panel A shows predictions from OLS (both pre- and post-policy) and the Tobit model and Heckman model (with the instrument). Panel B shows predictions from OLS (both pre- and post-policy) and the Newey and Powell models (both with the nonparametric first stage). These results mirror those shown in table 3 for mean squared difference (bottom row). In spite of capturing the mean and fraction college-ready similarly to other corrections, Tobit does poorly approximating the post-policy "true" distribution. The Powell and Newey models do not outperform the Heckman model.

Figure V provides a compact summary of our findings. We show each of the 24 predicted ACT means generated by a combination of the 8 selection correction models and 3 predictor sets in a bias-variance scatterplot (panel A). This allows us to visually compare the bias, variance, and mean squared error 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 true ACT mean. The predictions that also use student test scores are consistently much less biased and do not

[34] The predicted scores include the residual-adding procedure described in section IIb.

have higher variance. Within each covariate set, there is little variation in bias or variance across different selection correction mechanisms, except the Powell correction using a first stage series logit, 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.

Figure V panel B presents the same set of results for the fraction of students with college-ready ACT scores, instead of the mean ACT score. Once again, richer covariates lead to estimates with lower bias and no larger variance. Different correction models make little difference to the bias but can substantially influence the variance, particularly for the Powell model with the series logit first stage. The variance is, however, orders of magnitude lower than the squared bias and so is perhaps a less salient consideration. The pattern for the predicted test score distribution is very similar (figure V panel C), with the exception that the Tobit and semiparametric Powell models perform poorly using the rich set of predictors, and the variance is smaller for all corrections using these predictors.

What do our results imply for the interpretation of prior research using selection correction methods with student-level data in education? The common practice in prior research has been to use trimming methods such as the Tobit model, Heckman-style selection models, or bounding methods (Lee, 2009; Manski, 1990). Few papers have access to lagged or contemporaneous scores on other tests. For example, Angrist *et al*. (2006) address selection into Colombian college entrance exam-taking using Tobit models and Lee-style bounds. Krueger and Whitmore (2001) address selection into ACT- and SAT-taking using Heckman models without instruments and trimming methods similar to Tobit models. Both studies show that their results are robust to different modeling choices but use only student demographics and school

characteristics as predictors. Our results suggest that their findings should be interpreted with caution and that the robustness of their results to different modeling choices is not necessarily reassuring.

IVc. Comparing Individual-Level Selection Bias Corrections for Different Subgroups

We also evaluate how well individual-level selection correction models can predict the latent test score distribution for selected subgroups of the population. This is of interest for two reasons. Researchers, administrators, and policymakers are interested in mean scores and the college-readiness rate for the full population as well as for key student subgroups. Econometricians, applied and theoretical, are interested in how well selection correction models perform across different data generating processes. The latent ACT score distributions differ substantially for black, white, low-income, and higher-income students (appendix figure I). If the main pattern of results that we find for the overall sample holds across these subgroups, then concerns about the generalizability of our results are slightly reduced.

Reassuringly, we find that the main results hold across different subgroups. We estimate all eight selection correction models separately by race and free-lunch status using our three sets of model predictors (see appendix figures VI and VII). For all subgroups, as in the overall sample, we find that the choice of correction method makes little difference, but that the accuracy of the predictions increase substantially when including richer covariates. This robustness across different data generating processes addresses some concerns about the generality 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. We report the predicted mean ACT score

and fraction of students scoring at the college-ready level in table 4 and summarize the results in figure VI. There are large gaps in mean observed ACT scores and the college-readiness rate between black and white students and between low-income and higher-income students in the pre-policy period (row 3; columns 3, 6, 9, and 12). 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 and the college-readiness rate for all four subgroups. All selection correction models, applied to all four subgroups, successfully raise the predicted mean score and predicted college-readiness rate relative to the observed data. However, many of the models overestimate these measures of performance, 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 in the economics of education shows that the college application behavior of high-achieving, disadvantaged students do not match what would be predicted based on their academic performance (Avery & Hoxby, 2013; Radford, 2013; Hyman, forthcoming; Dillon & Smith, forthcoming). In other words, past achievement is less predictive of college application behavior among disadvantaged groups. This is consistent with our pattern of 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 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 natural experiment that we use in our analysis. Under a mandatory test-taking regime, some students induced to take the test may not exert the same amount of effort as those who take the test voluntarily.[35] We interpret distance to the nearest ACT test-taking center as a cost-shifting instrument. Conditional on other unobserved characteristics, students near a test center are more likely to take the ACT due to a lower cost. Conditional on observed characteristics and the instrument, students with larger utility gains from test-taking will be more likely to take the ACT. Thus, the selection-corrected distribution of pre-policy test scores reflects a hypothetical policy experiment that drives the conditional cost of ACT-taking so low that all students take the ACT and, presumably have some incentive to exert effort on the test.

In contrast, the observed distribution of post-policy test scores arises from mandatory ACT-taking and may include students who have little or no incentive to exert effort on the test. There is no accountability-related incentive or graduation requirement that depends on a student's ACT score. The ACT score can be used as a placement exam score at most Michigan community colleges, and as an entrance exam score at most 4-year colleges, but some students may know they will not attend any postsecondary school. The prediction model based on observed variation in testing costs from the distance instrument may over-predict latent scores for such students, predicting the latent score for those students *had they exerted full effort*. This may partly explain the robust overestimation of ACT scores for low-income and black students

[35] We contacted ACT Inc., asked them for the ACT score that a student would receive if they filled in answers at random, and found no large increase in the prevalence of this score post-policy, which we take as evidence that few students are exerting the minimum possible effort.

(figure VI). If these students face relatively low gains to ACT scores, perhaps due to information or financial constraints to college enrollment, then they will have relatively low incentive to exert effort on the test. Their observed scores will be lower than those predicted by the models based on a cost-shifting instrument, which extrapolates from a population who take the ACT voluntarily.[36]

IVd Comparing Group-Level Corrections

Many researchers using test scores as a dependent variable observe only students who take the exam and so cannot estimate individual probabilities of test-taking (Card & Payne 2002; Rothstein 2006). The individual-level corrections discussed thus far are infeasible in this case. In section IIf we laid out Card and Payne's (2002) strategy for estimating selection-corrected models using mean group scores and mean group test-taking rates. We estimate group-level selection models of the form of equation (4) using pre-policy data, generate the predicted distribution of group mean ACT scores, and compare this to the distribution 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 $\bar{X}_g$ 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

[36] The same argument implies that the reference distribution of ACT scores for our overall sample should perhaps have a higher mean and fraction college ready than the observed post-policy ACT score distribution. This is consistent with the pattern of results in table 3: we robustly predict a higher mean and higher fraction college ready than we observe in the post-policy data. However, our primary finding remains that the predictions vary little across correction methods but do vary substantially across covariate set.

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 functional form of the control function $h(.)$, while defining groups as schools. We set $h(k) = 0$ (i.e. no correction), k, $k + k^2 + k^3$, $ln(k)$, and $IMR(k)$. 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 (2009) 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 table 5. 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 true value.[37] 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

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

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, defining groups as demographic and academic subgroups within schools. With these less aggregated groups, the predictor vector $\bar{X}_g$ contains 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 table 5. 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 5. 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.[38] The predictions are substantially better with this less refined data and some fall

[38] 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.

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 figure VII, 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 conclude that none of the functional form choices for the selection correction term robustly 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.

V. Explaining the Similarity of Predicted Test Scores across Different Models

The results in section IVb show that the different sample selection correction methods generate similar distributions of predicted ACT scores. The different methods use different equations to model the probability of taking the ACT and the ACT scores themselves, so we should not necessarily expect the methods to deliver such similar outcomes. In this section we first diagnose whether the similar predictions arise from similarities in the student-level predicted probability of ACT-taking across different models, similarities in the student-level predicted

ACT scores across different models, or both. We then consider possible economic and statistical explanations for the similarities.[39]

Table 6 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 sets of predictors. There are two clear patterns in the table. First, the student-level predicted probabilities are very similar across the series logit and the two probit models, with correlation coefficients of at least 0.93. The predicted probabilities from the Heckman model with and without the instrument are almost identical. The nonparametric model generates slightly different predicted probabilities but they still have a correlation of at least 0.84 with the predicted probabilities from each of the other models. The central percentiles of the predicted probabilities are also similar across models, though there is more difference in the tails. These patterns go some way toward explaining the similarity of the predicted ACT score distributions across different selection correction models.

Second, the models with a rich set of predictors generate predicted probabilities that cover the whole unit interval. When only student demographics are used as predictors, the predicted values from all four models are bounded away from 0 and 1, other than a few outliers in the series logit. The demographics are all discrete, so they must produce a finite and small number of distinct values for the predicted probabilities. The three models that include the instrument can generate a larger number of predicted values as the instrument is continuous. All

[39] There are other approaches to nonparametric identification of sample selection models without instruments that we do not include in the paper (Lewbel, 2007; D'Haultfoeuille & Maurel, 2013). 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.

models generate predicted probabilities that span the unit interval when test scores and/or school- and district-level characteristics are included as predictors.

The student-level predicted ACT scores are also very highly correlated across models, as we show in table 7. The different selection correction models generate predicted ACT scores with correlations of at least 0.97 when using only student demographics as predictors. Including student test scores and school- and district-level characteristics marginally reduces these correlations, but all of them remain above 0.96.

So the similarity in predicted distributions of ACT scores is explained by very similar student-level predicted probabilities of test-taking and predicted ACT scores. But why do the different models deliver such similar predictions? We consider 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 sample selection corrections are unnecessary and will not improve on the simple OLS prediction model. We can clearly reject this explanation. The distributions of observed and latent test scores in figure II show clear evidence of negative selection into test-taking. The selection correction terms in both the Heckman and Newey models are large and significant predictors of ACT scores (see appendix tables 3, 4, and 5).[40] We conclude that there is strong evidence of sample selection, so some form of selection correction is required.

[40] 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 find that the coefficients are large and significantly different to zero. We reject the hypothesis of a zero coefficient for models with all combinations of the predictors and the instrument ($p < 0.001$). Moving from the 5th to the 95th percentile of the predicted probability of ACT-taking shifts the ACT score by 12.6 and 10 points, using the selection correction coefficients from the model with all predictors and respectively with and without the instrument. We implement an analogous test for the Newey model by testing if the coefficients on all four of the polynomial correction terms are zero. We again reject this hypothesis for all combinations of predictors ($p < 0.005$).

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. Figure II clearly shows that the latent test score distribution is not normal, and we verify this with parametric (skewness-kurtosis) and nonparametric (Kolmogorov-Smirnov) normality tests.[41] The latent distribution is also non-normal conditional on demographic characteristics (appendix figure I) 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). [42] We conclude that the structure of the selection problem, given the specification of the predictors, does not satisfy the joint normality assumption. There may exist some specification that satisfies the assumption, but this would not explain the central result of the paper.

Prior studies using Monte Carlo simulations or real data without quasi-experimental benchmarks found mixed evidence about the robustness of results from parametric and semiparametric models. Monte Carlo simulations by Goldberger (1983), Heckman *et al.* (2003), and Paarsch (1984) find that some parametric models perform poorly when the parametric assumptions are violated. But Newey *et al.* (1990) and Vella (1998) find that parametric and semiparametric selection models produce very similar results using real data. LaLonde (1986)

[41] The rejection of normality is not explained by our relatively large sample size. We also repeatedly reject normality for random 10% and 1% subsamples of the data.
[42] Joint normality is sufficient for identification in the Heckman model but is not necessary, so rejecting residual normality is not conclusive evidence against the appropriateness of the Heckman model.

compares treatment effects of employment training programs from experimental methods and non-experimental methods, including the parametric normal selection correction model. He finds that the selection-corrected estimates are not robustly similar to the experimental estimates. We offer two possible explanations for the difference between LaLonde's and our results, aside from the obviously different context. First, the predictors used in LaLonde's analysis may be less closely linked to the outcome of interest, even though he used lagged outcome variables as predictors. Second, we show in appendix II that the parametric normal selection model generally performs worse on parameter estimation than outcome prediction, unlike OLS and the semiparametric selection models.

Third, 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. We show in section IVb that the instrument is a robustly strong predictor of test-taking and does not predict other 11^{th} grade test scores. However, the instrument does not satisfy the identification at infinity assumption discussed in appendix I. In the probit specification with the rich set of predictors, moving from the 5^{th} to the 95^{th} percentile of instrument (14.1 miles) shifts the probability of test-taking by 12 percentage points. The relationships are similar for the series logit and the nonparametric estimators. This is substantially smaller than the shift from 0 to 100 percentage points required for identification of the intercept term in equation (1a) (Andrews and Schafgans, 1998; Chamberlain, 1986; Heckman, 1990). We can identify the slope coefficients in β but cannot separately identify the intercept coefficient in β from the level of the selection correction term.[43] The summary statistics

[43] The parametric additive structure of the model for latent ACT scores in equation (1a) may be important here. If the true model is not additive in observed and unobserved ACT predictors, then the limited range through which the instrument moves the probability of ACT-taking may be a larger problem.

shown in table 3 for the Newey and Powell models may thus have the wrong level. 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 & Rouse, 1995). However, we acknowledge that our ranking of different selection models may differ when an extremely strong instrument is available that permits identification of the intercept term in β.

