Essays on Educational Testing in an Era of Higher ("College-Ready") Standards

Essays on Educational Testing in an Era of Higher ("College-Ready") Standards

Yi Xe Thng

Andrew Dean Ho
Luke W. Miratrix
Eric Taylor

A Thesis Presented to the Faculty
of the Graduate School of Education of Harvard University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Education

2019

Acknowledgements

I am grateful to my advisor, Andrew Ho, for his relentless commitment to my success, his enthusiasm as I explore various dissertation projects, and his confidence in my ability to pull them together. Andrew gave me the opportunity to work on the project that eventually led to Essay 1. He has shared much sound advice with me, but one particularly has guided and encouraged me throughout the dissertation process: "listen to the data and don't be afraid of the answers that you find".

I am also thankful for my committee members, Eric Taylor and Luke Miratrix, who provided much constructive feedback and support that I needed to progress through the work. From both Luke and Eric, I have learnt that more complex methods need not always be better, and simpler methods may sometimes help the audience to understand the work better. It has been a delightful experience working with this committee.

Special thanks also to Dan Koretz and Richard Murnane, who listened to my ideas and read drafts of research proposals in the early days when I was conceptualizing the dissertation projects. They gave thoughtful advice that eventually led to Essay 3.

Essay 2 would not have been possible without the SLEDS data access granted by the Minnesota Office of Higher Education. Thanks to Meredith Fergus, Kara Arzamendia, Katherine Edwards, Anita Larson, and the other educators from the public school system in Minnesota for helping me to understand the data and how exit exams work in Minnesota. I have also received generous financial support from the HGSE Dean's Summer Fellowship to conduct the research for Essay 2.

Contents

Essay 1 Tables, Figures, and Appendices

Essay 2 Tables, Figures, and Appendices

Essay 3 Tables, Figures, and Appendices

Abstract

I present three essays on educational testing in an era of "college-ready" standards.

My first essay evaluates evidence-based standard setting methods that select passing or “college-ready” cut scores using regression-based predictive relationships between test scores and college outcomes. I investigate the forms of evidence that can be derived using predictive methods, whether such evidence can enhance score interpretations, and if so, how such evidence can be used to inform standard setting. I find that to compensate for the poor predictive utility, cut scores derived from predictive methods may be overly stringent or lenient than the stringency required of the standard. This may result in "college-ready" cut scores that are higher than warranted.

My second essay uses the case of Minnesota which set a high passing standard for its math high school exit exam, but later waived the passing requirement for obtaining a high school diploma and required failing students to take remediation. I investigate the impact of barely failing on students’ high school and college outcomes. I find some evidence that within this context, there may be a cohort-dependent impact on on-time high school graduation and enrollment in 4-year colleges for students who score barely below versus barely above the math passing score. If the second essay shows that the passing score has consequences, the first study may help to advance wiser selection of cut scores.

It is well-documented that gains on high-stakes state tests for low-income children and racial minority children are not matched on state-level audit tests such as the NAEP (Ho, 2007) or other low-stakes tests in districts (Jacob, 2005). This raises concerns about

the generalizability of findings from one test used to another. Using the case of Texas where the curriculum standards stayed the same but the newly introduced high-stakes test focused more on "college-readiness" standards, my third essay investigates whether measured district-SES score gaps change when the test changes. I find that after the assessment focused more on "college-readiness" standards, district-SES score gaps between the 90th and 10th district-SES percentiles widen slightly, but are of smaller magnitude than previously found using low-stakes audit tests for students.

Introduction

Educational testing is used as a tool to motivate student and system efforts towards improving test scores as a proxy for improving learning and education (Haertel, 2013). High-stakes testing changes the behavior of actors in the system and has both intended effects and unintended consequences (Koretz, 2013). Gathering evidence on the consequences of test use is one of the five sources of evidence recommended by the *Standards for Educational and Psychological Testing* to support the proposed interpretation of test scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Amidst recent educational reforms aimed at developing students who are college-ready by the end of high school (American Diploma Project, 2004; Jerald, 2008; U.S. Department of Education, 2010), it is not surprising that these reforms are closely tied to assessments. In particular, states are called to develop more rigorous "college-readiness" standards that reflect the knowledge and skills that students require in order to be successful in college. One implication of upgrading standards is that existing standards may be low, and a substantial proportion of students may not be able yet to meet the higher standards. It would take time for the system and students to catch up to these higher standards. Within such a policy environment, I present three essays that broadly aim to identify the consequences of educational testing at a time of when educational systems are shifting towards higher educational standards.

The first essay explores the use and development of cut scores in testing. It investigates predictive standard setting and more generally empirical-based standard setting. Both are relatively new standard setting methods developed to set standards that

purport to indicate whether students are "on-track" to college readiness. One implication of setting "rigorous" college-readiness standards is that they are often geared to move students towards higher levels of performance, and are hence of a higher level of stringency compared to existing standards. Such levels of performance would typically be above the average level of performance in the student population.

One finding from my study is that when the correlation between test scores and the predicted outcome scores is less than unity, predictive standard setting will yield cut scores that are overly stringent or lenient to compensate for the poor predictive utility. When predictive standard setting is used to identify cut scores for targeted outcome performance that is above-average, the resulting cut scores will be even more stringent than the targeted level of performance. In other words, college readiness cut scores will be more stringent than existing standards because (i) that is what is called for and implied by setting more rigorous standards, but also because (ii) the short-comings of predictive standard setting will further elevate the stringency of the identified cut score. Both these trends will result in "college-readiness" standards that are far more stringent than existing standards. One consequence of setting such (overly) stringent standards is that the percentage of students who fail to meet the standards will be higher than when standards were previously set lower.

My second study is situated in such a context, where the passing score on a newly introduced math high school exit exam in Minnesota is set very high, such that the percentage of students who cannot pass the test is higher than that under the old test. Although students were originally required to pass the exam in order to graduate from

high school, the state eventually waived the passing requirement after the first cohort of students sat for the test.

Once the cut score is set, it introduces a discontinuity that divides the test score distribution into distinct categories. From a skills perspective, there should not be any differences, on average, between students who score barely above or below the cut score. Any consequences of barely missing the cut score thus represents an effect of the testing policy (Papay, Murnane, & Willett, 2010). Within the context where a high passing standard is set, but the stakes for students are made more lenient (in the form of removing the passing requirement in order for students to be eligible for a high school diploma), I ask whether we still observe consequences of scoring barely below versus barely above a cut score. I find some evidence that in some cohorts, scoring barely below the math passing score has an impact on on-time high school graduation and 4-year college enrollment.

My third essay looks at another type of discontinuity: whether there are disruptions in test score trends when the assessment program changes. Under No Child Left Behind, states are required to use assessments to measure student performance in reading and math. State assessments should presumably measure what students know and can do broadly in the subject domain, as opposed to that on a specific test. Large-scale trends and gaps should not differ substantially when the measurement tool is changed provided that both are aligned to relevant content standards. Yet, changes to assessment programs have been associated with drops in nationally norm-referenced performance, arguably because of an over-focus on test-taking rather than learning improvements (Koretz, Linn, Dunbar, & Shepard, 1991).

The third essay is situated within the context where Texas changed its assessment to one that emphasizes college-readiness. In the context of a shift towards a more rigorous assessment and performance standards, I ask whether there are differences in district-SES gaps in performance measured using two different high-stakes tests in Texas. I find that after the assessment shifted towards measuring "college-readiness", the gap in performance between high-SES and low SES districts in the 90^{th} versus 10^{th} district-SES percentile widened slightly, but the magnitude is far smaller than the discrepancies found in student-level audit tests.

The last two essays contribute to evidence on whether high-takes testing causes unintended consequences at the student level or shapes unintended behavior at the district level respectively.

Essay 1

An Investigation of Methods to Incorporate Evidence into Standard Setting

An Investigation of Methods to Incorporate Evidence into Standard Setting

I. Introduction

Rigorous college-readiness standards have been the focus of recent educational reform (U.S. Department of Education, 2010), in particular the Common Core State Standards (Common Core State Standards Initiative, 2010). These standards "must be based on evidence regarding what students must know and be able to do at each grade level to be on track to graduate from high school college- and career-ready" (U.S. Department of Education, 2010, p.8). The emphasis on predictive evidence that addresses whether a student is "on track" has motivated so-called "evidence-based methods" for setting performance standards (McClarty, Way, Porter, Beimers, & Miles, 2013), including predictive standard setting (ACT, 2004; Kobrin, 2007).

Benchmarks set by predictive standard setting largely take the form, "a student with current score *x* has a *p%* probability of exceeding a future score *y*." For example, the SAT defines its benchmark score as a score that "predict(s) a 65% probability or higher of getting a first-year college grade point average of either 2.7 or higher (approximately a B average)" (Kobrin, 2007, p.2). For the ACT, "the ACT College Readiness Benchmarks represent the level of achievement required for students to have a high probability of success (a 75 percent chance of earning a course grade of C or better, a 50 percent chance of earning a B or better) in such credit-bearing college courses as English Composition, Algebra, and Biology. The benchmarks correspond to ACT Assessment scores on the English, Mathematics, and Science tests, respectively" (ACT, 2004).

These methods contrast with traditional standard setting methods such as the Angoff method and bookmark method. Traditional standard setting methods rely

predominantly on the judgment of standard setting panels regarding what students should know and be able to do at a particular performance level. In judgmental-based standard setting, the performance of students with respect to an external criterion is occasionally used post-hoc as external validity evidence to check the reasonableness of the performance standard set. In predictive standard setting, the external criterion is incorporated directly as an outcome for predictive models.