Fourth, there may be a sample selection problem whose structure violates the assumptions of the parametric models, but there may not be enough variation in the predictors of test-taking for the semiparametric models to perform well. Some semiparametric models are identified only if at least one predictor is strictly continuous (Ichimura 1993; Klein & 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. Discrete and coarse data can also generate predicted probabilities that do not span the unit interval, limiting the effective variation in the selection correction terms.[44] 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 in the ACT scores predicted by different 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, each accounting for less than respectively 1.3% and 2.5% of all observations.

[44] If the support of the predicted probabilities is narrow, the selection correction terms may also be highly correlated with the ACT score predictors. Appendix figure II plots the inverse Mills ratio from the probit first stage against the index predictions, $X_i\hat{\delta} + Z_i\hat{\gamma}$. The relationship is approximately linear with only student demographics and the distance instrument included in the model. With the richer predictors, the relationship becomes nonlinear, though the nonlinearity is driven more by the predictors than by the instrument.

We conclude that there is a sample selection problem whose structure is not consistent with the assumptions of the parametric models and that the data are reasonably well-suited to semiparametric analysis. There are good theoretical reasons to expect the two-stage semiparametric models to perform better than the two-stage parametric models, and for the latter to perform better than the single-stage correction models. In this setting, it simply appears that the violations of the assumptions made by the more restrictive models are innocuous.

VI. Conclusion

College entrance exam scores on the ACT and SAT provide a useful measure of college-readiness. States, districts, and schools can use these scores to diagnose their performance at preparing their students for postsecondary education. Researchers can and do use these scores to evaluate the effects of education interventions (Krueger & Whitmore, 2001; Card & Payne, 2002; Angrist *et al.*, 2006; Rothstein, 2006). The main drawback is that less than half of public high school students nationwide take the ACT or SAT. Without knowing the nature of selection into test-taking, this limits an administrator, policymaker, or researcher's ability to harness these scores as a true measure of college-readiness that is representative of the overall population. Researchers have attempted to control for the resulting selection bias using a variety of parametric methods.

These attempts form part of a larger literature in econometrics and statistics that analyzes sample selection problems. These problems arise when dependent variables of interest are not observed for part of the population, and this part of the population may be systematically different on unobserved characteristics. Researchers have proposed a range of parametric and

semiparametric methods to address sample selection bias but there is little consensus on their relative merits in practical applications.

We use the implementation of a policy in Michigan requiring all 11th graders to take the ACT to compute the distribution of latent scores of students who were not taking the test prior to the policy. We show that the assumptions made by common selection correction and bounding methods are not correct in this setting. In particular, there is positive selection into test-taking but there is no value of the observed score distribution above which a trivial number of non-takers would score. We then use the near complete distribution of ACT scores post-policy to approximate the true distribution of latent scores pre-policy. We compare the ability of various sample selection corrections to match the true distribution of latent test scores.

We show that none of the sample selection corrections do well at matching the latent distribution of test scores when they use only basic demographic information to predict test scores and test-taking behavior. With more information about students, particularly measures of achievement such as state-administered standardized test scores, simple OLS does well at correcting for selection bias. There is little gain from using more flexible selection correction methods: censored regressions, bivariate normal selection models, and two-stage semiparametric selection models. We also show that the 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, we show that group-level correction methods using a control function evaluated at the test-taking rate of the group perform poorly. This poor performance is robust across different control function specifications, casting doubt on prior work claiming that robust results across control functions was reassuring. Aggregating the groups to increasingly

refined cells, in particular cells defined by prior test scores, substantially improves predictive accuracy.

What, if any, more general implications can be drawn from our findings? We do not claim that predictions from different selection corrections models will also be similar in very different settings, such as selection into wage employment (Heckman 1974, 1976), selection into education levels (Willis & Rosen, 1979), selection into different occupations or industries (Roy 1951; French & Taber 2011), or firm entry and exit (Olley & Pakes, 1996). However, three aspects of our results may provide some guidance for practice. First, we find that predictive accuracy 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 0.134, 0.198, and 0.614 respectively. Regressing ACT-taking on the instrument and the three sets of predictors yields pseudo-R2 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 for selection will work far better when lagged wage data is available as a predictor.

Second, our findings are not limited to settings where the assumptions of parametric selection correction models fail. 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 have sufficient variation 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 predictive gains from using these more flexible models are quantitatively small, particularly relative to the

gains from obtaining richer data. Researchers who believe that parametric model assumptions do not fit their application should not necessarily conclude that they will do better by estimating more flexible models.

Third, our findings are not limited to outcome prediction. Many selection correction applications focus on estimation of the parameters of some outcome model, rather than outcome prediction. The two problems are not statistically equivalent but our findings also apply to parameter estimation in this setting, as we discuss in appendix II.

We conclude that the richness of the data used to model selection into college entrance exam-taking matters far more than the econometric method used to correct for selection. This may come as a disappointment to researchers working with data that do not contain pre-treatment measures of the outcome or other covariates that should be very highly predictive of the outcome. In an education setting, this is good news for state and school district administrators and policymakers hoping to gauge their students' college-readiness. While these administrators and policymakers should not rely on being able to accurately forecast differences in scores across groups, simple OLS regression using data that contain students' test score information can come very close to providing the overall pictures of college-readiness that they seek. Our findings also suggest that researchers analyzing samples with selectively missing data will experience larger gains from seeking richer datasets than from more flexible econometric models. This reinforces results in the treatment effects literature emphasizing the importance of rich data for estimating treatment effects in non-experimental settings (Heckman *et al.*, 1998; Heckman & Smith 1999).

References

ACT Inc. 2002. “Interpreting ACT Assessment Scores.” Research Publication. http://www.act.org/research/researchers/briefs/2002-1.html#UItAIYq5fw [accessed May 4, 2013].

Ahn, Hyungtaik and James Powell. 1993. “Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism.” *Journal of Econometrics*, 58: 3-29.

Altonji, Joseph, Hidehiko Ichimura and Taisuke Otsu. 2012. “Estimating Derivatives in Nonseparable Models with Limited Dependent Variables.” *Econometrica*, 80(4): 1701-1719.

Angrist, Joshua, Eric Bettinger and Michael Kremer. 2006. “Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia.” *American Economic Review*, 96(3): 847-862.

Arellano, Manuel and Stephan Bonhomme. 2015. “Quantile Selection Models.” Working paper.

Avery, Christopher and Caroline 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.

Bettinger, Eric, Bridget Long, Phil Oreopoulos, and Lisa Sanbonmatsu. 2012. “The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment.” *Quarterly Journal of Economics*, 127(3):1205–1242.

Borra, Simone and Agostino 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, George. 2015. “The Effect of Access to College Assessments on Enrollment and Attainment.” *American Economic Journal: Applied Economics*, 7(4): 1-36.

Card, David. 1995. “Using Geographic Variation in College Proximity to Estimate the Returns to Schooling.” In Louis Christofides, Kenneth Grant, and Robert Swidinsky (eds) *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*. University of Toronto Press, pp. 201-222.

Card, David and Abigail 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, Gary. 1986. “Asymptotic Efficiency in Semiparametric Models with Censoring.” *Journal of Econometrics*, 32: 189-218.

Chen, Songnian and Shakeeb Khan. 2003. "Semiparametric Estimation of Heteroskedastic Sample Selection Models.” *Econometric Theory*, 19: 1040-1064.

Clark, Melissa, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2009. “Selection Bias in College Admissions Test Scores.” *Economics of Education Review*, 28: 295-207.

Das, Mitali, Whitney Newey, and Frank Vella. 2003. "Nonparametric Estimation of Sample Selection Models." *Review of Economic Studies*, 70: 33-58.

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

D'Haultfoueille, Xavier and Arnaud Maurel. 2013. "Another Look at Identification at Infinity of Sample Selection Models." *Econometric Theory*, 29(1) 213-224.

Dillon, Eleanor and Jeffrey Smith. Forthcoming. "The Determinants of Mismatch between Students and Colleges." *Journal of Labor Economics*.

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

Donald, Stephen. 1995. "Two Step Estimation of Heteroskedastic Sample Selection Models." *Journal of Econometrics*, 65: 347-380.

French, Eric and Christopher Taber. 2009. "Identification of Models of the Labor Market." In Orley Ashenfelter and David Card (eds) *Handbook of Labor Economics*, 4A: 537-617.

Gallant, Ronald and Douglas Nychka. 1987. "Semi-nonparametric Maximum Likelihood Estimation." *Econometrica*, 55: 363-390.

Goldberger, Arthur. 1983. "Abnormal Selection Bias." In Karlin, Samuel, Takeshi Amemiya, and Leo Goodman (eds) *Studies in Econometrics, Time-Series and Multivariate Statistics*. New York: Academic Press.

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

Hanson, Samuel and Adi Sunderam. 2012. "The Variance of Nonparametric Treatment Effect Estimators in the Presence of Clustering." *Review of Economics and Statistics*, 94(4): 1197-1201.

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

Heckman, James. 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.

Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, 47(l): 153-161.

Heckman, James and Bo Honore. 1990. "The Empirical Content of the Roy Model." *Econometrica*, 58: 1121-1149.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, 66: 1017-1098.

Heckman, James and Richard Robb Jr. 1985. "Alternative Methods for Evaluating the Impact of Interventions." in *Longitudinal Analysis of Labor Market Data*, Eds, James J. Heckman and Burton Singer, Cambridge University Press.

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

Hyman, Joshua M. Forthcoming. "ACT for All: The Effect of Mandatory College Entrance Exams on Postsecondary Attainment and Choice." *Education Finance and Policy*.

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

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

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

Krueger, Alan and Diane 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, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, 76: 604-620.

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

Lee, Lung-Fei. 1978. "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variable." *International Economic Review*, 19, pp. 415-433.

Lee, Fung-Lei. 1982. "Some Approaches to the Correction of Selectivity Bias." *Review of Economic Studies*, 49: 355-372.

Lee, Fung-Lei. 1983. "Generalized Econometric Models with Selectivity." *Econometrica*, 51(2) 507-512.

Lewbel, Arthur. 2007. "Endogenous Selection or Treatment Model Estimation." *Journal of Econometrics*, 141: 777-806.

Mammen, Enno, Christoph Rothe, and Melanie Schienle. Forthcoming. "Semiparametric Estimation with Generated Covariates." *Econometric Theory*.

Manski, Charles. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources*, 24: 343-360.

Manski, Charles. 1990. “Nonparametric Bounds on Treatment Effects.” *American Economic Review*, 80(2): 319-323.

Melenberg, Bertrand and Arthur Van Soest. 1996. “Parametric and Semi-Parametric Modeling of Vacation Expenditures.” *Journal of Applied Econometrics*, 11: 59-76.

Mroz, Tom. 1987. “The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions.” *Econometrica*, 55: 765-800.

Newey, Whitney. 1997. “Convergence Rates and Asymptotic Normality for Series Estimators.” *Journal of Econometrics*, 79: 147-168.

Newey, Whitney. 2009. “Two Step Series Estimation of Sample Selection Models.” *Econometrics Journal*, 12: S217-S229.

Newey, Whitney, James Powell, and James Walker. 1990. “Semiparametric Estimation of Selection Models: Some Empirical Results.” *American Economic Review Papers and Proceedings*, 80(2): 324-328.

Olley, G. Steven and Ariel Pakes. 1996. “The Dynamics of Productivity in the Telecommunications Equipment Industry.” *Econometrica*, 64(6): 1263-1297.

Olsen, Randall. 1980. “A Least Squares Correction for Selectivity Bias.” *Econometrica*, 48(7): 1815-1820.

Paarsch, Harry. 1984. “A Monte Carlo Comparison of Estimators for Censored Regression Models.” *Journal of Econometrics*, 24: 197-213.

Powell, James. 1987. “Semiparametric Estimation of Bivariate Latent Variable Models.” *Working Paper 8704*, Social Systems Research Institute, University of Wisconsin, Madison, WI.

Puhani, Patrick. 2002. “The Heckman Correction for Sample Selection and its Critique.” *Journal of Economic Surveys*, 14(1): 53-68.

Radford, Alexandria. 2013. *Top Student, Top School? How Social Class Shapes Where Valedictorians Go to College*, University of Chicago Press, Chicago, IL.

Rothstein, Jesse. 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.

Rubin, Donald. 1976. “Inference and Missing Data.” *Biometrika*, 63(3): 581-592.

Smith, Jeffrey and Petra Todd. 2005. “Does Matching Overcome LaLonde’s Critique of Nonexperimental Estimators?” *Journal of Econometrics*, 125: 305-353.

Stock, James and Motohiro Yogo. 2005. “Testing for Weak Instruments in Linear IV Regression.” In James Stock and Donald Andrews (eds) *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*. Cambridge University Press.

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

Vella, Frank. 1998. “Estimating Models with Sample Selection Bias: A Survey.” *Journal of Human Resources*, 33(1), pp. 127-169.

Willis, Robert and Sherwin Rosen. 1979. “Education and Self-Selection.” *Journal of Political Economy*, 87(5), pp. S7-S36.

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
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
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.

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)	-0.008	-0.025***	-0.019***	-0.024***	0.030***	-0.003	0.001	0.002
	(0.006)	(0.006)	(0.004)	(0.005)	(0.007)	(0.005)	(0.002)	(0.002)
Distance Squared (/ 10)	-0.000	0.007***	0.008***	0.010***	-0.014***	-0.002	-0.000	-0.001
	(0.003)	(0.002)	(0.002)	(0.002)	(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) are 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.

Table 3. Mean Latent ACT Score, Fraction College-Ready, and Quantile Differences by Correction Method and Control Variables

	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)
Panel A: X = Student Demographics												
E[ACT*]	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)
Fraction ACT*>=20	0.440	0.482	0.451	0.588	0.554	0.559	0.554	0.541	0.540	0.541	0.546	0.554
			(0.010)		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.008)
Quantile Differences	0.000	0.000	0.300	1.687	1.323	1.276	1.334	1.302	1.292	1.307	1.265	1.400
			(0.028)		(0.148)	(0.151)	(0.148)	(0.149)	(0.148)	(0.149)	(0.184)	(0.152)
Panel B: X = …Plus School-Level Covs												
E[ACT*]	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)
Fraction ACT*>=20	0.440	0.482	0.468	0.588	0.532	0.536	0.533	0.532	0.532	0.532	0.535	0.533
			(0.011)		(0.008)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Quantile Differences	0.000	0.000	0.325	1.687	1.058	1.053	1.078	1.062	1.073	1.070	1.127	1.097
			(0.022)		(0.128)	(0.130)	(0.129)	(0.128)	(0.130)	(0.129)	(0.132)	(0.129)
Panel C: X = …Plus Student Test Scores												
E[ACT*]	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)
Fraction ACT*>=20	0.440	0.482	0.468	0.588	0.469	0.460	0.463	0.463	0.460	0.463	0.497	0.479
			(0.011)		(0.008)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)	(0.010)	(0.009)
Quantile Differences	0.000	0.000	0.324	1.687	0.623	1.382	0.444	0.453	0.419	0.525	1.084	0.721
			(0.016)		(0.033)	(0.108)	(0.031)	(0.031)	(0.030)	(0.033)	(0.081)	(0.040)

Notes: The sample is as in Table 1, except only the 2005 and 2008 cohorts. Columns (1) and (3) 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, fraction scoring greater than or equal to 20, and quantile differences for the predicted ACT score from regressions of ACT scores on covariates. The predicted ACT score is calcuated for ACT-takers and non-takers. Standard errors calculated using 500 bootstrap replications resampling schools.

Table 4. Race and Poverty Gaps in Mean Latent ACT Scores and Fraction College-Ready by Correction Method

	E[ACT*]						Fraction ACT*>=20					
	Black	White	Gap	Poor	Non-Poor	Gap	Black	White	Gap	Poor	Non-Poor	Gap
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Post-Policy												
Raw	15.61	19.98	4.38	16.77	20.19	3.42	0.124	0.506	0.383	0.224	0.522	0.298
DFL	15.95	20.28	4.33	16.84	20.46	3.62	0.156	0.532	0.376	0.232	0.545	0.313
OLS	15.86	20.27	4.41	16.78	20.43	3.65	0.129	0.516	0.387	0.208	0.528	0.320
	(0.26)	(0.11)	(0.28)	(0.08)	(0.12)	(0.12)	(0.024)	(0.009)	(0.025)	(0.007)	(0.010)	(0.010)
Pre-Policy												
Raw	16.76	21.44	4.68	18.29	21.28	3.00	0.201	0.647	0.446	0.350	0.628	0.278
OLS	16.04	20.07	4.03	17.21	20.12	2.91	0.127	0.516	0.389	0.246	0.520	0.274
	(0.19)	(0.08)	(0.20)	(0.08)	(0.09)	(0.10)	(0.017)	(0.007)	(0.017)	(0.008)	(0.008)	(0.009)
Tobit	15.87	19.79	3.92	16.94	19.90	2.95	0.152	0.500	0.348	0.266	0.508	0.242
	(0.19)	(0.08)	(0.20)	(0.09)	(0.09)	(0.11)	(0.017)	(0.006)	(0.017)	(0.008)	(0.007)	(0.009)
Heckman	16.08	20.18	4.10	17.31	20.22	2.91	0.127	0.511	0.385	0.243	0.515	0.271
(with IV)	(0.18)	(0.08)	(0.19)	(0.09)	(0.09)	(0.11)	(0.017)	(0.007)	(0.018)	(0.008)	(0.008)	(0.009)
Newey -	16.05	20.22	4.17	17.31	20.25	2.93	0.127	0.509	0.382	0.243	0.511	0.267
Series Logit	(1.42)	(0.08)	(1.43)	(0.10)	(0.09)	(0.11)	(0.017)	(0.007)	(0.018)	(0.009)	(0.008)	(0.010)
Newey -	16.00	20.16	4.16	17.15	20.18	3.03	0.126	0.514	0.387	0.241	0.516	0.275
Nonparametric	(0.18)	(0.08)	(0.19)	(0.09)	(0.09)	(0.11)	(0.017)	(0.007)	(0.017)	(0.008)	(0.008)	(0.009)
Powell -	16.27	20.41	4.14	17.39	20.46	3.06	0.128	0.543	0.415	0.269	0.547	0.277
Series Logit	(0.20)	(0.11)	(0.22)	(0.11)	(0.10)	(0.13)	(0.017)	(0.010)	(0.019)	(0.011)	(0.009)	(0.013)
Powell -	16.12	20.17	4.05	17.41	20.22	2.80	0.135	0.523	0.389	0.265	0.528	0.263
Nonparametric	(0.19)	(0.08)	(0.21)	(0.09)	(0.09)	(0.11)	(0.017)	(0.007)	(0.018)	(0.010)	(0.008)	(0.011)

Notes: The sample is as in Table 3. Columns (1) - (6) report 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. Columns (7) - (12) report the fraction of the predicted ACT scores that are greater than or equal to 20. 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.

Table 5. Group-Level Mean Latent ACT Score and Fraction College-Ready by Control Function and Level of Aggregation

	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.

Table 6. 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)
Smallest Values												
1	0.295	0.251	0.279	0.165	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.004
2	0.295	0.251	0.279	0.174	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.007
3	0.295	0.251	0.279	0.177	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.008
4	0.295	0.251	0.279	0.182	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.008
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
Largest Values												
1	0.822	0.841	0.911	0.950	0.992	0.992	1.000	0.997	1.000	1.000	1.000	1.000
2	0.822	0.841	0.997	0.950	0.999	0.999	1.000	0.997	1.000	1.000	1.000	1.000
3	0.822	0.890	0.999	0.960	1.000	1.000	1.000	0.997	1.000	1.000	1.000	1.000
4	0.822	0.921	1.000	0.960	1.000	1.000	1.000	0.997	1.000	1.000	1.000	1.000
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	
Non-Parametric	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

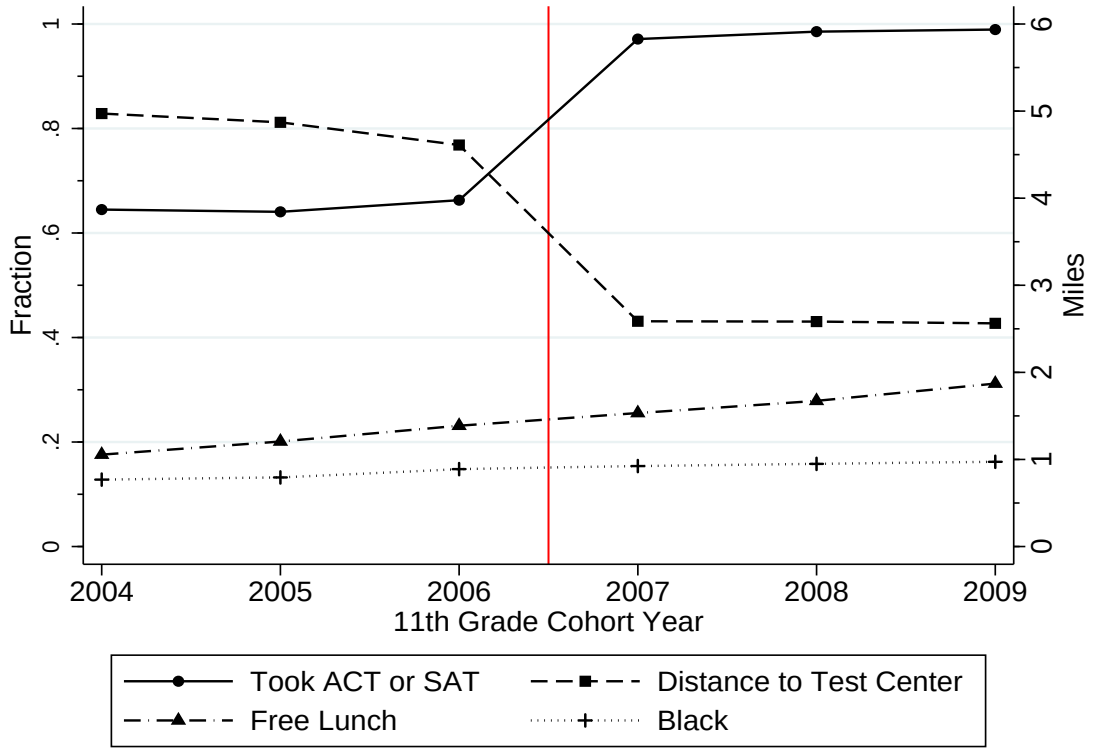
Notes: Table reports descriptive statistics and correlations of the first stage predicted probabilities across selection models and by covariate set included as regressors. The fraction of correct predictions is the fraction of predicted probabilities that after rounded to 0 and 1 using a cutoff of 0.36,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

Table 7. 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.

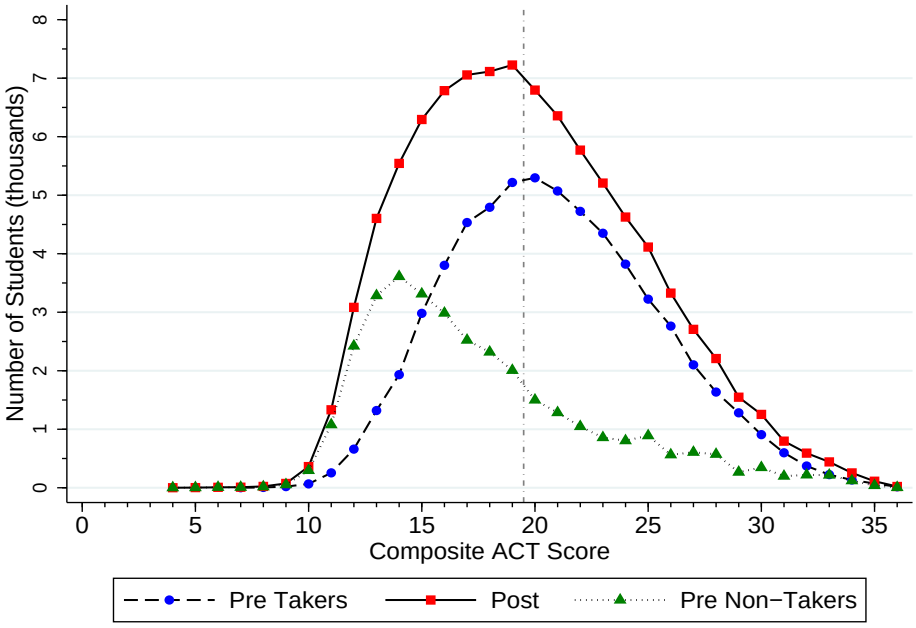
Figure I: Sample Characteristics by Cohort



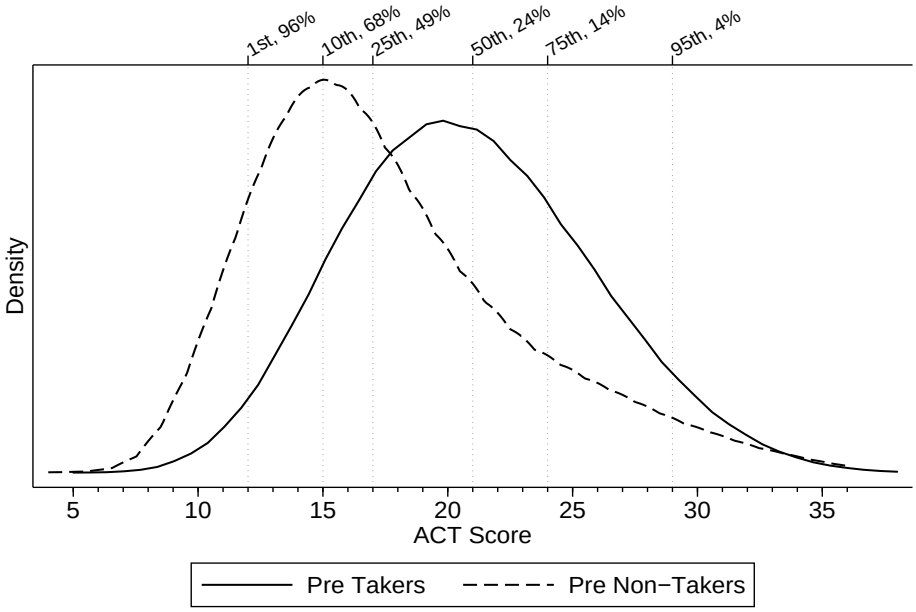
Notes: Figure shows means of four characteristics by eleventh grade cohort year: 1) Fraction taking the ACT or SAT, 2) Driving miles to nearest ACT test center, 3) Fraction free or reduced-price lunch, and 4) Fraction black. The vertical red line indicates the implementation of the mandatory ACT policy.