The emphasis on external criterion-related validity evidence forms the basis for the validity argument for evidence-based standard setting (EBSS) (McClarty et al., 2013). Evidence-based standard setting systematically collects cut scores identified by empirical methods, including predictive standard setting, using various sources of external evidence to identify neighborhoods where candidate cut scores fall. According to the EBSS argument, convergence of candidate cut scores in a particular region provides stronger validity evidence to support setting the eventual cut score in that region.

Predictive cut scores and evidence-based cut scores seem appealing because their basis in predictive and empirical evidence makes them appear less arbitrary than cut scores set by traditional judgmental-based standard setting methods. Predictive cut scores also provide an interpretation of the cut score that is connected to a meaningful criterion outcome, such as college-readiness, making it seem more interpretable and relevant (Beaton, Linn, & Bohrnstedt, 2012).

Even though predictive standard setting appears more evidence-based and objective than judgmental-based standard setting, judgment is still required for a number of inputs to the method. These include selecting the prediction method, selecting the probability associated with the prediction, selecting the outcome measure and criterion

score, and forming the analytic sample. Additionally, for evidence-based standard setting, judgment is still required in cases where there is no clear agreement of cut scores across different outcomes, and to select the final cut score within the neighborhood range.

Ho (2012) demonstrates that predictive cut scores are a function of both the stringency of the target outcome score, as well as the correlation between the test on which cut scores are set and the outcome test. Thus, students may be required to reach a higher or lower cut score not because of the standard required, but because of a stronger or weaker predictive relationship between the two tests. Even though predictive cut scores have been around for some time (see ACT, 2004; Kobrin, 2007; McClarty et al., 2013) these problems have not been well-identified.

In this paper, I discuss issues associated with predictive standard setting and by extension, evidence-based standard setting. I then investigate an alternative that attempts to combine the strength of judgmental-based standard setting and predictive standard setting. In this alternative, cut scores are first set by judgmental-based methods. Then, empirically based statements generated by predictive methods can be attached post-hoc to the identified cut scores.

II. Research Questions

The overarching research question in this study concerns the use of empirical methods in standard setting. In research question (RQ) 1, I first conduct an empirical investigation, before using the results to make a conceptual investigation in RQ2. I state RQ1 here and will clarify them in section (V) after explaining the terminology and predictive methods.

RQ1a: How do predictive cut scores deviate from a stringency-only equipercentile cut score when focal-outcome test correlations vary?

RQ1b: How dependent are predictive cut scores on judgments made regarding inputs to the predictive method? Specifically,

RQ1bi: How much do predictive cut scores vary using different predictive methods?

RQ1bii: How much do predictive cut scores vary across different probabilities/ quantiles of prediction?

RQ1biii: How much do predictive cut scores vary across different criterion scores?

RQ1c: How do empirical cut scores compare, using impact data and misclassification rates?

RQ2a: Can predictive standard setting be used as a stand-alone standard setting method?

RQ2b: What are the forms of empirically-based statements that can accompany any score on the predictor test regarding a future outcome score? What is the suitability of attaching empirically-based statements to judgmental-based cut scores as a standard setting method?

III. Background and Terminology

Modern standards-based reform efforts include content standards (the skills and knowledge that test-takers are to acquire); tests that measure these standards; and performance standards (levels of competence that test-takers can or should achieve). Performance standards are operationalized by cut scores on the test score scale (Kane,

1994), and performance at or in each level is described by "performance level descriptors" (Hambleton, Pitoniak, & Copella, 2012). Cut scores indicate the minimum score that test-takers must obtain on a test to demonstrate that they meet the performance standards. Setting cut scores is the goal of standard setting methods (Cizek, 2012) and is the focus of this paper.

I refer to predictive standard setting as an umbrella of methods that use regression as the basis for prediction. I refer to the test on which the cut score is set as the focal test, or the predictor test. The focal test scores are used as the predictor variable in the regression model. I refer to the test that measures the eventual outcome of interest as the outcome test. The outcome test scores are used as the dependent, or predicted variable. In the case of college-readiness standards, the ultimate outcome is often some form of college-level grade, such as first-year grade point average (first-year GPA) or course grade on a college-level course. I refer to the level of performance expected on the outcome test as the criterion score, e.g. B+ or C. The predictive cut score is the minimum score on the focal test identified using a regression-based prediction method that predicts the criterion score on the outcome test.

I use the term empirical methods to refer to prescribed procedures that use empirical data and a quantitative method to identify a cut score. Empirical methods may differ by the quantitative method, such as a regression-based method (ordinary least squares (OLS) regression, logistic regression or quantile regression), or an equipercentile linking method. The studies in which empirical methods are applied to identify an empirical cut score may differ by the external evidence used, such as concurrent studies (measurements taken from the students around the same time, or concurrently, as the

focal test) or predictive studies (measurements taken from the students at a future time, so the focal test is used to predict the future performance). I define predictive standard setting by the use of regression-based methods, which may use either concurrent or future outcomes.

I use the terms identified cut score or candidate cut score to refer to the cut score derived through an empirical method, that is under consideration as a potential cut score, but which has not yet been put forward as the "final" recommended cut score by the standard setting panel.

In the empirical investigations, I use the following notation: Let X and Y be the distribution of test scores on the focal test and outcome test respectively, X_c be the identified cut score on the focal test, and Y_c be the criterion score on the outcome test.

IV. Empirical Methods for Identifying Cut Scores

I discuss two empirical methods that can be used to identify cut scores: a general set of regression-based methods and a contrasting equipercentile linking method. Regression-based methods are the basis of predictive standard setting. I present the equipercentile linking method as a method that maintains "equal stringency" rather than "best prediction."

IV.1 Regression-based predictive methods

OLS regression. OLS regression regresses outcome test scores on the focal test scores:

$$OutcomeScore = \beta_0 + \beta_1 Predi\ ctorScore + \varepsilon \quad (1)$$

For a given criterion score, Y_c, that is the targeted level of performance on the outcome test, the predictive cut score, X_c, on the focal test can be calculated. For a given

pair of focal test scores and outcome test scores, and a given criterion score, there is only one score on the focal test that predicts the criterion score. There is a linear relationship between (Y_c, X_c) that satisfies the above regression equation (see Appendix A for depicting diagram).

Logistic regression. Logistic regression predicts the probability p of scoring at or above the criterion score on the outcome test. The corresponding logit function is:

$$l\,n\left(\frac{P(OutcomeScore \geq TargetScore)}{1-P(OutcomeScore \geq TargetScore)}\right) = \beta_0 + \beta_1 Predi\ ctorScore \quad (2)$$

The predictive cut score is calculated for a given criterion score and probability. I first look at a range of probabilities (50%, 65%, 75%) (LR50, LR65, and LR75) that have been used as acceptable margins of accuracy over a range of standardized scores on the outcome test (ACT, 2004; Kobrin, 2007).

For a given pair of focal test X and outcome test Y and a given criterion score Y_c, there is a joint distribution of probability p and predictive cut score, X_c, (p, X_c) that satisfies the above logit function. Similarly, for a given probability p of scoring above the criterion score Y_c, where there is a linear relationship between (Y_c, X_c) that satisfies the linear equation, here, for a given predictive cut score, X_c, there is a non-linear relationship between (p, Y_c) that satisfies the equation. To identify a cut score X_c, the probability p has to be specified. The converse is also true, we can fix X_c and solve for the probability p.

Quantile regression. McClarty, Murphy, Keng, Turhan, and Tong (2012) evaluate quantile regression as a method for predictive standard setting. Whereas OLS regression estimates the conditional mean of the outcome test score, quantile regression (Koenker & Hallock, 2001) estimates the conditional median or some other specified quantile:

$$Q(q|x_i) = \beta_0 + \beta_1 Predi\ ctorScore \quad (3)$$

where $Q(\mathrm{q}|x_i)$ is the q^{th} quantile of outcome test score. Quantile regression is more robust to outliers in the outcome score than OLS regression. McClarty et al. (2012) looks at quantiles over 40th, 50th, and 60th percentile, but quantile regression has so far not been used in the literature for predictive standard setting. To facilitate comparison to logistic regression, I look at quantiles "numerically equivalent" to the probabilities of scoring above a criterion score in logistic regression, where for the (1-q)th quantile regression, q% of students score above the criterion score i.e., 50th, 35th, and 25th quantiles (QR50, QR35, and QR25).

For a given pair of focal test and outcome test and a given criterion score, there is a joint distribution of (q, X_c) that satisfies the above quantile regression equation, i.e., we can specify any quantile q and identify the corresponding cut score X_c. The converse is also true: we can fix a particular X_c and solve for the corresponding quantile q. Similarly, for a given quantile of performance, there is a joint distribution of (Y_c, X_c) that satisfies the equation; for a given predictive cut score, there is a joint distribution of (q, Y_c) that satisfies the equation.

IV.2 Equipercentile method

The equipercentile method (Lord, 1955) identifies the cut score on the focal test that has equivalent percentile rank as the criterion score on the outcome test. This is an expression of "equal stringency" that does not depend on the relationship between the tests nor the construct each of them measures. I refer to the stringency level of a performance standard broadly as the difficulty level (see Phillips, 2014), measured by the percentage of students who are able to score above a given level on the test. In this sense,

the equipercentile cut score is used as a stringency-only reference point to predictive-based cut scores.