Figure II: Observed and Latent ACT Scores Pre- and Post-Mandatory ACT

(a) Calculating Latent Scores Pre-Policy (Frequency Distributions)



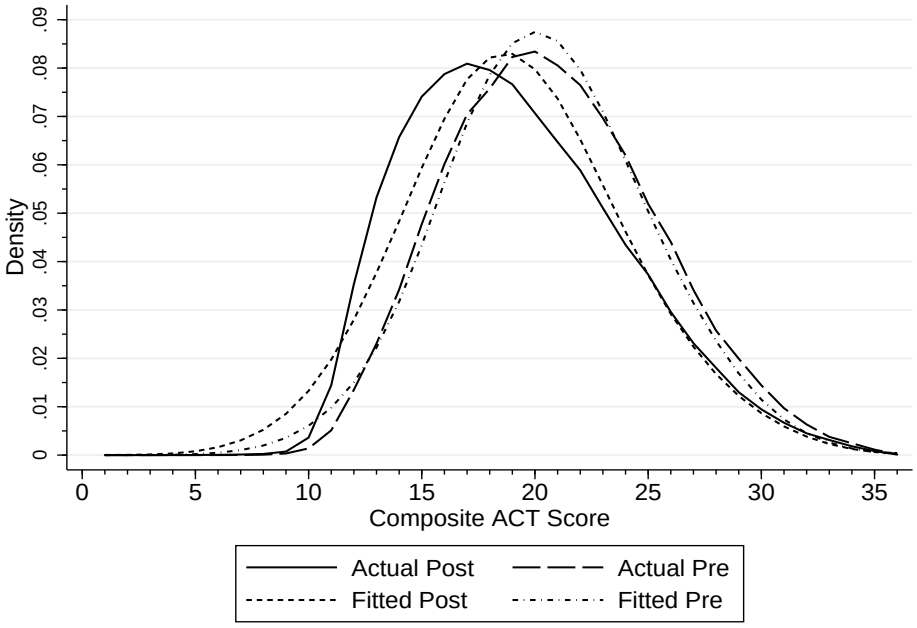
(b) Pre-Policy Observed and Latent Score Densities



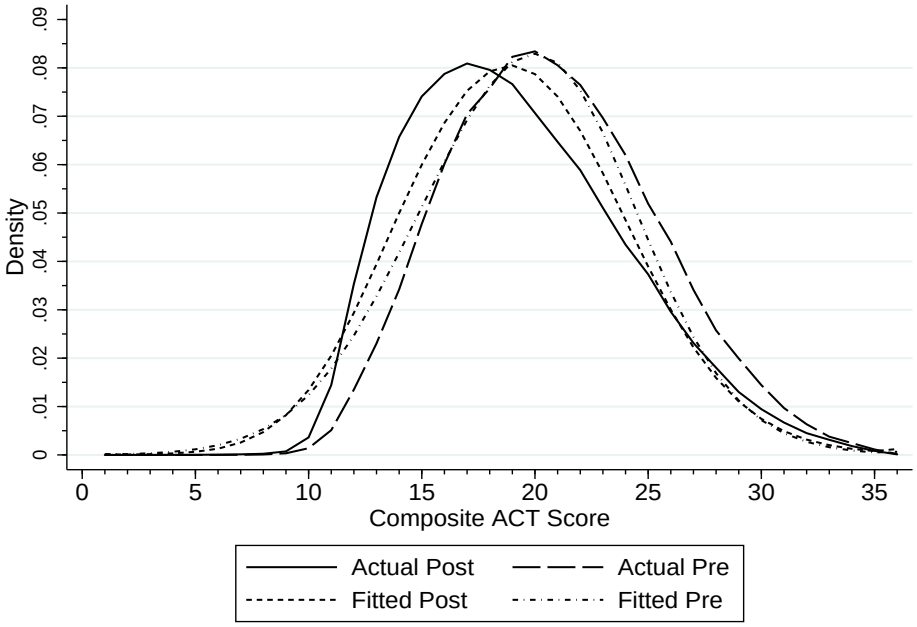
Notes: Figure (a) shows: 1) the distribution of ACT scores pre-policy, 2) the post-policy distribution reweighted following DiNardo, Fortin, and Lemieux (1996) to resemble the pre-policy cohort, and 3) the difference between (1) and (2), which is the latent score distribution among non-takers in the pre-period. Figure (b) plots kernel densities of (1) and (3). Along the top of the figure are percentiles of (1) followed by the fraction of (3) that has a latent score higher than that value.

Figure III: Observed and Predicted ACT Scores Pre- and Post-Policy

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



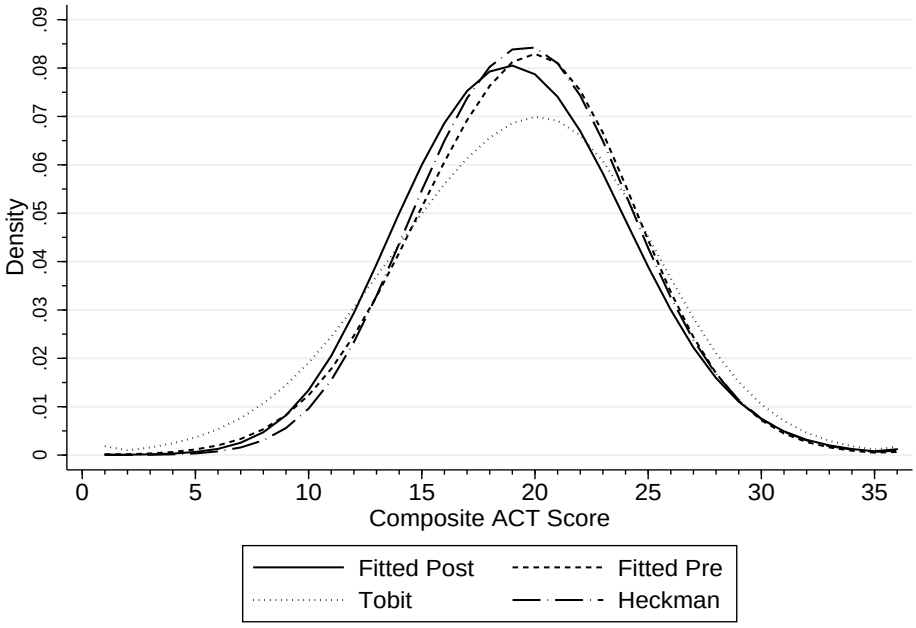
(b) Predicting ACT Scores Using Student 8th and 11th Grade Test Scores



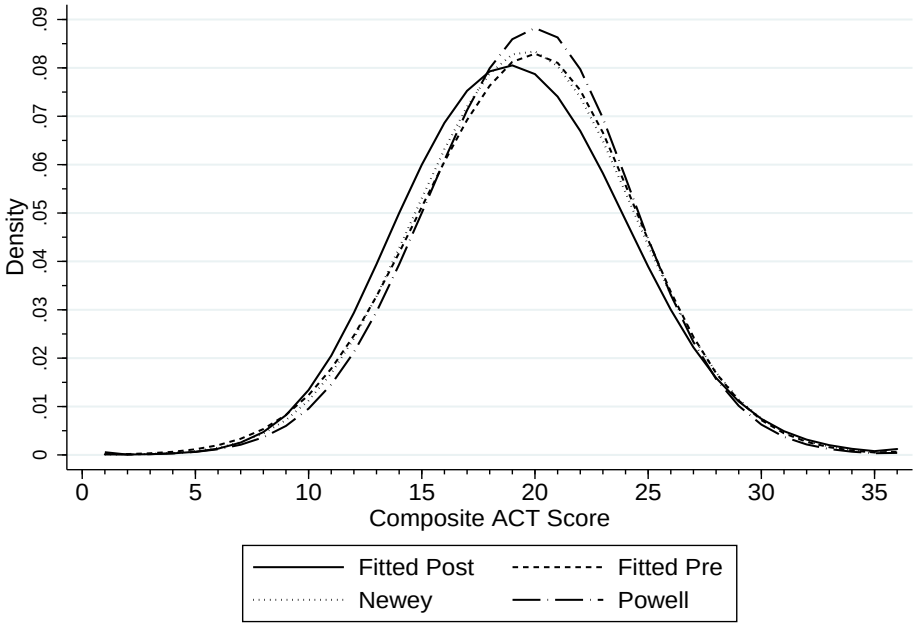
Notes: Figure (a) shows pre- and post-policy raw ACT scores and fitted values from regressions of ACT scores on student-level demographics and school- and district-level demographics and test scores. The post-policy regressions are DFL-weighted. The pre-policy fitted values are predicted out of sample 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.

Figure IV: Comparing the Performance of Sample Selection Corrections

(a) Parametric Corrections



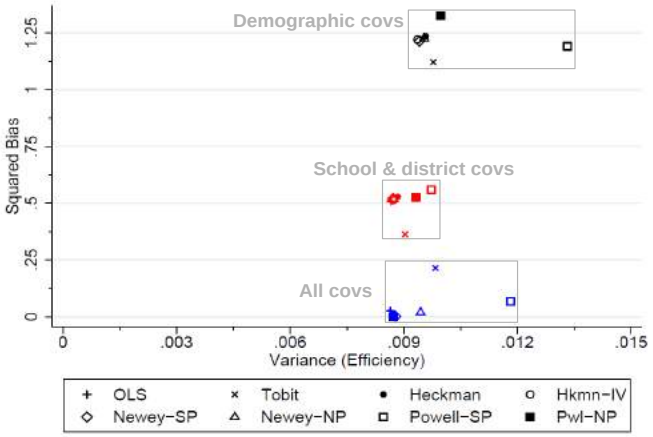
(b) Semiparametric Corrections



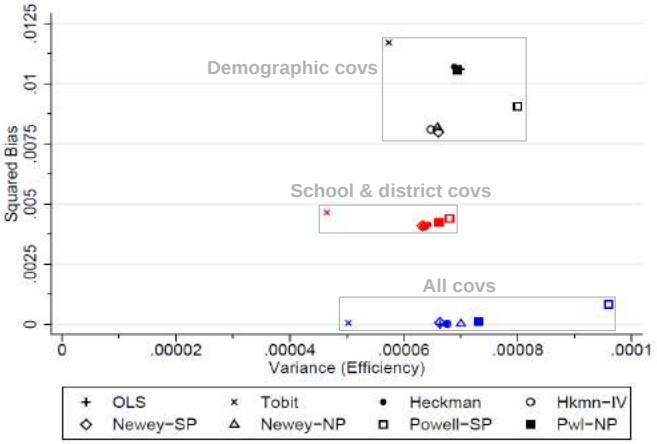
Notes: Figure shows pre- and post-policy fitted values from regressions of ACT scores on student-, school-, and district-level demographics, and 8th and 11th grade test scores. The post-policy regressions are DFL-weighted. The pre-policy fitted values are predicted out of sample to all students. Draws 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.