For a given pair of focal test and outcome test and a given criterion score, there is only one score on the focal test that has equal stringency as the criterion score. There is a typically non-linear relationship between (Y_c, X_c) that satisfies the equal stringency relationship.

V. Predictive Standard Setting

I refer to predictive standard setting as a procedure that uses regression-based methods for identifying cut scores on a test. The regression-based methods seek a score on the focal test that "predicts" a given criterion score on the outcome test. It typically takes the form of a predictive statement based on logistic regression: a student who scores at the cut score X_c on the focal test has p% probability of achieving the future score Y_c on the outcome test (see ACT, 2004; Kobrin, 2007).

Predictive standard setting has often been put forth as an objective, evidence-based alternative over traditional judgmental-based standard setting. However, predictive standard setting requires specification of a number of inputs to the predictive equation. In this section, I present a conceptual framework to investigate predictive standard setting. At the end of the section, I explain in detail the research questions (RQ1) for the empirical investigation.

V.1 Predictive standard setting: Conceptual framework

Dependency of predictive cut scores on focal-outcome test correlations

Performance standards specify the level of performance required on a test in order for a test taker to be classified into a particular performance category (Cizek, 2012). The

level of stringency required should correspond to the level of knowledge, skills, or ability required to demonstrate the performance level. Predictive cut scores, however, are a function of both the stringency of the standard, and the predictive strength, i.e., correlation, between the predictor and outcome test (Ho, 2012).

As a point of reference, the equipercentile linking method establishes the score correspondence between X and Y that have equivalent percentile on either scale. Equipercentile cut scores (X_c^{equi}) identified this way have the same "passing" percentage on either scale, and hence reflect a "stringency-only" relationship (Ho, 2012). Where X and Y have standard bivariate normal distributions with correlation ρ_{XY}, Y_c and X_c^{equi} are related by:

$$Y_c = X_c^{equi} \quad (4)$$

In the case where OLS regression is used as the prediction method, the relationship between the OLS predictive cut score X_c^{OLS} and the criterion score Y_c is a function of the correlation ρ_{XY} between the two tests (Ho, 2012):

$$X_c^{OLS} = \frac{Y_c}{\rho_{XY}} \quad (5)$$

or

$$X_c^{OLS} = \frac{X_c^{equi}}{\rho_{XY}} \quad (6)$$

Whereas the equipercentile linking method identifies a cut score that reflects only stringency (equation 4), the OLS-regression based method identifies a cut score that deviates from the stringency-only cut score by a factor of $1/\rho_{XY}$ (equation 6). When the target outcome score is below average on the z-scale, the predictive cut score would be lower, or more lenient than the equipercentile cut score by a factor of $1/\rho_{XY}$. When the

target outcome score is above average on the z-scale, the predictive cut score would be higher, or more stringent than the equipercentile cut score by a factor of $1/\rho_{XY}$.

In the case of predicting college-readiness, the typical raw correlation between high school GPA or SAT scores and first-year college GPA is from +0.30 to +0.40 (Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Shaw, 2015). Based on the above relationships, the OLS regression-based cut score for these correlation ranges will be 2 to 3 times higher on the z-scale than the equipercentile-based cut score.

Cut scores derived by probabilistic regression methods are thus affected by a confounded relationship between the stringency of the standard and the predictive strength of the focal-outcome tests. If a student is not "on track" to a future cut score, it can be due to a combination of facts: a) the future standard is high; b) the predictor test does not predict the outcome; or both.

In summary, predictive standard setting results in cut scores that confound the level of knowledge and skills required of students to demonstrate "college-readiness" with the predictive strength of relationship between the focal-outcome tests. In other words, the predictive cut score is affected both by the level of student performance, and the joint relationship between the focal (predictor) and outcome tests. In this paper, I first empirically demonstrate the sensitivity of predictive cut scores to focal-outcome tests with different correlations under conditions of bivariate normality. Comparing this to the equipercentile cut score as a stringency-only reference point, I then ask how sensitive predictive cut scores are to focal-outcome tests with different correlations when X and Y depart from bivariate normality. I answer this under simulated conditions and in an actual empirical dataset.

Judgment required for inputs to predictive standard setting methods

Judgment is required for selecting the specific predictive standard setting method, and for selecting inputs to implement the method. First we need to identify the outcome measure and select a criterion score on the measure. Next, we need to select a prediction method to use. Finally, depending on the prediction method chosen, we need to select the type of summary statistic being modeled. For OLS regression, this is set at the mean. For logistic regression, we need to select the probability of scoring above the criterion score. For quantile regression, we need to select the conditional quantile of performance being modeled. I ask how sensitive predictive cut scores are when the above inputs are varied.

V.2 Empirical investigation of predictive standard setting (RQ1)

In this section, I discuss the empirical investigations that I conduct to answer RQ1. I conduct the investigations for RQ1a and RQ1b under the following scenarios: (1) under conditions of standard bivariate normality; (2) under simulated conditions with departures from standard bivariate normality; and (3) using an actual empirical dataset. I use Cohen's effect size guidelines to reference whether the magnitude of the difference between the identified predictive cut score and the stringency-only equipercentile cut score is small (±0.2), medium (±0.5), or large (±0.8) in terms of standard deviation units[1].

To answer RQ1c, I also evaluate the cut scores using impact data and classification accuracy (see also McClarty et al., 2012). Impact data refers to the percentage of students who will exceed the cut score. If the impact data (i.e. passing rate) is similar to what policymakers expect based on their point of reference to previous tests or other tests, then the cut score may be assumed to be reasonable (Hambleton &

[1] No extra calculation is required for the simulated data since the scale is already in standard deviation units for a bivariate standard normal distribution.

Pitoniak, 2006). Impact data is a commonly used statistic in the standard setting literature to evaluate cut scores (Hambleton & Pitoniak, 2006). Impact data is regularly reported in NAEP, and has been used by states in cut score studies (e.g. see Texas Education Agency, 2013).

Classification accuracy looks at the percentage of accurate classifications for a given cut score and criterion score. Classification accuracy is commonly used to evaluate cut scores in professional licensure and certification tests where criterion-referenced testing is more popular (Clauser, Margolis, & Case, 2006), but less commonly used for K-12 standard setting. With the increasing use of external criteria in empirical standard setting for college-readiness, classification accuracy might receive more attention in K-12 standard setting. Below, I adapt the discussion of indices for measuring accurate classifications from the professional licensure and certification test literature.

Within the classical test theory framework, a student whose level of proficiency meets the standard will have a true score at or above the cut score associated with the standard. The score that the student receives on the test used to measure proficiency is the observed test score. Classification errors occur when the proficiency classification based on the observed score does not match the proficiency classification based on the true score (see Appendix B for depicting diagram). Misclassification arises in two ways within this framework. False negative arises when students are classified as not proficient on the observed test when their true score in fact meets the proficiency standard. False positive arises when students are classified as proficient on the observed test when their true score falls below the proficiency standard.

Hanson and Brennan (2004) propose a set of indices to look at the size of classification errors (or misclassification rates) relative to the population. The false-positive rate (fp) refers to the proportion of students who are not proficient but received passing status on the test; the false-negative rate (fn) refers to the proportion of students who are proficient but received failing status on the test. In this study, I calculate the misclassification rate as the sum of the false positive rate and the false negative rate.

V.3 Predictive standard setting: Empirical investigation methods (RQ1)

Data

Standard bivariate normal distributions. I start first by generating two variables, *X* and *Y* to respectively represent focal test scores, the test on which standards are being set, and outcome test scores, the test on which future performance is being targeted with criterion scores. *X* and *Y* are generated as standard bivariate normal distributions with correlations between 0 and 1 (r = +0.9, +0.7, +0.5, and +0.3).

Departure from standard bivariate normality. Next, I simulate conditions that depart from standard bivariate normality by using log-normal distributions to approximate skewed distributions (see Ho & Yu, 2015; Reardon & Ho, 2015).

I follow the procedure described in Reardon and Ho (2015) where log-normal distributions are used to approximate skewed distributions. For our skewed distribution of interest, *X*, there is a log-normal distribution X^* that is normally distributed, $X^* \sim N(\mu, \sigma)$ where

$$X = e^{\mu+\sigma X^*} \quad (7)$$

The distribution of *X* has a skewness of $\gamma = sign(c)(e^{c^2}+2)\sqrt{(e^{c^2}-1)}$ determined by the parameter *c*. I generate a standard normal distribution of $X^* \sim N(0, 1)$ to derive the corresponding skewed distribution. By constraining

$$a = -si\ gn(c)(e^{c^2} - 1)^{-0.5}$$

$$b = si\ gn(c)\ (e^{2c^2} - e^{c^2})^{-0.5}$$

X will have a standardized (mean 0, variance 1) log-normal distribution. I generate distributions with skewness γ of ±0.10, ±0.30, and ±0.50, all well within the range of skewness statistics of ±0.5, which has been found to be rarely exceeded by K-12 state-level tests (Ho & Yu, 2015).

Empirical dataset. Finally, I use an empirical dataset containing City University of New York (CUNY) first-year GPAs[2] (outcome test) and New York Regents high school tests in math and ELA (focal tests) from nearly 50,000 students during graduation years 2011 and 2012. The test score distributions generally have a negative skew with magnitude of skew lower than 0.5. The Regents high school ELA has greater skew around -0.80. The first-year GPA distribution has a bimodal peak that includes modes at a non-zero value as well as a mode at 0.

In the analyses, I use the original first-year GPA scale to retain its meaning and standardized the Regents high school score scales to have a mean of 0 and standard deviation of 1.

Methods

I use the empirical methods described under Section IV to identify predictive cut scores.

To set predictive cut scores where the data departs from bivariate normality, the usual procedures of checking and dealing with model misfit can apply. Hence, if

[2] Student's weighted freshman GPA at CUNY. The total number of grade points earned by each student is divided by the total number of credits attempted during their freshman year. Only non-remedial courses and courses that provided grades are included in the calculations.

distributions are skewed, they could be transformed so that the regression residuals are normally distributed. I do not perform these transformations here because, in the case of a generalized logarithmic transformation as in Equation (7) above, they would lead back to the standard normal bivariate distributions. Thus, this study illustrates a situation where regression models are applied without standard diagnostic checks. Although this would represent poor statistical practice, imperfect model fitting methods are often used in operational settings for consistency and transparency. Even assuming that flexible predictive models are not used, untransformed distributions illustrate how nonstandard distributions affect the predictive cut score. This can inform how cut scores would be affected for situations where the departures are not serious enough to violate the regression residuals normality assumptions and invoke transformations to the affected variable.

Where applicable, I pay special attention to results for data with correlations of r= +0.3 or r= +0.5. These correlations are typical for that observed between college first-year GPA with high school GPA and college admissions tests (ACT, 2007; Kobrin, 2007).

V.4 Predictive standard setting: Results

V.4.1 Results under standard bivariate normality

RQ1a: How do predictive cut scores deviate from a stringency-only equipercentile cut score when focal-outcome test correlations vary?

Figure 1 and Table 1 shows the cut scores identified by each predictive method across a range of criterion scores on the outcome test when the focal-outcome test correlation changes. They also show the equipercentile cut score as a stringency-only

reference. For clarity of presentation, I focus on predictive methods that model some measure of central tendency in the outcome variable – OLS regression, logistic regression that predicts 50% probability of scoring above the criterion score (LR50), and quantile regression that predicts the median performance in outcome (QR50).

Stringency-only reference cut score. Figure 1 Panel A and Table 1 Panel A show that when the equipercentile method is used, the identified cut score on the z-scale is identical (within sampling limits) to the criterion score, regardless of the correlation between the focal test and outcome test. In the rest of the discussion for results under standard bivariate normality, I take this as the stringency-only reference point to compare other cut scores.

Criterion score at average of outcome distribution. Figure 1 Panels B-D show that when the criterion score is targeted at the average of the outcome test distribution, the identified cut score is also around the average of the focal test distribution for all the methods that model some form of central tendency – OLS regression, logistic regression with 50% probability, and quantile regression at the median – regardless of the correlation between the focal test and outcome test.

Criterion score above (below) average of outcome distribution. When the criterion score is targeted to be above (below) the average of outcome distribution, the regression-based methods will identify cut scores that deviate above (below) the stringency-only equipercentile-based cut score. In the case of OLS regression, the identified cut scores deviate by a factor of $1/\rho_{XY}$ as predicted by equation 6. When the strength of the focal-outcome test correlation drops, the severity of the deviation increases. This is because to compensate for the poor test prediction, the cut score has to

be set even more stringent (lenient) than the level of performance expected by the criterion score.

RQ1bi: How much do predictive cut scores vary using different predictive methods?

Table 2 and Figure 2 groups the results from Table 1 to show how predictive cut scores vary by prediction method. In general, the predictive cut scores identified by OLS regression and median quantile regression are very similar to each other, since there are no outliers in a standard normal distribution to skew the conditional outcome distributions.

Holding the test correlation and criterion score fixed, the predictive cut scores identified by regression-based methods that model some form of mean outcome (OLS), or median outcome (QR50) differs somewhat from that for logistic regression which models 50% probability of achieving the outcome (LR50). Using Cohen's *d* as a guide to the size of difference between the predictive cut scores, at r=0.7 and above, the difference in predictive cut scores for each prediction method is very small. At r=0.5 and below, the difference in predictive scores is also very small when the criterion score is within ±1 standard deviation units of the average outcome score. At r=0.5, the difference in predictive scores is small-sized for criterion scores between 1 to 2 standard deviation units away from the average outcome score, and medium-sized for r=0.3 or below.

RQ1bii: How much do predictive cut scores vary across different probabilities of prediction?

Table 3 and Figure 3 show that when logistic regression is used, predictive cut scores vary across different specified probabilities of scoring above the criterion score.

As expected, the stringency of the logistic regression cut score increases as the specified probability of scoring above the criterion score increases.

At correlation r=0.5, the difference in predictive cut scores for probability set at 75% versus 65% is about 0.5 standard deviation units when criterion scores are targeted within ±1 standard deviation units of the average outcome score. At correlation r=0.3, the difference in predictive cut scores is about 1 standard deviation units. These differences range from medium- to large-sized.

Table 4 and Figure 4 shows corresponding results when quantile regression is used. The difference in predictive cut scores predicted by 25th quantile regression and 35th quantile regression is about 0.5 standard deviation units at correlation r=0.5, and about 1 standard deviation units at correlation r=0.3 for criterion scores targeted within ±1 standard deviation units of the average outcome score. These differences range from medium- to large-sized.

RQ1biii: How much do predictive cut scores vary across different criterion scores?

The results presented in Table 1 show that across all predictive methods, for a given change in targeted criterion score, the predictive cut score increases by more than that magnitude on the z-scale for all prediction methods when the focal-outcome test correlation is less than 1. At lower correlations, the difference in predictive cut score for a given change in targeted criterion score is wider than that at higher correlations.

RQ1c: How do empirical cut scores compare, using impact data and misclassification rates?

Finally, I evaluate the predictive cut scores and equipercentile cut scores using two approaches commonly used to provide external validity evidence in standard setting.

I focus our discussion for the case when correlation r=0.3 but the results are generally true for correlations below unity. To recap, we found earlier that at r=0.3 (Figure 2 Panel D), the predictive cut score is lower than the equipercentile cut score when the criterion score is below average of the outcome distribution, and higher than the equipercentile cut score when the criterion score is above average.

Figure 5 shows the impact data over the range of criterion scores when cut scores are identified using the equipercentile method, and the regression-based methods that measure some form of central tendency (OLS, LR50, or QR50). When the criterion score is targeted above the average of the outcome score distribution, the percentage of passing students for the corresponding predictive cut scores are lower than that for the corresponding equipercentile cut score. This happens because predictive cut scores are set more stringent than the equipercentile cut score in order to compensate for the low predictive utility of the focal test scale. Conversely, when the criterion score is targeted below average of the outcome score distribution, the percentage of passing students is higher when predictive methods are used than when the equipercentile method is used, because the former sets cut scores that are more lenient.

Figure 6 shows the misclassification rates for the corresponding cut scores over the range of criterion scores. The misclassification rate for equipercentile cut scores is higher for all criterion scores, other than that targeted at the average of the outcome distribution. Again, this is mainly a function of the relative location of the identified cut score. Figure 7 shows why this is so. Figure 7 shows how misclassification rates vary as a function of cut scores on the X score scale at different target values of the criterion score. Generally at low correlations, for criterion scores above average, the higher the cut score,

the lower the misclassification rate; for criterion scores below average, the lower the cut score, the lower the misclassification rate. Since predictive cut scores are higher than equipercentile cut scores for criterion scores above average and vice versa for criterion scores below average, the misclassification rates for predictive cut scores are always lower than those for equipercentile cut scores. However, as Figure 2 Panel D reminds us, the lower misclassification rates for predictive cut scores come at the expense of overly stringent or lenient cut scores compared to the stringency-only equipercentile cut scores.

V.4.2 Results under simulated departures from standard bivariate normality

In this section, I present identified predictive cut scores when the focal test scores depart from standard normal distributional assumptions. These results are illustrative in nature, because proper application of regression models require initial data exploration and alternative model specifications to deal with non-linear and non-normal relationships.

Skewness. Figure 8 shows how predictive cut scores and equipercentile cut scores vary when the focal test score distribution varies from negative skewness of -0.50 to +0.50 for correlation values of r=0.3 and r=0.5.

In general, even as the skewness of the distributions vary, the broad findings under distributions with bivariate normality assumptions still hold true: predictive cut scores differ substantially from equipercentile cut scores as the focal-outcome test correlation drops; logistic regression and quantile regression cut scores depends on the specified probability of correct prediction or quantile of performance modeled, and the targeted criterion score. Hence, I focus the discussion on how the cut scores vary within prediction method and specified probability/quantile or criterion score as skewness varies.

The results in Figure 8 Panels A and B show that equipercentile cut scores are sensitive to skewness in the focal test score distribution. This occurs because skewness in the focal test score distribution shifts the density of the distribution relative to the outcome score distribution. Even so, the differences in equipercentile cut scores are very small (compared to equipercentile cut scores from a non-skewed distribution) when the criterion score is within ±1 standard deviation units of the average outcome score over skewness ranges of γ within ±0.50.

Both OLS and quantile regression cut scores are robust to differences in skewness of the focal test score distribution, since the outcome score distribution is still normally distributed over all values of focal test scores.

Logistic regression cut scores are more sensitive to skewness in the focal test score distribution than OLS and quantile regression methods. When the probability is set at 50%, and the criterion score is within ±1 standard deviation units of the average outcome score, the differences in LR50 cut scores are very small (compared to logistic regression cut scores from a non-skewed distribution) as skewness values vary over the range within ±0.50. When the probability is set at 65% or 75%, the differences in logistic regression cut scores for skewed distributions beyond ±0.30 (compared to logistic regression cut scores from a non-skewed distribution) are still small if the criterion score is within ±1 standard deviation units of the average outcome score and reach medium-sized if the criterion score lies within 1 to 2 standard deviation units above or below the average outcome score. Data tables are available upon request.