Figure V: MSE Comparison Across Correction Methods and Covariates

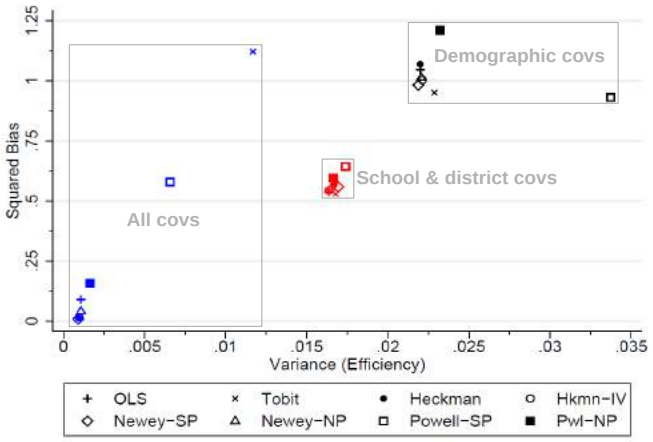
(a) Mean ACT Score



(b) Fraction College-Ready



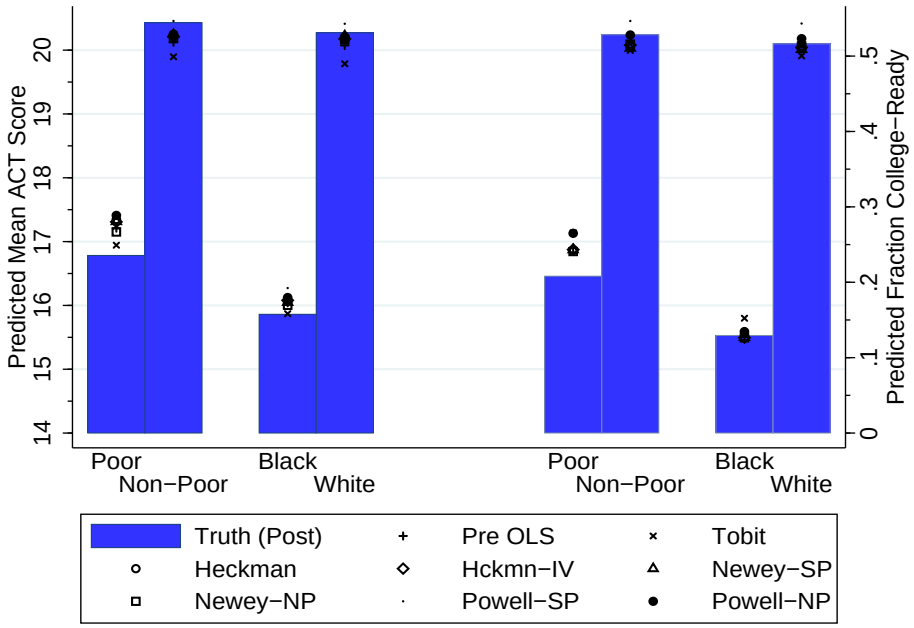
(c) Test Score Distribution



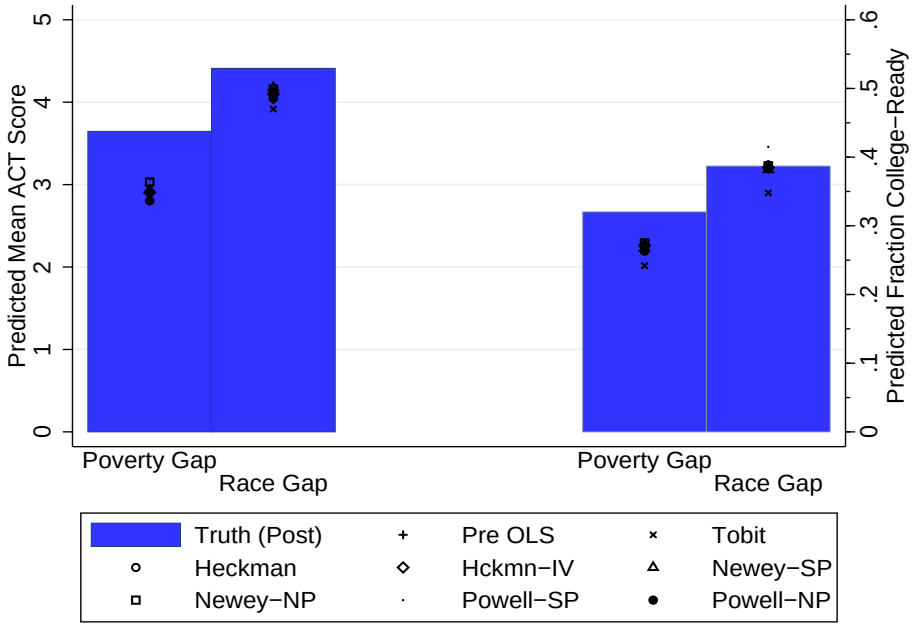
Notes: Figure shows the mean squared error of each combination of correction method and covariate set from Table 3. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the 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.

Figure VI: Race and Poverty Gaps in Mean Latent ACT Scores and Fraction College-Ready

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



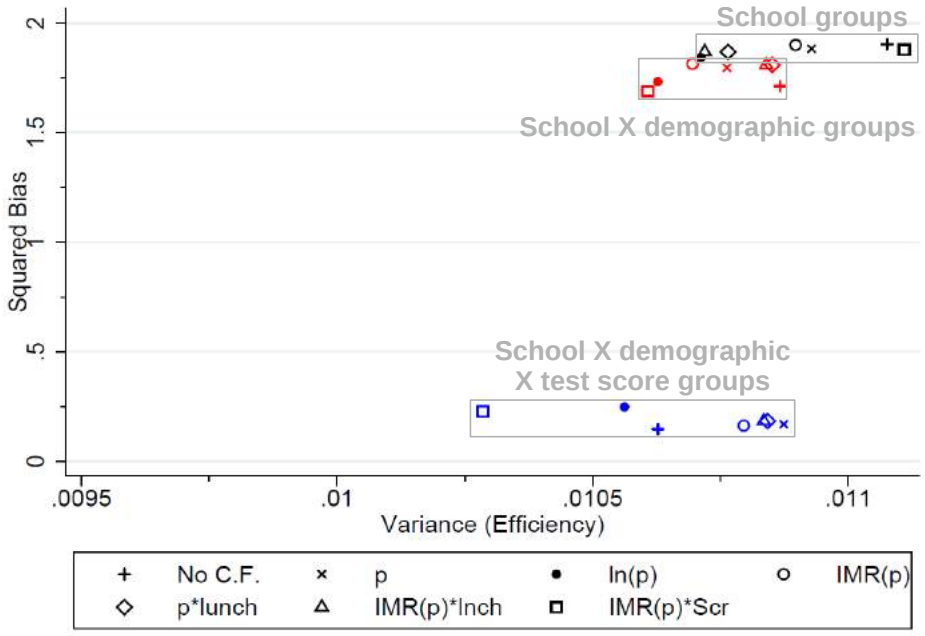
(b) Predicted Poverty and Race Gaps



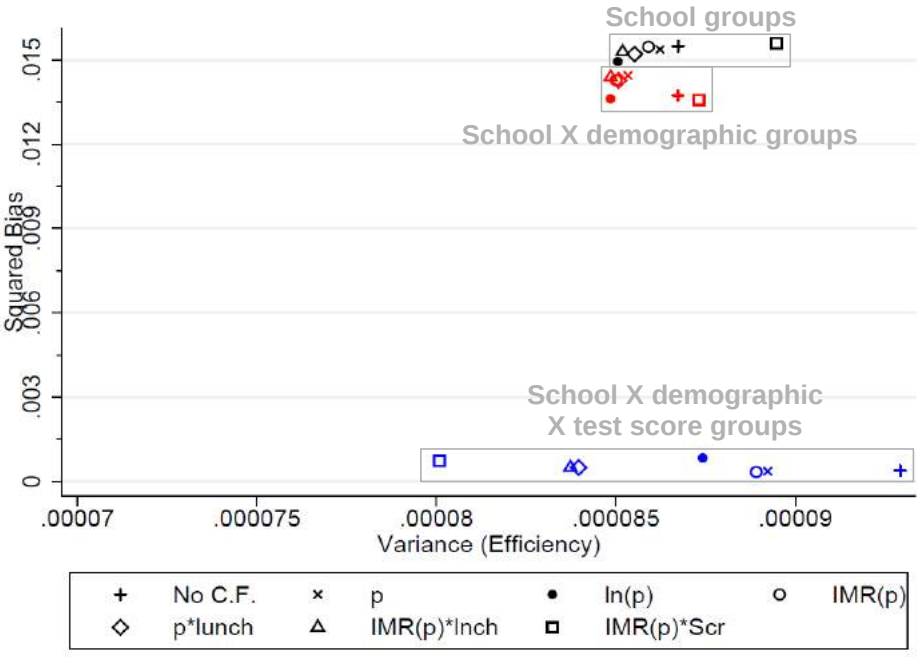
Notes: Figure (a) shows the "true" and predicted mean latent ACT score and fraction-college ready across correction methods by poverty status and race. Figure (b) shows the "true" and predicted gaps in these measures across correction methods by poverty status and race. All specifications include basic student demographics, school- and district-level covariates, and student 8th and 11th grade test scores.

Figure VII: 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 I: Selection Correction Models

This appendix elaborates on section IIb 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.

AIa. 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

$$ACT_i = X_i\beta + \varepsilon_i \quad \text{(A1)}$$

for the test-takers. This is a special case of system (1) where u_i and ε_i are independent and $Pr(TAKE_i = 1|X_i) > 0$ 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 adjustment assumes that ε_i is homoskedastic and normally distributed and that students take the ACT if and only if their latent scores exceed some threshold value $\overline{\overline{ACT}}$. Under these assumptions, we can assign the threshold score $\overline{\overline{ACT}}$ to all students who do not take the ACT, where $\overline{\overline{ACT}}$ is the lowest score obtained by any test-taker. In practice, researchers generally set $\overline{\overline{ACT}}$ higher than the minimum observed value and then assign the score $\overline{\overline{ACT}}$ to

both students with missing scores and students with non-missing scores below $\overline{\overline{ACT}}$. This necessarily discards information for some test-takers, and discards more information as $\overline{\overline{ACT}}$ is set higher. Under these assumptions, the parameter vector β equals the minimizer of the likelihood function

$$\mathcal{L}(\beta, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{\sigma} \varphi \left(\frac{TAKE_i - \beta X_i}{\sigma} \right) \right)^{TAKE_i} \left(1 - \phi \left(\frac{\beta X_i - \overline{\overline{ACT}}}{\sigma} \right) \right)^{1 - TAKE_i} \quad \text{(A2)}$$

where the first and second terms of the likelihood reflect the observed ACT scores and the probability of taking the ACT respectively. $\varphi(.)$ 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 66%. Results reported in section IV are robust to substantial changes in this threshold.

AIb. 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). In both cases, we estimate the systems:

$$ACT_i = X_i \beta + \frac{\varphi(X_i \hat{\delta} + Z_i \hat{\gamma})}{\phi(X_i \hat{\delta} + Z_i \hat{\gamma})} \theta + \varepsilon_i \quad \text{(A3a)}$$

$$TAKE_i^* = X_i\delta + Z_i\gamma + u_i \tag{A3b}$$

$$TAKE_i = \begin{cases} 1 & \text{if } TAKE_i^* \geq 0 \\ 0 & \text{if } TAKE_i^* < 0 \end{cases} \tag{A3c}$$

$$ACT_i = \begin{cases} ACT_i^* & \text{if } TAKE_i^* \geq 0 \\ . & \text{if } TAKE_i^* < 0 \end{cases} \tag{A3d}$$

where ε_i and u_i are jointly normally distributed and homoskedastic, and $\varphi(.)$ and $\phi(.)$ are the standard normal density and distribution functions respectively. Under the assumption of joint normality, the non-zero conditional mean error function $E[ACT_i|X_i] = X_i\beta + E[\varepsilon_i|u_i > -X_i\delta - Z_i\gamma]$ is a linear function of the inverse Mills ratio $\frac{\varphi(X_i\hat{\delta}+Z_i\hat{\gamma})}{\phi(X_i\hat{\delta}+Z_i\hat{\gamma})}$. Hence, estimating a probit regression of $TAKE_i$ on (X_i, Z_i) and equation (A3a) by ordinary least squares provides a consistent estimator of β.[1] We estimate equation (A3b) using only X_i as predictors ("Heckman") and also including a set of instruments Z_i that are excluded from equation (A3a) 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).

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 section AIc are all attempts to relax these distributional assumptions.[2]

[1] Alternatively, (A3a) and (A3b) can be jointly estimated by maximum likelihood, yielding a more efficient estimator. This maximum likelihood approach converges very slowly in our data, so we report only results from the two-stage OLS-ML estimation.

[2] 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.

AIc: Semiparametric Multiple-Equation Corrections for Sample Selection Bias ("Newey" and "Powell")

We now consider models of the form

$$ACT_i = X_i\beta + h(\widehat{TAKE}_i^*) + \varepsilon_i \quad \text{(A4a)}$$

$$TAKE_i^* = g(X_i, Z_i) + u_i \quad \text{(A4b)}$$

$$TAKE_i = \begin{cases} 1 & \text{if } TAKE_i^* \geq 0 \\ 0 & \text{if } TAKE_i^* < 0 \end{cases} \quad \text{(A4c)}$$

$$ACT_i = \begin{cases} ACT_i^* & \text{if } TAKE_i^* \geq 0 \\ . & \text{if } TAKE_i^* < 0 \end{cases} \quad \text{(A4d)}$$

where $g(.,.)$ and $h(.)$ are potentially unknown functions, and we do not assume a specific distribution for ε_i or u_i. There are a wide range of semiparametric sample selection correction models (Pagan & 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 $h(\widehat{TAKE}_i^*)$. 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 *et al.* (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:

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

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.[3] 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 & 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 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. [4] This yields 17, 585, and 731 terms when using the basic, school/district, and

[3] There does not appear to be a consensus procedure for choosing of series orders in nonlinear regression models, even though series logit models feature in several important econometric theory papers (Hirano, Imbens & Ridder 2003). Leave-one-out cross-validation is computationally burdensome in large datasets like ours.

[4] 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.

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 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 table 7 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 $E[TAKE_i|X_i, Z_i] = g(X_i, Z_i)$ rather than approximating it with a regression model. We start by calculating the Mahalanobis distance between every pair of observations *i* and *j*: $D_{i,j} = \sqrt{(W_i - W_j){V_W}^{-1}(W_i - W_j)'}$. 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: $\widehat{TAKE_i} = \sum_{k=1}^{K} \omega_{i,k} TAKE_k$. The weighting function $\omega_{i,k} = \frac{1/(1+d_{i,k})}{\sum_{k=1}^{K} 1/(1+d_{i,k})}$ assigns

more weight to observations with a lower Mahalanobis distance to i.[5] This estimator directly constructs the conditional mean $E[TAKE_i|W_i = w]$ at each value w without making assumptions about the nature of 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 $\widehat{TAKE_i^*}$, the predicted probability of test-taking (Newey 2009).[6] We select the order of the series using leave-one-out cross-validation. We then estimate equation (A4a) 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 (A4a) using a differencing approach (Ahn & Powell, 1993; Powell 1987). We calculate $dACT_i = ACT_i - \frac{1}{N-1}\sum_{j\neq i} w(i,j)ACT_j$ and $dX_i = X_i -$

[5] We use $1/(1 + d_{i,k})$ in the weighting function rather than $1/d_{i,k}$. Some pairs of observations have identical values for all elements in W_i and W_k so $d_{i,k} = 0$.