V.4.3 Results using empirical data

Table 5 and Table 6 show the equipercentile and predictive cut scores identified using 2010 Regents high school Math and ELA score as the focal test respectively, and first-year GPA as the outcome score for criterion scores between 1.7 to 3.3, corresponding to grades between C-to B+. These grades correspond to above average first-year GPA scores in the empirical dataset, except for C and C-, which are slightly below average. The results show a similar pattern as that observed in the simulated dataset. I illustrate the magnitude of the differences in empirical cut scores using Regents high school math as an example, but the patterns apply for Regents high school ELA as well.

When criterion scores are targeted above the average outcome score, the predictive cut scores are more stringent than that of equipercentile cut scores. For a criterion score targeted at B+, the predictive cut scores are at least 0.90 standard deviation units higher than the equipercentile cut scores when the predictive method models some measure of central tendency in the outcome score (OLS, LR50 or QR50). In the empirical dataset, the OLS cut score is quite different from the median quantile regression cut score because the score distributions depart from bivariate normality.

For differences in specified probabilities (50%, 65% or 75%) or quantiles (50^{th}, 35^{th}, or 25^{th} percentile) in logistic regression and quantile regression respectively, the differences in cut score can range between 0.40 to 0.80 standard deviation units from one specified probability/quantile to the next.

Over the ranges of criterion scores from C- to B+, the difference in equipercentile cut score for each half step grade difference (e.g. C- to C, B to B+) is between 0.25 to 0.5 standard deviation units; the difference in identified equipercentile cut score for a

criterion score within the range of C- and B+ is about 1.7 standard deviation units. For predictive cut scores, each half step grade difference in criterion score can result in differences in identified predictive cut score between 0.5 to 1.0 standard deviation units. When criterion scores range between C- and B+, the predictive cut scores can reach a difference of about 2.0 to 4.0 standard deviation units.

V.5 Predictive standard setting: Discussion

RQ2a: What is the suitability of predictive standard setting as a stand-alone prediction method?

In this study, we find that regression-based predictive cut scores differ substantially from equipercentile cut scores that provide a "stringency-only" reference when focal-outcome test score correlations are weak. This issue goes undetected when checks using misclassification rates are used. For a given criterion score, the corresponding predictive cut scores yield lower misclassification rates over corresponding equipercentile cut scores, but this is not surprising given the optimization problem that regression models solve. The cost of predictive accuracy is unnecessary stringency (or leniency) of cut scores as students are penalized (or given credit for) the imperfect predictive relationship of the test. Regression-based predictive cut scores can also differ substantially from each other depending on the specified probability of correct prediction (logistic regression) or quantile of performance (quantile regression) modeled, and the criterion score specified. These issues have not been widely studied in the standard setting literature. In this section, I discuss the implications of the above findings for predictive standard setting.

V.5.1 Dependency on stringency of performance and test correlations

The empirical results from RQ1a show that for all focal-outcome test correlations below unity, and at criterion scores targeted at levels other than the average of the outcome distribution, predictive cut scores deviate from the equipercentile cut score that represents a stringency-only performance level. The weaker the focal-outcome test correlations are, the greater the deviation from the stringency-only equipercentile cut score.

If the focal test used is imperfectly correlated to the outcome test, and the criterion score is above the average of the outcome score distribution, students would be expected to reach a higher level of performance to be classified as "college-ready" when predictive methods were used to identify cut scores than if the equipercentile method were used. There will be a group of students who would be classified differently depending on the empirical method used to set cut scores. The weaker the correlation of the focal-outcome test, the greater the percentage of affected students. Thus, there will be a group of students penalized due to the imperfect focal-outcome test correlation, a joint property of the tests, rather than because they cannot reach the level of performance required by the stringency of the criterion score.

Conversely, when the criterion score is below the average of the outcome score distribution, a group of students would be classified as "college-ready" by the predictive cut score not because they have met the stringency required of the criterion score, but because the predictive cut score is more lenient than it should be in order to compensate for the imperfect focal-outcome test correlation. Again, the weaker the correlation of the focal-outcome test, the greater the percentage of affected students. Thus, there will be a group of students who will miss out on opportunities, allocated on the basis of their

performance on the focal test relative to the cut score, to help them become "college-ready".

The above issues would not be detected if misclassification rate is used to evaluate the adequacy of the cut scores, because this statistic is driven by the relative location of the cut score with respect to the criterion score. In fact, misclassification rate statistics would appear more favorable for predictive cut scores at the expense of more stringent (or lenient) cut scores than the stringency-only standards.

It is somewhat surprising that the deviation of predictive cut scores away from a "stringency-only" cut score has not been picked up by the examination of impact data. Our simulations suggest that at low correlations, predictive cut scores would lead to more extreme passing rates than equipercentile cut scores, depending on the location of the criterion score. It is possible that other issues with the GPA scale commonly used to measure college-readiness, and representativeness of the analytic sample may mask the differences in predictive cut scores and equipercentile cut scores when correlations are low. I discuss these issues further in Appendix C.

V.5.2 Suitability of predictive standard setting as a stand-alone method

Predictive standard setting may not be suitable as a stand-alone standard setting method for a number of reasons. First, predictive standard setting may not be as objective as it seems. It has been put forth as an alternative over judgmental-based methods because it is perceived not to be judgment-driven. In Appendix C, I explore in greater detail how predictive standard setting requires judgments about the choice of outcome measure, analytic sample, and prediction method, as well as inputs, to the prediction equation including the criterion score and probability of achieving the criterion score (or

quantile of performance). The differences in identified predictive cut scores as a result of different decisions may often be non-trivial.

Second, predictive standard setting rests on the quality of the evidence. For the predictive models to work, the focal test score and outcome test score distributions should meet bivariate normality assumptions. When these assumptions are not met, the identified predictive cut score is misleading. Furthermore, the lack of a well-defined construct measured on a well-defined scale (see Appendix C.4 for further details), as well as issues in constructing a representative analytic sample (see Appendix C.5 for further details) may also affect the level of the identified cut score.

Third, predictive standard setting is conceptually incoherent with the purposes of standard setting. The level of performance identified through predictive standard setting reflects not only the stringency of performance required of a standard, but also the correlation between focal test and outcome test scores. Thus, construct irrelevant factors to academic readiness may also affect the location of the predictive cut score.

Finally, the likely interpretation of predictive cut scores is not supported by the prediction process. Once predictive cut scores are set as a college-readiness standard, it is likely that the resulting classification would be interpreted as an attribute of the student, when in actual fact, the classification reflects student performance as well as test correlations, a function of the joint properties of the focal test and outcome test (see Appendix C.1).

V.5.3 Implications for using predictive standard setting as basis to identify neighborhood for cut scores in evidence-based standard setting

The quality of the cut score recommended by evidence-based standard setting is only as strong as the evidence that it rests on. The discussion about the issues with predictive standard setting also applies when it is used to identify the neighborhood for the eventual cut score such as that used in empirical-based standard setting (see Appendix D for a more detailed discussion). Any study that uses regression-based prediction for standard setting, whether it is based on a future outcome or a concurrent outcome would be subject to similar issues that predictive standard setting faces.

Equipercentile cut score appears to address the problem of confounding stringency with predictive strength of tests. However, it is sensitive to the choice of outcome measure (Appendix C.4), and choices in the formation of the analytic sample (Appendix C.5). The stringency of the criterion score depends on the outcome scale used in the analytic sample, as well as the overall distribution of performance in the analytic sample. Furthermore, depending on the test difficulty, the specification of what students know and are able to do may differ from test to test.

Hence, even if there is clear convergence of empirical cut scores around a region of the focal test score scale, it is not clear whether they are converging around a "correct" cut score that accurately reflects the standards of performance required for a student to be considered college-ready. In the case where there is no clear convergence of empirical cut scores, judgment will be required to prioritize some evidence over others in order to arrive at the final cut score.

VI. "Empirically-Based Statements" to Accompany Judgmental-Based Standard Setting

In this section, I explore an alternative that uses empirical methods to generate "empirically-based statements" which can then be attached post-hoc to cut scores identified by judgmental-based standard setting. This method seeks to combine the strengths of both standard setting methods. I briefly recap the pros and cons of each method.

Predictive standard setting is appealing because it provides an interpretation to the test score scale that references a meaningful external criterion. However, as a stand-alone standard setting method, the cut scores identified by predictive standard setting deviate from a stringency-only standard due to imperfect predictive utility of the score scale. The definition of what students should know and are able to do to perform at a particular performance level is derived post-hoc after the cut score is identified, and thus is circumscribed by the predictive cut score and the test. Moreover, judgment is also required to select the prediction method and the inputs to use.

Judgmental-based standard setting on the other hand, focuses on defining what students should know and are able to do at a given performance level as measured by the focal test, to the best judgment of the standard setting panel. The disadvantage of the method is that is that it can be subjective.

The alternative proposal seeks to combine the strengths of predictive standard setting and judgment-based standard setting. In this presented alternative, cut scores can first be set using judgmental-based standard setting based on the standard setting panel's best judgment of the knowledge and skills required for a given performance level.

"Empirically-based statements" generated by prediction methods can then be attached to these identified cut scores. These statements facilitate checking the reasonableness of the cut score with respect to an external criterion, and also becomes part of the interpretation of the score scale that future users can reference to.

In the next few sections, I present the forms of "empirically-based statements" that can be generated for any given score *x* on the predictor test, before critiquing this alternative.

RQ2b: What are the forms of empirically-based statements that can accompany any score on the focal test regarding a future outcome score?

Data and Methods

To answer RQ2b, I use the same data and methods as for RQ1. The only difference is that here, I calculate the expected outcome score given a particular score on the focal test, or the combination of expected outcome score and probability (logistic regression) or quantile (quantile regression).

Results

Table 7 shows the future outcome score predicted by OLS and quantile regression for a range of cut scores set at *x* on the focal test, and a range of focal-outcome test correlations. The equipercentile scores are also included as a reference point that indicates a corresponding stringency-only relationship. Table 8 shows the logistic regression-derived probabilities of predicting a corresponding future outcome score over ±1 standard deviation units of the average outcome score for a range of cut scores set at *x* on the predictor test. Using these two tables, we can generate empirically-based statements for specific scores on the focal test.

For example, using a theoretical dataset where *X* and *Y* are both standard normal distributions, and the correlation between *X* and *Y* is +0.3, the following set of statements about a student scoring at +1.0 SD on the focal test are all empirically-based:

- A student who scored +1.0 SD on the predictor test scored 84.1% higher than all students on the predictor test. A student who scored +1.0 SD on the outcome test would also score 84.1% higher than all students on the outcome test (equipercentile linking-derived statement).
- A student who scored +1.0 SD on the predictor test is predicted to score +0.30 SD on the outcome test, on average (OLS regression-derived statement).
- The median score on the outcome test for a student who scored +1.0 SD on the predictor test is predicted to be +0.30 SD (quantile regression-derived statement).
- A student who scored +1.0 SD on the predictor test has a 23.0% probability of scoring +1.0 SD or higher on the outcome test (logistic regression-derived statement).
- A student who scored +1.0 SD on the predictor test has a 62.4% probability of scoring 0.0 SD or higher on the outcome test (logistic regression-derived statement, approximately LR65).
- A student who scored +1.0 SD on the predictor test has a 76.8% probability of scoring -0.4 SD or higher on the outcome test (logistic regression-derived statement, approximately LR75).

Note that for any given score on the predictor test, there is an infinite number of (p, Y_c^{LR}) and (q, Y_c^{QR}) pairs, i.e. there is an infinite number of "empirically-based statements" that can be attached to any given X value.

RQ2b: What is the suitability of attaching "empirically-based statements" to judgmental-based cut scores as a standard setting method?

"True to the extent that regression assumptions are not violated". The empirical-based statements generated by prediction methods are true "to the extent that the modeling assumptions are not violated".

Perceptions influenced by predictive utility of scale. Even though prediction is no longer driving the identification of cut scores, the "empirically-based statements" are still influenced by the predictive utility of the scale. These statements and the level of the criterion score referenced to will still be influenced by the focal-outcome test correlations. There is still a possibility that these statements may drive judgments of the reasonableness of judgmental-based identified cut score.

Take the case where the judgmental-based cut score is at +1.0 SD on the focal test scale. The logistic-regression derived statement is: "A student who scored +1.0 SD on the predictor test has a 76.8% probability of scoring -0.4 SD or higher on the outcome test." A high probability of scoring a below average predicted score would hardly be palatable for those seeking high standards. On the other hand, an alternative statement for an identical cut score: "A student who scored +1.0 SD on the predictor test has a 23.0% probability of scoring +1.0 SD or higher on the outcome test" would seem to stand on shaky grounds for predicting with only 23.0% probability for a reasonably high target score.

Both of the above statements are true for the same cut score, and both might invoke a response to increase the correct probability of prediction, or the predicted criterion score higher, both of which would work to shift the judgmental-based cut score upwards. But the perception of low standards or poor prediction probability of the judgmentally identified cut score is not indicative of low standards of the judgmental-based standard setting panel, but influenced by the poor predictive utility of the focal test for the outcome.

Consider an alternative scenario where the focal-outcome test correlation is r=0.9, the following empirical-based statements are true:

- A student who scored +1.0 SD on the predictor test scored 84.1% higher than all students on the predictor test. A student who scored +1.0 SD on the outcome test would also score 84.1% higher than all students on the outcome test (equipercentile linking-derived statement).
- A student who scored +1.0 SD on the predictor test has a 40.1% probability of scoring +1.0 SD or higher on the outcome test (logistic regression-derived statement).
- A student who scored +1.0 SD on the predictor test has a 77.5% probability of scoring +0.6 SD or higher on the outcome test (logistic regression-derived statement).

For a similar cut score, the focal test with higher correlation with the outcome test would be perceived as setting “higher standards” than the one with lower correlation. Hence, even though the use of “empirically-based statements” does not drive the initial identification of cut scores (via judgmental-based standard setting), weak focal-outcome

test correlations still drive the relationships between any score on the focal test with a future outcome score, which may in turn influence perceptions of whether "high" standards are set. This is a testable hypothesis and can be studied.

Camara (2013) writes that:

> "Phillips (2012) states that "it is uncritically accepted that the performance standards must be based on the content standards and the PLDs written by the content experts, and that they should not be contaminated by empirical data" (p.323). Panelists in each state likely believe they have set rigorous standards and cut scores based on their experiences in the classroom, but without any external referent the validation argument is based solely on one line of evidence (i.e., content). … Given the intended purposes of CCR assessments, if performance levels and benchmarks are inconsistent with empirical data on performance in college and career-training programs, they will not only lack credibility but would raise concerns about the validity of the interpretative argument." (p.23)

The converse is also true. If empirical-based cut scores are uncritically accepted to reveal "correct" cut scores, then they may cast doubt on judgmental-based cut scores when in fact, the empirical-based cut scores are the ones deviating away from a stringency-only standard.

Performance standard circumscribed by test. Earlier, I alluded that defining what students should know and be able to do receives greater focus in judgmental-based standard setting than predictive standard setting. In actual practice, standards set in judgmental-based standard setting are still limited to knowledge and skills that are included in the test (Haertel, Beimers, & Miles, 2012). When the standard setting panel

specifies a performance standard or achievement level descriptor that includes knowledge or skills not tested, the panel has to revise the standard and descriptor. I contend that this is a key weakness in current standard setting. The main focus of standard setting has been on setting the cut score, while defining the standard and achievement level descriptors take a secondary focus. Even if both receive the same amount of attention and time, the eventual standard and achievement level descriptors are still subject to what is included in the test. I suggest that to set college-readiness standards that are driven by the level of performance required to be "college-ready", standard setting needs to shift its focus from setting cut scores on a given test, to integrate standard setting with test development and construction, so that tests can be developed to measure different levels of performance well.

VII. Summary and Conclusion

Traditional methods for setting performance standards have been widely critiqued because judgment of the standard setting panel is the main basis for setting standards. With the focus on setting college-readiness standards, there seems to be promise for a new approach to use external criterion evidence to drive empirical standard setting.

However, as Ho (2012) demonstrates, and which I which show the results empirically in this study, predictive methods will confound the stringency of the standard with the predictive utility of the focal test when the correlation between the focal test scores and predicted test scores are less than 1. At levels of focal-outcome test correlations typically observed between K-12 tests and measures of college-readiness, the deviation of predictive cut scores from stringency-only standards can be quite severe. This will lead to cut scores that are too demanding or too lenient than what the level of

performance for college-readiness calls for. Judgment is also required for specifying every input to the prediction model.

Although equipercentile cut scores avoid confounding between the stringency of standards and the predictive utility of tests, it is not immune to problems with the score scales. In particular, the level of performance specified by equipercentile cut scores may be biased when score scales are not well-defined or when there is bias in the analytic sample. Furthermore, the level of performance specified by equipercentile cut scores are circumscribed by the content of the test in practice, rather than defined by knowledge and skills that being college-ready calls for.

Collectively, these issues with the various empirical methods for standard setting also affect evidence-based standard setting. Even if cut scores identified by various empirical methods using different outcome tests are in close agreement with one another and may converge on a particular region of focal test scores, it is also possible that they are converging on a spurious cut score, i.e. being precisely inaccurate.

In this study, I also explore whether empirically-based statements can be attached to judgmental-based cut scores to add meaning to the score scale and anchor interpretation of scores in a relevant outcome. I demonstrate how these empirically-based statements are still driven by the predictive relationships between focal and outcome tests. False impressions about standards or predictive accuracy being set too low may arise not because the standards of the judgmental-based standard setting panel is low, but because of the poor predictive utility of the focal test.

The findings from this study suggests that predictive standard setting does not solve the reliance on judgment that is commonly associated with judgemental-based

standard setting. Given what we learn about cut scores identified by predictive standard setting methods – that they will be overly stringent when the criterion score is above average performance and overly lenient when the criterion score is below average performance, we might use the longitudinal data as a kind of reasonableness check for the upper or lower limit for identified cut scores. That is, if the proposed cut score is even more stringent than the cut score identified by regression-based predictive methods when the targeted criterion score is above average performance, that might serve as a warning that the cut score may not be warranted. The longitudinal data may serve as external validity evidence to gauge the reasonableness of cut scores, but there will still be a need to focus standard setting on defining the knowledge and skills that students need to have in order to be ready for college.