[6] Newey (2009) proposes using polynomials in either $\widehat{TAKE_i}$ or $\widehat{TAKE_i^*}$, the latent index that determines test-taking. 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 $E[\varepsilon_i|u_i > -X_i\delta - Z_i\gamma]$.

$\frac{1}{N-1}\sum_{j\neq i} w(i,j)X_j$, where $w(i,j)$ is a kernel or weighting function that is decreasing in the difference between i and j. 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 (A4a) as

$$dACT_i = \beta dX_i + d\varepsilon_i$$

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 are close together. In practice, we sort the data by the predicted probability of test-taking and use a weight function that equals $1/(1 + |\hat{p}_i - \hat{p}_j|)$ for $0 < |i - j| < 5$ and zero otherwise. We then estimate the differenced equation using weighted least squares, with higher weights assigned to observations that have "close" matches on the predicted probability of test-taking: $1/\left(\sum_{i-j=-4}^{4}|\hat{p}_i - \hat{p}_j|\right)$. We obtain essentially identical 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 omitting the weights when estimating the differenced equation.[7]

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 instrument: an

[7] 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. Simulations on a dataset with moments matched to our primary dataset show that the results are very robust to choices of different kernels.

observed variable that affects the probability of test-taking but does not directly affect test scores. Intuitively, the coefficient vector β and the selection term in (A4a) are separately identified only if there is additional information in the selection correction term (from an instrument) or the functional form of the term is known (from a set of parametric assumptions). The existence of an instrument is sufficient for identification of the slope coefficients in β but not the intercept. The intercept is identified when the instrument 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 instrument (Andrews & Schafgans, 1998; Chamberlain, 1986; Heckman, 1990). We note in section V that our proposed instrument does not satisfy this criterion and so we do not have a consistent semiparametric estimator of the intercept term in β. However, the relationship between our instrument and ACT-taking is similar to the relationships between other cost-based instruments and education participation (Card, 1995; Kane & Rouse, 1995; Bulman, 2015). The nonidentification of the intercept term does not stop the semiparametric models from delivering accurate predictions with sufficiently rich predictors.

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. Finally, 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 & Khan, 2003).

Appendix II: Prediction and Parameter Estimation

We evaluate selection correction models by running selection-corrected regressions of pre-policy ACT scores on a vector of predictors and comparing the predicted ACT scores to a reference distribution based on the post-policy ACT scores. Most theoretical papers and some empirical papers on selection corrections focus instead on parameter estimation. They try to correct the estimator of a specific parameter (or occasionally a vector of parameters) for selection bias. It is in principle possible that correction models' performance may be very different with respect to prediction and parameter estimation. In this appendix we consider the problem in more detail and show that our conclusions are robust across both prediction and parameter estimation.

To formalize this idea, note that the distribution of latent ACT scores $F_{ACT^*}(.)$ can be evaluated at any point a as $F_{ACT^*}(a) = E_X\left[F_{\varepsilon|X}(a - X\beta)\right]$, 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 $F_{\varepsilon|X}(.)$. The residual-adding procedure we discuss in section IIc effectively assumes that the error distribution does not vary with X or with ACT-taking: $F_{ACT^*}(a) = E_X\left[F_{\varepsilon,D=1}(a - X\beta)\right]$. This is a strong assumption. In particular, the assumptions of the Tobit and Heckman models imply that the error distribution should differ between ACT-takers and non-takers. The accurate predictions reported in sections IV and V suggest that with sufficiently rich predictors this assumption is innocuous.

We could instead adopt a parametric approach to identification of $F_{\varepsilon|X}(.)$. 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. So we could sample from this parametric distribution, rather than sampling from the empirical

distribution $\hat{F}_{\varepsilon|TAKE=1}(.)$. This would introduce another difference between the parametric (Tobit and Heckman) and semiparametric (Newey and Powell) selection correction models and might provide more scope for the semiparametric models to outperform the parametric models. The predicted values in simulations are unchanged when we sample from the parametric or empirical distribution.

Given that prediction entails arguably stronger identifying assumptions, would we obtain different results if we had focused instead on the relationship between "true" and selection-corrected parameters? In column 1 of appendix tables 3, 4, and 5 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.[8] 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.

First, the performance of all models is 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

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

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 0.38 to 0.18. This pattern is entirely consistent with the pattern across predictoions reported in section IVb. The only difference is that bias reduction from school- and district-level predictors is slightly larger for parameter estimation than for prediction.

Second, the more general 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 with an IV (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. 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.

Appendix III: Robustness Checks

We estimate several additional specifications to establish the robustness of these results. The results from these robustness checks are shown in tables 6 and 7 of the appendix. The point estimates include the residual-adding process but we omit standard errors for readability. We present three candidate "reference distributions." In column 1 we present summary statistics for the raw post-policy ACT score distribution. In column 2 we present summary statistics for the post-policy ACT score distribution adjusted using inverse probability weights to equate the distribution of race, gender, poverty, and selected school- and district-level measures to the pre-policy distribution. In column 3 we present summary statistics again using inverse probability weights and predicting the 1.5% of ACT scores that are missing in the post-treatment period using weighted least squares regression. The reweighting raises ACT scores slightly because the poverty rate in Michigan rose from 2005 to 2008. Including the predicted missing values lowers ACT scores slightly because special education students are overrepresented in this group. In all of the robustness checks, we compare the selection-corrected distribution to the reweighted reference distribution in column 2. The findings are unchanged if we use the raw or reweighted and predicted reference distributions.

First, we estimate all models with a complete set of interactions between the predictors and squares of all continuous predictors in both the first and second stages (table 6, panel 1).[9] 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.

[9] 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.

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 (table 6, 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 (table 6, 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 (table 6, 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.[10]

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

[10] 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.

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.[11]

Sixth, we predict the ACT score distributions from the Heckman and Newey models using $X_i\widehat{\beta}$ for prediction, instead of respectively $X_i\widehat{\beta} + \frac{\varphi(X_i\hat{\delta}+Z_i\hat{\gamma})}{\phi(X_i\hat{\delta}+Z_i\hat{\gamma})}\widehat{\theta}$ and $X_i\widehat{\beta} + \widehat{h}(\widehat{TAKE}_i^*)$. We show summary statistics for these predicted distributions in appendix table 7, columns 7 to 10. The correction terms do not technically "belong" in the prediction model, as their only role in the model is to permit identification of β. However, excluding these terms from the prediction model yields summary statistics that are generally farther from the summary statistics for the reference distributions. With the sparse sets of predictors, OLS outperforms both the Heckman and Newey models (appendix table 7, panels 1 and 2). With the rich set of predictors, OLS and the Newey model both outperform the Heckman model (appendix table 7, panel 3). These results show that the relative performance of OLS and the Heckman and Newey models, documented in section IV is not explained by the inclusion of the selection correction term in the prediction model.

[11] See Arellano and Bonhomme (2015) and Altonji *et al*. (2012) for exceptions.

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.

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 Table 3. 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 4. 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 5. 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.

Appendix Table 6. 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

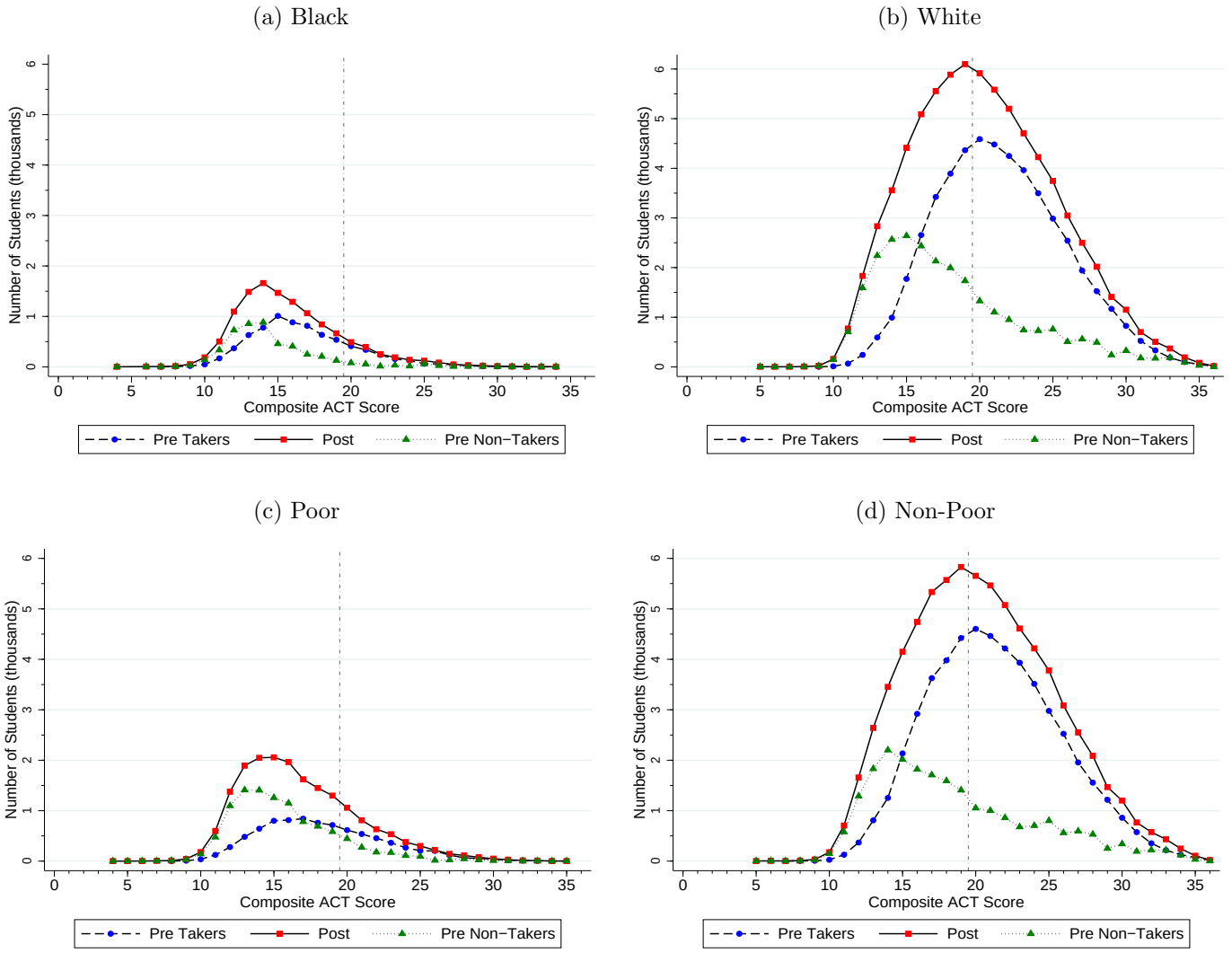
Notes: Table presents estimated parameters as in Table 3, but with slightly altered specifications. Standard errors are nearly identical to Table 3 and omitted for readability. Specification check 1 includes interactions between the predictors as well as squares of any continuous variables. Specification 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.

Appendix Table 7. Specification Check: Predicting ACT Scores Excluding Correction Terms

						Pre-Policy, by Correction Method			
	Post-Policy ("Truth")			Pre-Policy (Biased)		Heckman		Newey	
	Raw	DFL	OLS	Raw	OLS	No IV	With IV	Series Lgt	N.P.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
X = Student Demographics									
E[ACT*]	19.25	19.73	19.56	20.86	20.67	21.00	20.86	20.73	20.71
Fraction ACT*>=20	0.440	0.482	0.451	0.588	0.554	0.575	0.561	0.551	0.549
Quantile Differences	0.000	0.000	0.300	1.687	1.323	1.858	1.673	1.413	1.406
X = …Plus School-Level Covs									
E[ACT*]	19.25	19.73	19.77	20.86	20.48	20.95	20.63	20.52	20.50
Fraction ACT*>=20	0.440	0.482	0.468	0.588	0.532	0.559	0.540	0.535	0.533
Quantile Differences	0.000	0.000	0.325	1.687	1.058	1.771	1.342	1.144	1.100
X = …Plus Student Test Scores									
E[ACT*]	19.25	19.73	19.69	20.86	19.52	21.08	20.76	19.96	19.69
Fraction ACT*>=20	0.440	0.482	0.468	0.588	0.469	0.565	0.548	0.498	0.479
Quantile Differences	0.000	0.000	0.324	1.687	0.623	2.006	1.606	1.107	0.715

Notes: Table presents estimated parameters as in Table 3, but when predicting the ACT score, excludes the correction terms from the prediction. Standard errors are nearly identical to Table 3 and omitted for readability.