Tables and Figures

Table 1. How predictive cut scores vary with focal-outcome test correlations, by prediction method

Correlation	Criterion Score																		
	-2.0	**-1.6**	**-1.4**	**-1.2**	**-1.0**	**-0.8**	**-0.6**	**-0.4**	**-0.2**	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**1.2**	**1.4**	**1.6**	**2.0**
	A. Equipercentile "stringency-only" reference																		
r=0.9	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
r=0.7	-2.00	-1.60	-1.41	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
r=0.5	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
r=0.3	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
	B. OLS																		
r=0.9	-2.22	-1.78	-1.56	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00	0.22	0.44	0.67	0.89	1.11	1.33	1.56	1.78	2.22
r=0.7	-2.86	-2.29	-2.00	-1.71	-1.43	-1.14	-0.86	-0.57	-0.29	0.00	0.29	0.57	0.86	1.14	1.43	1.71	2.00	2.29	2.86
r=0.5	-4.00	-3.20	-2.80	-2.40	-2.00	-1.60	-1.20	-0.80	-0.40	0.00	0.40	0.80	1.20	1.60	2.00	2.40	2.80	3.20	4.00
r=0.3	-6.67	-5.33	-4.67	-4.00	-3.33	-2.67	-2.00	-1.33	-0.67	0.00	0.67	1.33	2.00	2.67	3.33	4.00	4.67	5.33	6.67
	C. LR50																		
r=0.9	-2.20	-1.76	-1.55	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00	0.22	0.44	0.67	0.89	1.11	1.33	1.55	1.77	2.20
r=0.7	-2.68	-2.19	-1.94	-1.68	-1.41	-1.13	-0.85	-0.57	-0.28	0.00	0.29	0.57	0.85	1.13	1.40	1.67	1.93	2.19	2.68
r=0.5	-3.51	-2.91	-2.61	-2.28	-1.93	-1.57	-1.19	-0.80	-0.40	0.00	0.40	0.79	1.18	1.56	1.92	2.27	2.59	2.91	3.48
r=0.3	-5.48	-4.63	-4.19	-3.70	-3.15	-2.57	-1.96	-1.32	-0.66	0.00	0.67	1.31	1.95	2.55	3.12	3.65	4.12	4.60	5.39
	D. QR50																		
r=0.9	-2.22	-1.78	-1.55	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00	0.22	0.45	0.67	0.89	1.11	1.33	1.56	1.78	2.22
r=0.7	-2.85	-2.28	-2.00	-1.71	-1.43	-1.14	-0.85	-0.57	-0.28	0.00	0.29	0.57	0.86	1.15	1.43	1.72	2.00	2.29	2.86
r=0.5	-4.00	-3.20	-2.80	-2.40	-2.00	-1.60	-1.20	-0.80	-0.40	0.00	0.40	0.80	1.20	1.60	2.00	2.41	2.81	3.21	4.01
r=0.3	-6.66	-5.33	-4.66	-4.00	-3.33	-2.66	-2.00	-1.33	-0.66	0.00	0.67	1.34	2.00	2.67	3.34	4.00	4.67	5.33	6.67

Note: LR50 refers to logistic regression with 50% probability of scoring at or above the criterion score. QR50 refers to quantile regression at the median. Based on standard bivariate normal distributions.

Table 2. How predictive cut scores vary by prediction method, grouped within focal-test outcome correlations

Prediction Method	Criterion Score																		
	-2.0	**-1.6**	**-1.4**	**-1.2**	**-1.0**	**-0.8**	**-0.6**	**-0.4**	**-0.2**	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**1.2**	**1.4**	**1.6**	**2.0**
	Correlation = 0.9																		
Equi	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
OLS	-2.22	-1.78	-1.56	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00	0.22	0.44	0.67	0.89	1.11	1.33	1.56	1.78	2.22
LR50	-2.20	-1.76	-1.55	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00	0.22	0.44	0.67	0.89	1.11	1.33	1.55	1.77	2.20
QR50	-2.22	-1.78	-1.55	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00	0.22	0.45	0.67	0.89	1.11	1.33	1.56	1.78	2.22
	Correlation = 0.7																		
Equi	-2.00	-1.60	-1.41	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
OLS	-2.86	-2.29	-2.00	-1.71	-1.43	-1.14	-0.86	-0.57	-0.29	0.00	0.29	0.57	0.86	1.14	1.43	1.71	2.00	2.29	2.86
LR50	-2.68	-2.19	-1.94	-1.68	-1.41	-1.13	-0.85	-0.57	-0.28	0.00	0.29	0.57	0.85	1.13	1.40	1.67	1.93	2.19	2.68
QR50	-2.85	-2.28	-2.00	-1.71	-1.43	-1.14	-0.85	-0.57	-0.28	0.00	0.29	0.57	0.86	1.15	1.43	1.72	2.00	2.29	2.86
	Correlation = 0.5																		
Equi	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
OLS	-4.00	-3.20	-2.80	-2.40	-2.00	-1.60	-1.20	-0.80	-0.40	0.00	0.40	0.80	1.20	1.60	2.00	2.40	2.80	3.20	4.00
LR50	-3.51	-2.91	-2.61	-2.28	-1.93	-1.57	-1.19	-0.80	-0.40	0.00	0.40	0.79	1.18	1.56	1.92	2.27	2.59	2.91	3.48
QR50	-4.00	-3.20	-2.80	-2.40	-2.00	-1.60	-1.20	-0.80	-0.40	0.00	0.40	0.80	1.20	1.60	2.00	2.41	2.81	3.21	4.01
	Correlation = 0.3																		
Equi	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
OLS	-6.67	-5.33	-4.67	-4.00	-3.33	-2.67	-2.00	-1.33	-0.67	0.00	0.67	1.33	2.00	2.67	3.33	4.00	4.67	5.33	6.67
LR50	-5.48	-4.63	-4.19	-3.70	-3.15	-2.57	-1.96	-1.32	-0.66	0.00	0.67	1.31	1.95	2.55	3.12	3.65	4.12	4.60	5.39
QR50	-6.66	-5.33	-4.66	-4.00	-3.33	-2.66	-2.00	-1.33	-0.66	0.00	0.67	1.34	2.00	2.67	3.34	4.00	4.67	5.33	6.67

Note: Equi refers to equipercentile cut score. LR50 refers to logistic regression with 50% probability of scoring at or above the criterion score. QR50 refers to quantile regression at the median. Based on standard bivariate normal distributions.

Table 3. How predictive cut scores vary with probability *p* of scoring at or above the criterion score using logistic regression, by focal-outcome test correlations

Probability *p*	**Criterion Score**																		
	-2.0	**-1.6**	**-1.4**	**-1.2**	**-1.0**	**-0.8**	**-0.6**	**-0.4**	**-0.2**	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**1.2**	**1.4**	**1.6**	**2.0**
	Correlation = 0.9																		
p=50	-2.20	-1.76	-1.55	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00	0.22	0.44	0.67	0.89	1.11	1.33	1.55	1.77	2.20
p=65	-2.04	-1.60	-1.39	-1.17	-0.94	-0.72	-0.50	-0.27	-0.05	0.17	0.39	0.61	0.83	1.05	1.27	1.49	1.71	1.93	2.35
p=75	-1.92	-1.48	-1.26	-1.04	-0.82	-0.59	-0.37	-0.14	0.08	0.30	0.52	0.74	0.96	1.18	1.40	1.62	1.84	2.05	2.48
	Correlation = 0.7																		
p=50	-2.68	-2.19	-1.94	-1.68	-1.41	-1.13	-0.85	-0.57	-0.28	0.00	0.29	0.57	0.85	1.13	1.40	1.67	1.93	2.19	2.68
p=65	-2.38	-1.86	-1.60	-1.33	-1.06	-0.77	-0.49	-0.20	0.09	0.37	0.66	0.94	1.21	1.49	1.75	2.02	2.26	2.51	2.98
p=75	-2.15	-1.62	-1.35	-1.07	-0.78	-0.50	-0.20	0.09	0.37	0.66	0.94	1.22	1.50	1.76	2.02	2.28	2.52	2.76	3.21
	Correlation = 0.5																		
p=50	-3.51	-2.91	-2.61	-2.28	-1.93	-1.57	-1.19	-0.80	-0.40	0.00	0.40	0.79	1.18	1.56	1.92	2.27	2.59	2.91	3.48
p=65	-3.03	-2.39	-2.06	-1.70	-1.33	-0.95	-0.55	-0.15	0.25	0.65	1.05	1.43	1.81	2.17	2.51	2.84	3.14	3.44	3.96
p=75	-2.65	-1.98	-1.63	-1.25	-0.87	-0.47	-0.06	0.35	0.75	1.16	1.55	1.93	2.29	2.65	2.97	3.29	3.56	3.85	4.33
	Correlation = 0.3																		
p=50	-5.48	-4.63	-4.19	-3.70	-3.15	-2.57	-1.96	-1.32	-0.66	0.00	0.67	1.31	1.95	2.55	3.12	3.65	4.12	4.60	5.39
p=65	-4.63	-3.68	-3.19	-2.64	-2.05	-1.43	-0.79	-0.12	0.55	1.22	1.88	2.50	3.11	3.68	4.21	4.69	5.11	5.54	6.22
p=75	-3.98	-2.95	-2.41	-1.83	-1.20	-0.55	0.12	0.81	1.49	2.17	2.82	3.43	4.01	4.56	5.05	5.50	5.87	6.27	6.87

Note: Based on standard bivariate normal distributions. The predictive cut scores are derived from logistic regression with 50%, 65%, and 75% probability of scoring at or above the criterion score.