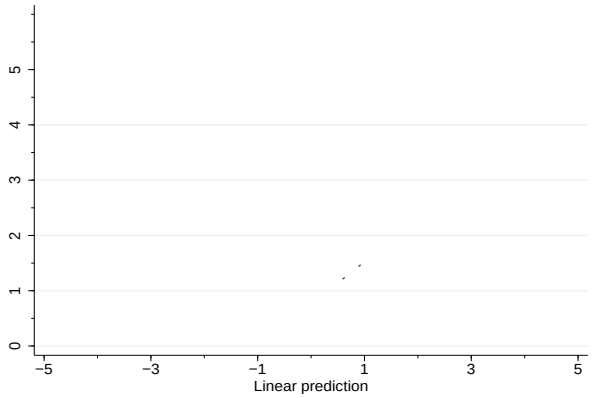
Appendix Figure I: Observed and Latent ACT Scores, By Subgroup



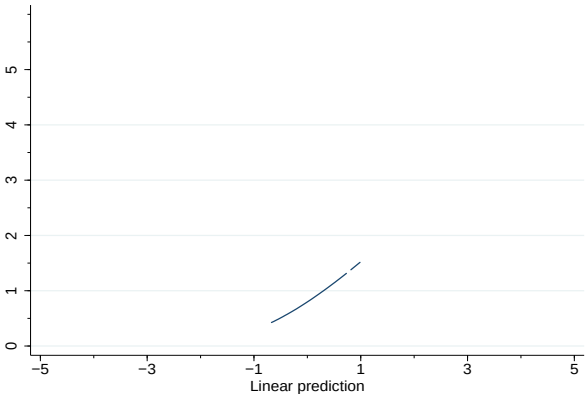
Notes: Figures show 1) the distribution of ACT scores pre-policy, 2) the distribution post-policy reweighted following DiNardo, Fortin, and Lemieux (1996) to resemble the pre-policy cohort, and 3) the difference between (1) and (2), which is the latent score distribution among non-takers in the pre-period.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

Appendix Figure II: IMRs vs Linear Predictions From Probits

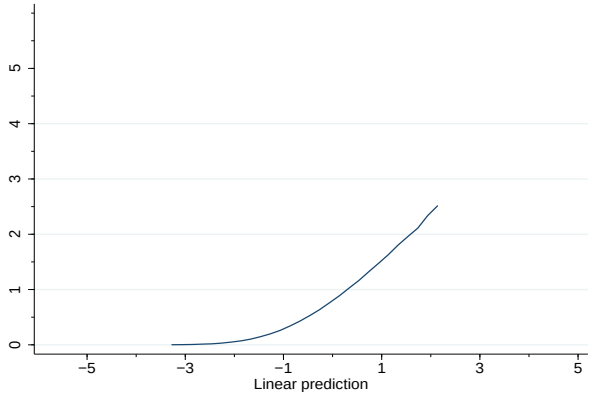
(a) No IV - Student Demographics



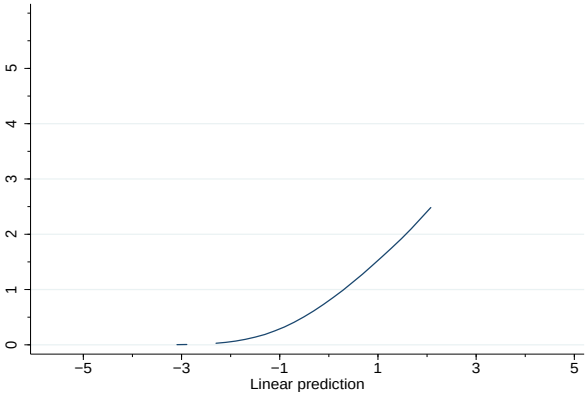
(b) IV - Student Demographics



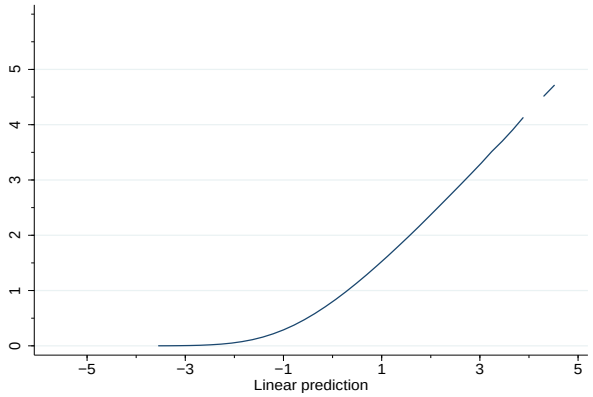
(c) No IV - Plus School Covariates



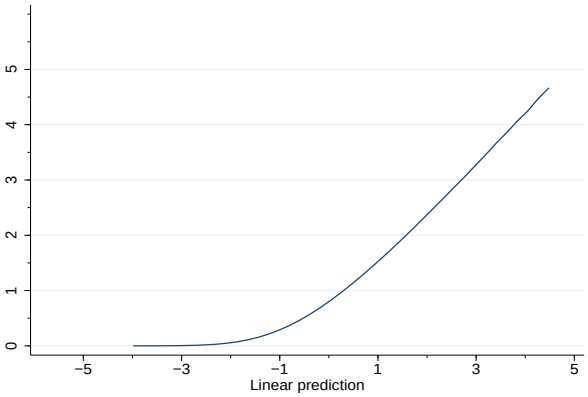
(d) IV - Plus School Covariates



(e) No IV - Plus Student Scores



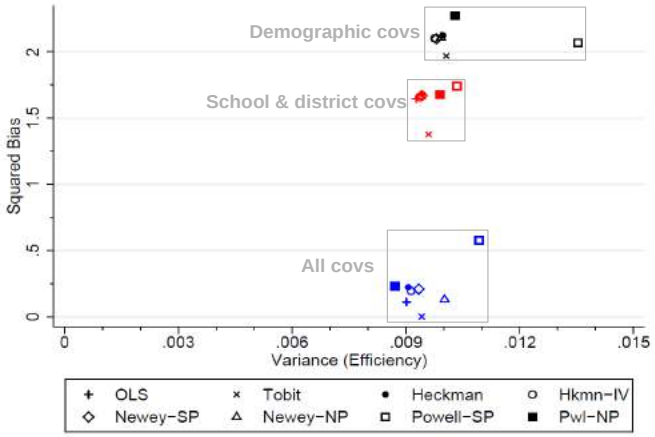
(f) IV - Plus Student Scores



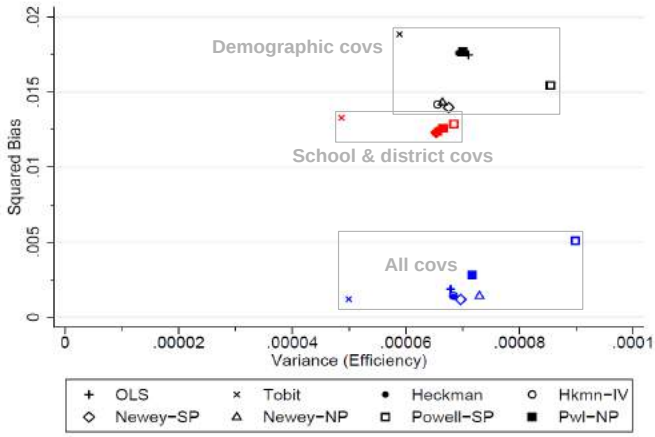
Notes: Figures plot the inverse Mills ratio against the linear prediction from the first stage Heckman corrections, with and without an IV and by covariate set.

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

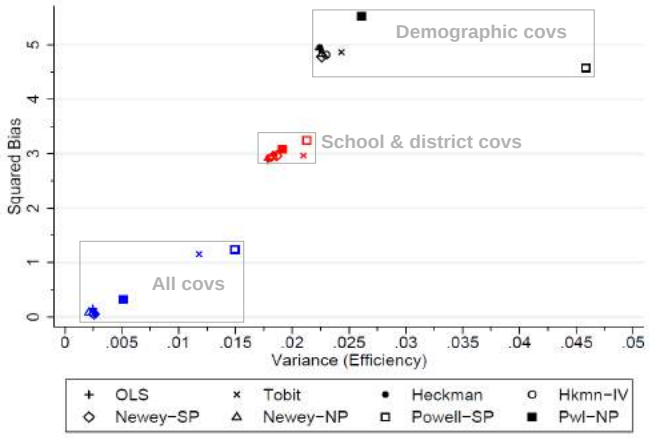
(a) Mean ACT Score



(b) Fraction College-Ready



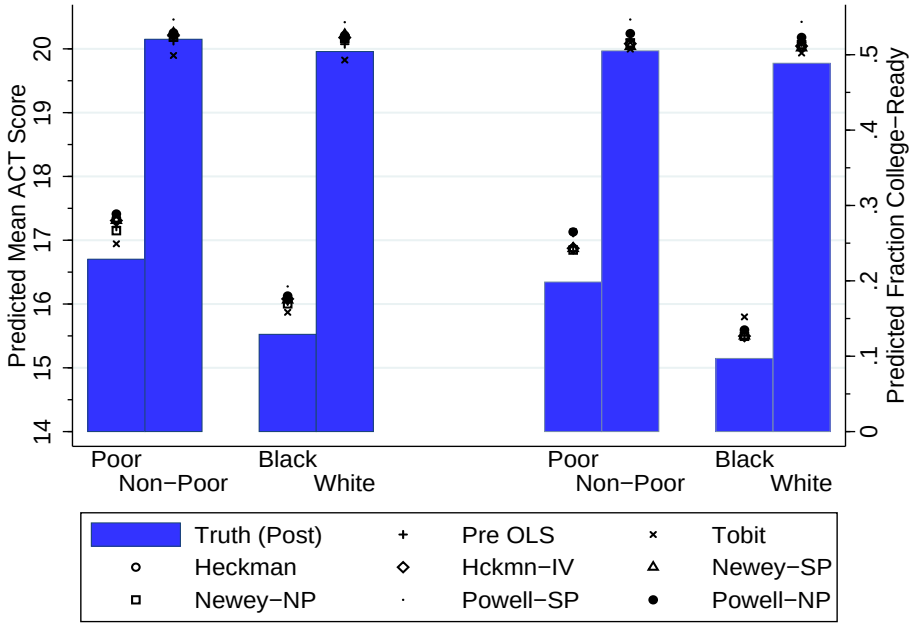
(c) Test Score Distribution



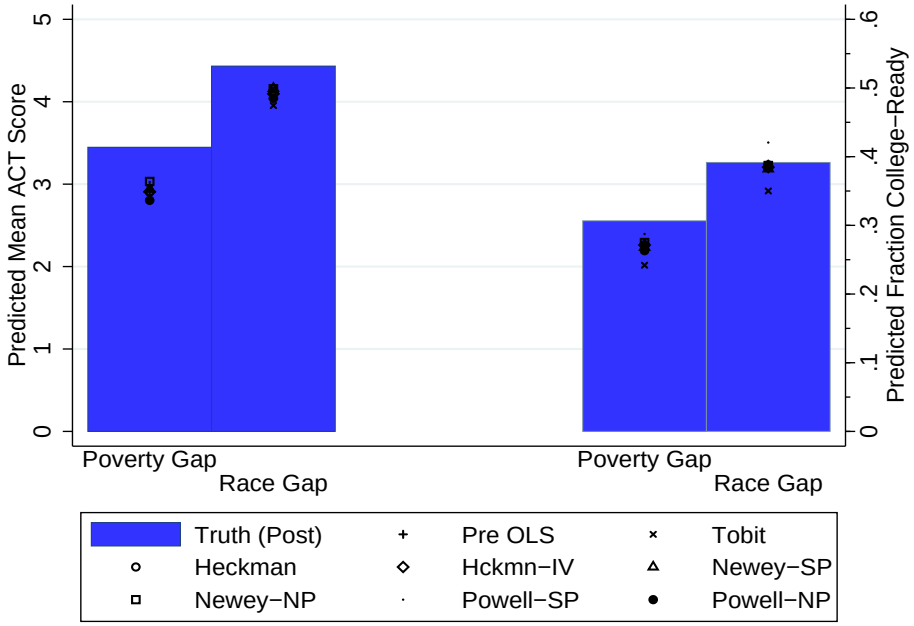
Notes: Figure shows the mean squared error of each combination of correction method and covariate set estimated without DFL weights. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. 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 IV: Score Gaps Compared to Post-Policy Distribution W/Out DFL Weights

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



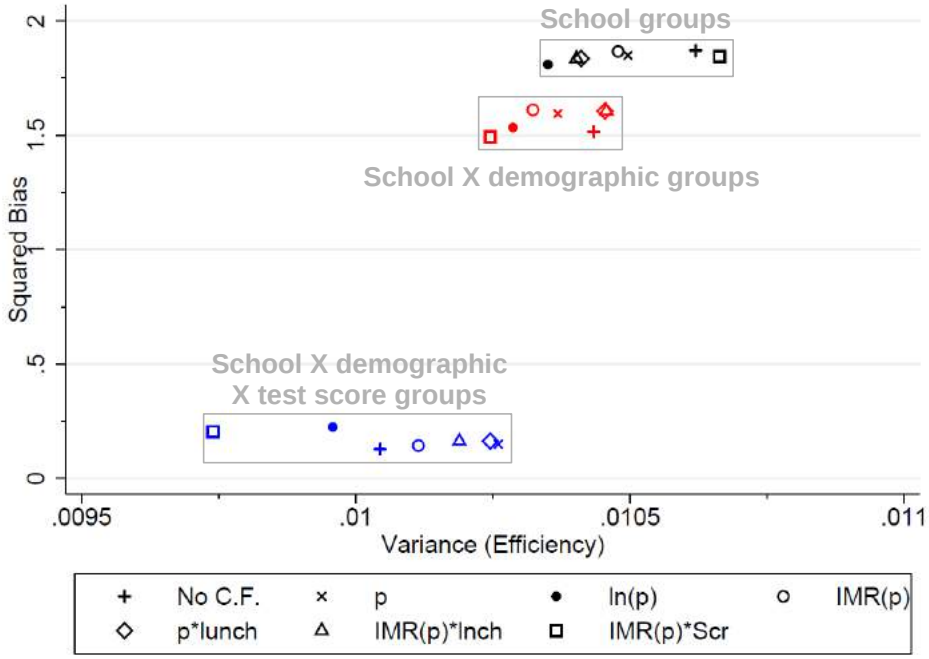
(b) Predicted Poverty and Race Gaps



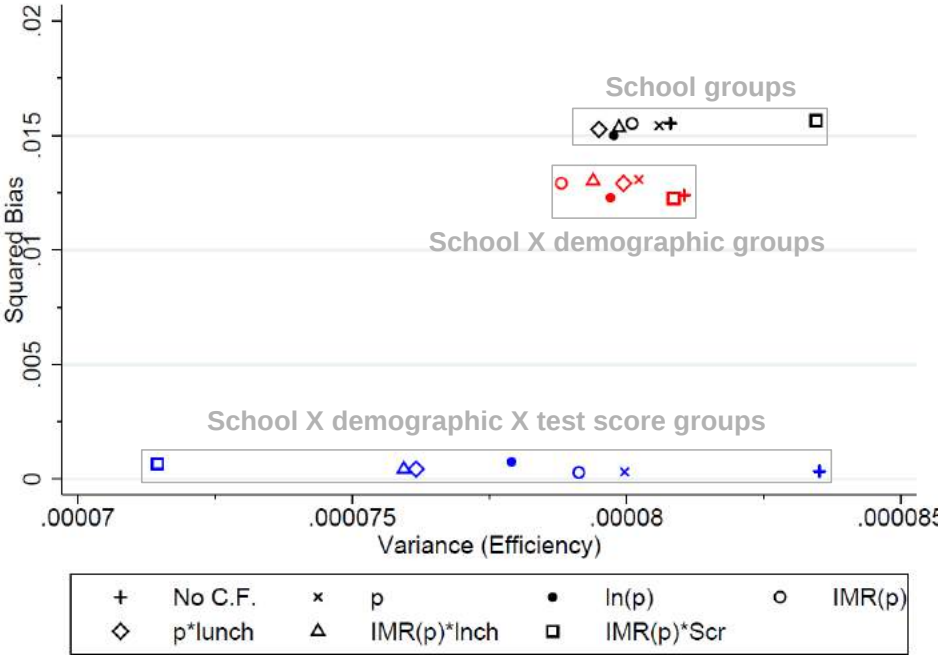
Notes: Figure (a) shows the "true" (using the fitted post-policy distribution without the DFL weights) and predicted mean latent ACT score and fraction college-ready across correction methods by poverty status and race. Figure (b) shows the "true" (using the fitted post-policy distribution without the 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 Figure V: Group-Level MSE Comparison W/Out DFL Weights

(a) Mean ACT Score



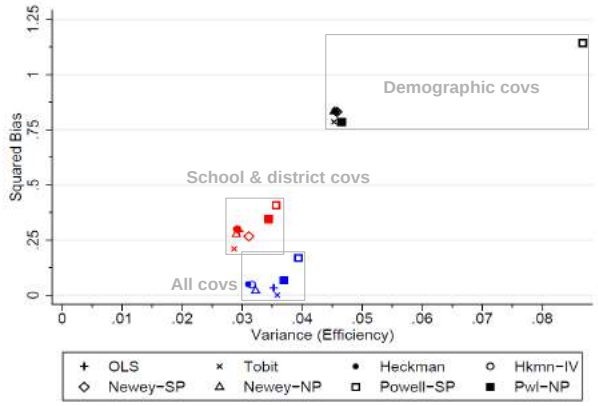
(b) Fraction College-Ready



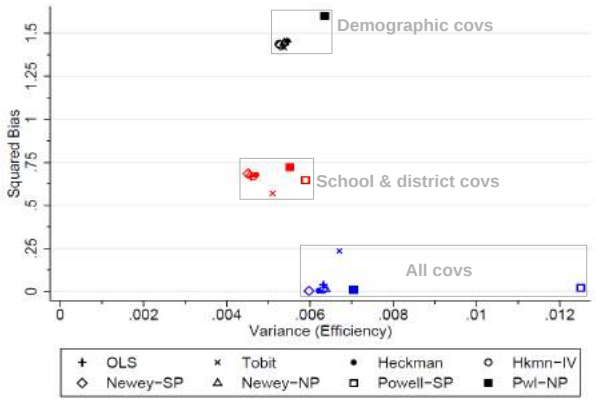
Notes: Figure shows the mean squared error of each combination of control function and data aggregation level for the group-level selection corrections from Table 5, fitting the post-policy distribution without DFL weights. Black (top of each figure): school-level; Red (middle): school*free lunch*minority-level; 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.

Appendix Figure VI: MSE Comparison by Race

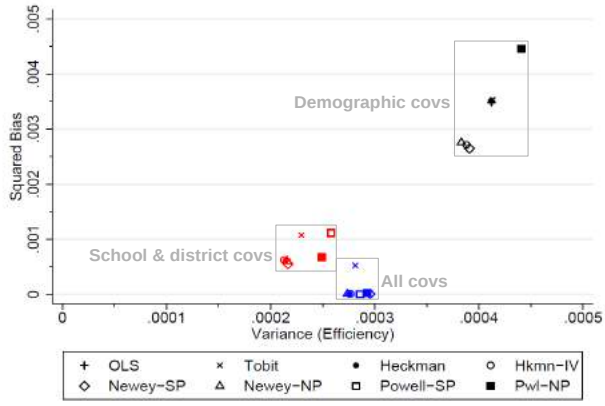
(a) Mean ACT Score - Black



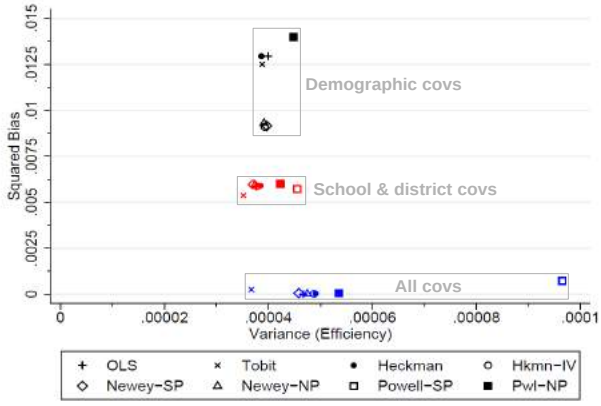
(b) Mean ACT Score - White



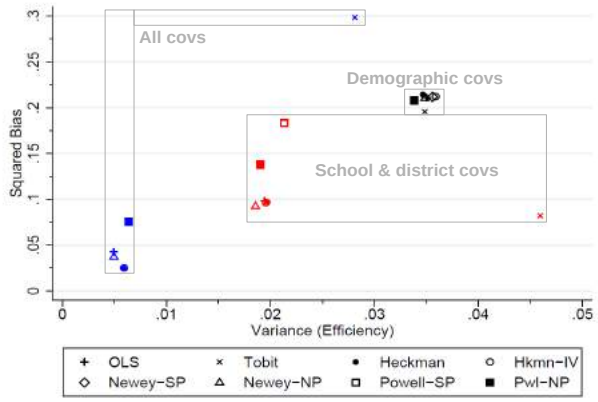
(c) Fraction College-Ready - Black



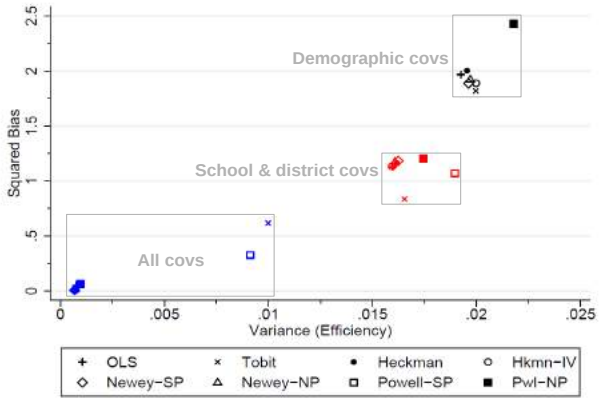
(d) Fraction College-Ready - White



(e) Test Score Distribution - Black



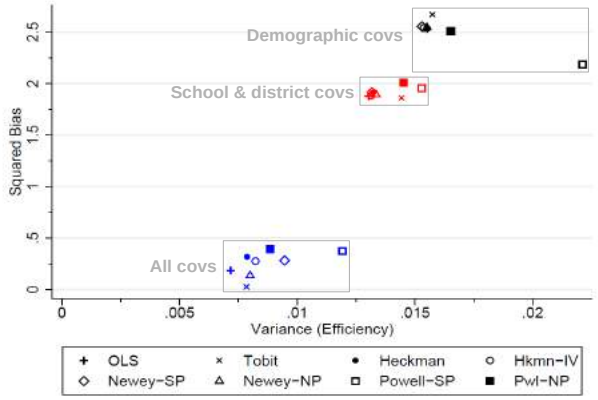
(f) Test Score Distribution - White



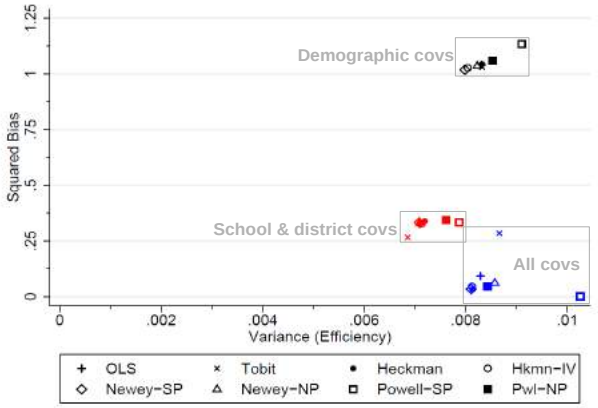
Notes: Figure shows the mean squared error of each combination of correction method and covariate set by race. Black (top of each figure): basic student demographics; Red (middle): plus school- and district-level covariates; Blue (bottom): plus student 8th and 11th grade test scores. Bias is the difference 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 VII: MSE Comparison by Poverty Status

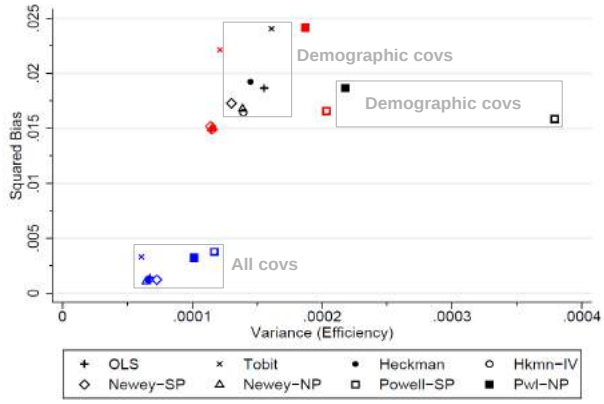
(a) Mean ACT Score - Poor



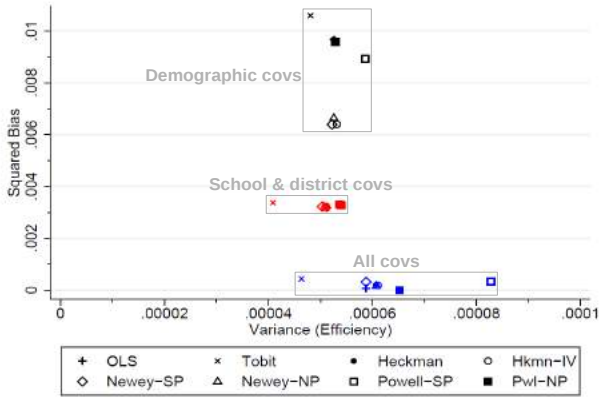
(b) Mean ACT Score - Non-Poor



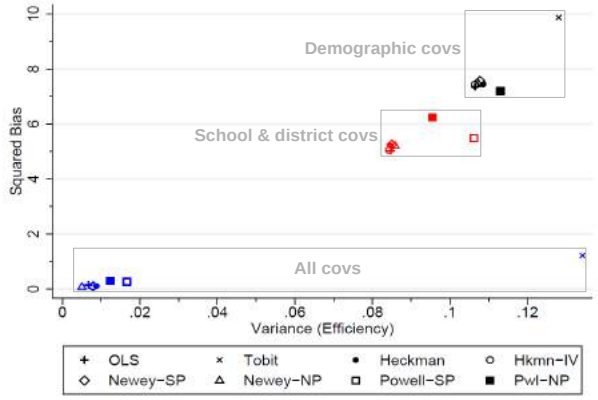
(c) Fraction College-Ready - Poor



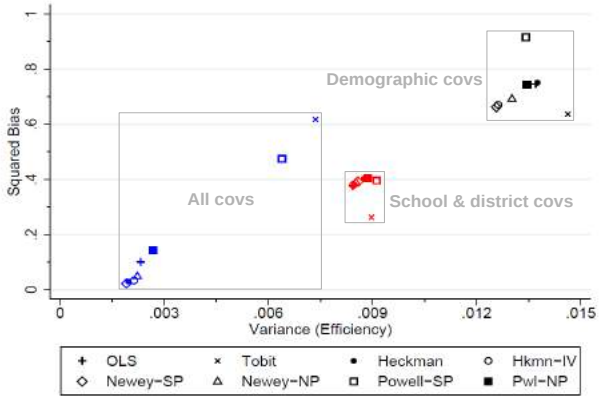
(d) Fraction College-Ready - Non-Poor



(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.