Table 4. How predictive cut scores vary with quantile q of quantile regression, by focal-outcome test correlations

Quantile *q*	**Criterion Score**																		
	-2.0	**-1.6**	**-1.4**	**-1.2**	**-1.0**	**-0.8**	**-0.6**	**-0.4**	**-0.2**	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**1.2**	**1.4**	**1.6**	**2.0**
	Correlation = 0.9																		
q=50	-2.22	-1.78	-1.55	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00	0.22	0.45	0.67	0.89	1.11	1.33	1.56	1.78	2.22
q=35	-2.04	-1.59	-1.37	-1.15	-0.92	-0.70	-0.48	-0.26	-0.04	0.19	0.41	0.63	0.85	1.08	1.30	1.52	1.74	1.96	2.41
q=25	-1.90	-1.45	-1.23	-1.01	-0.79	-0.56	-0.34	-0.12	0.10	0.33	0.55	0.77	0.99	1.22	1.44	1.66	1.88	2.10	2.55
	Correlation = 0.7																		
q=50	-2.85	-2.28	-2.00	-1.71	-1.43	-1.14	-0.85	-0.57	-0.28	0.00	0.29	0.57	0.86	1.15	1.43	1.72	2.00	2.29	2.86
q=35	-2.47	-1.90	-1.61	-1.32	-1.04	-0.75	-0.47	-0.18	0.11	0.39	0.68	0.96	1.25	1.54	1.82	2.11	2.39	2.68	3.25
q=25	-2.17	-1.60	-1.31	-1.03	-0.74	-0.45	-0.17	0.12	0.40	0.69	0.97	1.26	1.54	1.83	2.11	2.40	2.68	2.97	3.54
	Correlation = 0.5																		
q=50	-4.00	-3.20	-2.80	-2.40	-2.00	-1.60	-1.20	-0.80	-0.40	0.00	0.40	0.80	1.20	1.60	2.00	2.41	2.81	3.21	4.01
q=35	-3.33	-2.53	-2.13	-1.73	-1.33	-0.93	-0.53	-0.13	0.27	0.67	1.07	1.46	1.86	2.26	2.66	3.06	3.46	3.86	4.66
q=25	-2.84	-2.04	-1.64	-1.24	-0.84	-0.44	-0.04	0.36	0.76	1.17	1.57	1.97	2.37	2.77	3.17	3.57	3.97	4.37	5.17
	Correlation = 0.3																		
q=50	-6.66	-5.33	-4.66	-4.00	-3.33	-2.66	-2.00	-1.33	-0.66	0.00	0.67	1.34	2.00	2.67	3.34	4.00	4.67	5.33	6.67
q=35	-5.44	-4.11	-3.44	-2.78	-2.11	-1.44	-0.78	-0.11	0.56	1.23	1.89	2.56	3.23	3.89	4.56	5.23	5.89	6.56	7.89
q=25	-4.51	-3.18	-2.51	-1.85	-1.19	-0.52	0.14	0.81	1.47	2.13	2.80	3.46	4.13	4.79	5.45	6.12	6.78	7.44	8.77

Note: Based on standard bivariate normal distributions. The predictive cut scores are derived from quantile regression at the median, 35th, and 25th quantile of the outcome score that corresponds to the criterion score.

Table 5. Empirical cut scores identified using Regents high school Math exams (2010 data) as the focal test and first-year GPA as the outcome

Predictive Method	**Target First-Year GPA**					
	C-	**C**	**C+**	**B-**	**B**	**B+**
	1.7	**2.0**	**2.3**	**2.7**	**3.0**	**3.3**
Equi	-0.50	-0.25	0.01	0.44	0.78	1.20
OLS	-0.86	-0.19	0.48	1.37	2.04	2.71
LR50	-1.12	-0.53	0.09	0.89	1.49	2.07
QR50	-1.10	-0.50	0.11	0.91	1.51	2.11
LR65	-0.27	0.27	0.84	1.59	2.14	2.63
LR75	0.39	0.89	1.42	2.13	2.64	3.07
QR35	-0.22	0.31	0.84	1.55	2.08	2.62
QR25	0.41	0.90	1.40	2.05	2.55	3.04

Note: Equi refers to the equipercentile approach. LR50, LR65, and LR75 refers to logistic regression with 50%, 65%, and 75% probability of scoring at or above the criterion score. QR50, QR35, and QR25 refers to quantile regression at the median, 35th, and 25th quantile of the outcome test score (corresponding to the criterion score).

Table 6. Empirical cut scores identified using Regents high school ELA exams (2010 data) as the focal test and first-year GPA as the outcome

Predictive Method	**Target First-Year GPA**					
	C-	**C**	**C+**	**B-**	**B**	**B+**
	1.7	**2.0**	**2.3**	**2.7**	**3.0**	**3.3**
Equi	-0.41	-0.02	0.08	0.37	0.85	1.14
OLS	-0.88	-0.17	0.53	1.47	2.18	2.88
LR50	-1.12	-0.51	0.13	0.93	1.55	2.15
QR50	-1.08	-0.47	0.15	0.97	1.58	2.20
LR65	-0.23	0.33	0.91	1.65	2.22	2.74
LR75	0.45	0.98	1.51	2.21	2.74	3.19
QR35	-0.19	0.35	0.89	1.62	2.17	2.71
QR25	0.46	0.95	1.44	2.10	2.59	3.09

Note: Equi refers to the equipercentile approach. LR50, LR65, and LR75 refers to logistic regression with 50%, 65%, and 75% probability of scoring at or above the criterion score. QR50, QR35, and QR25 refers to quantile regression at the median, 35th, and 25th quantile of the outcome test score (corresponding to the criterion score).

Table 7. Future outcome scores predicted by cut scores, by predictive method and focal-outcome test correlation

Predictive Method	Cut Score																		
	-2.0	**-1.6**	**-1.4**	**-1.2**	**-1.0**	**-0.8**	**-0.6**	**-0.4**	**-0.2**	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**1.2**	**1.4**	**1.6**	**2.0**
	Correlation = 0.9																		
Equi	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
OLS	-1.80	-1.44	-1.26	-1.08	-0.90	-0.72	-0.54	-0.36	-0.18	0.00	0.18	0.36	0.54	0.72	0.90	1.08	1.26	1.44	1.80
QR50	-1.80	-1.44	-1.26	-1.08	-0.90	-0.72	-0.54	-0.36	-0.18	0.00	0.18	0.36	0.54	0.72	0.90	1.08	1.26	1.44	1.80
QR65	-1.63	-1.27	-1.09	-0.91	-0.73	-0.55	-0.37	-0.19	-0.01	0.17	0.35	0.53	0.71	0.89	1.07	1.25	1.43	1.61	1.97
QR75	-1.51	-1.15	-0.97	-0.79	-0.61	-0.43	-0.25	-0.07	0.11	0.29	0.47	0.65	0.83	1.01	1.19	1.37	1.55	1.73	2.09
	Correlation = 0.7																		
Equi	-2.00	-1.60	-1.39	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
OLS	-1.40	-1.12	-0.98	-0.84	-0.70	-0.56	-0.42	-0.28	-0.14	0.00	0.14	0.28	0.42	0.56	0.70	0.84	0.98	1.12	1.40
QR50	-1.40	-1.12	-0.98	-0.84	-0.70	-0.56	-0.42	-0.28	-0.14	0.00	0.14	0.28	0.42	0.56	0.70	0.84	0.98	1.12	1.40
QR65	-1.13	-0.85	-0.71	-0.57	-0.43	-0.29	-0.15	-0.01	0.13	0.27	0.41	0.55	0.69	0.83	0.97	1.11	1.25	1.39	1.67
QR75	-0.92	-0.64	-0.50	-0.36	-0.22	-0.08	0.06	0.20	0.34	0.48	0.62	0.76	0.90	1.04	1.18	1.32	1.46	1.60	1.88
	Correlation = 0.5																		
Equi	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
OLS	-1.00	-0.80	-0.70	-0.60	-0.50	-0.40	-0.30	-0.20	-0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	1.00
QR50	-1.00	-0.80	-0.70	-0.60	-0.50	-0.40	-0.30	-0.20	-0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	1.00
QR65	-0.67	-0.47	-0.37	-0.27	-0.17	-0.07	0.03	0.13	0.23	0.33	0.43	0.53	0.63	0.73	0.83	0.93	1.03	1.13	1.33
QR75	-0.41	-0.21	-0.11	-0.01	0.08	0.18	0.28	0.38	0.48	0.58	0.68	0.78	0.88	0.98	1.08	1.18	1.28	1.38	1.58
	Correlation = 0.3																		
Equi	-2.00	-1.60	-1.40	-1.20	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	2.00
OLS	-0.60	-0.48	-0.42	-0.36	-0.30	-0.24	-0.18	-0.12	-0.06	0.00	0.06	0.12	0.18	0.24	0.30	0.36	0.42	0.48	0.60
QR50	-0.60	-0.48	-0.42	-0.36	-0.30	-0.24	-0.18	-0.12	-0.06	0.00	0.06	0.12	0.18	0.24	0.30	0.36	0.42	0.48	0.60
QR65	-0.23	-0.11	-0.05	0.01	0.07	0.13	0.19	0.25	0.31	0.37	0.43	0.49	0.55	0.61	0.67	0.73	0.79	0.85	0.97
QR75	0.04	0.16	0.22	0.28	0.34	0.40	0.46	0.52	0.58	0.64	0.70	0.76	0.82	0.88	0.94	1.00	1.06	1.12	1.24

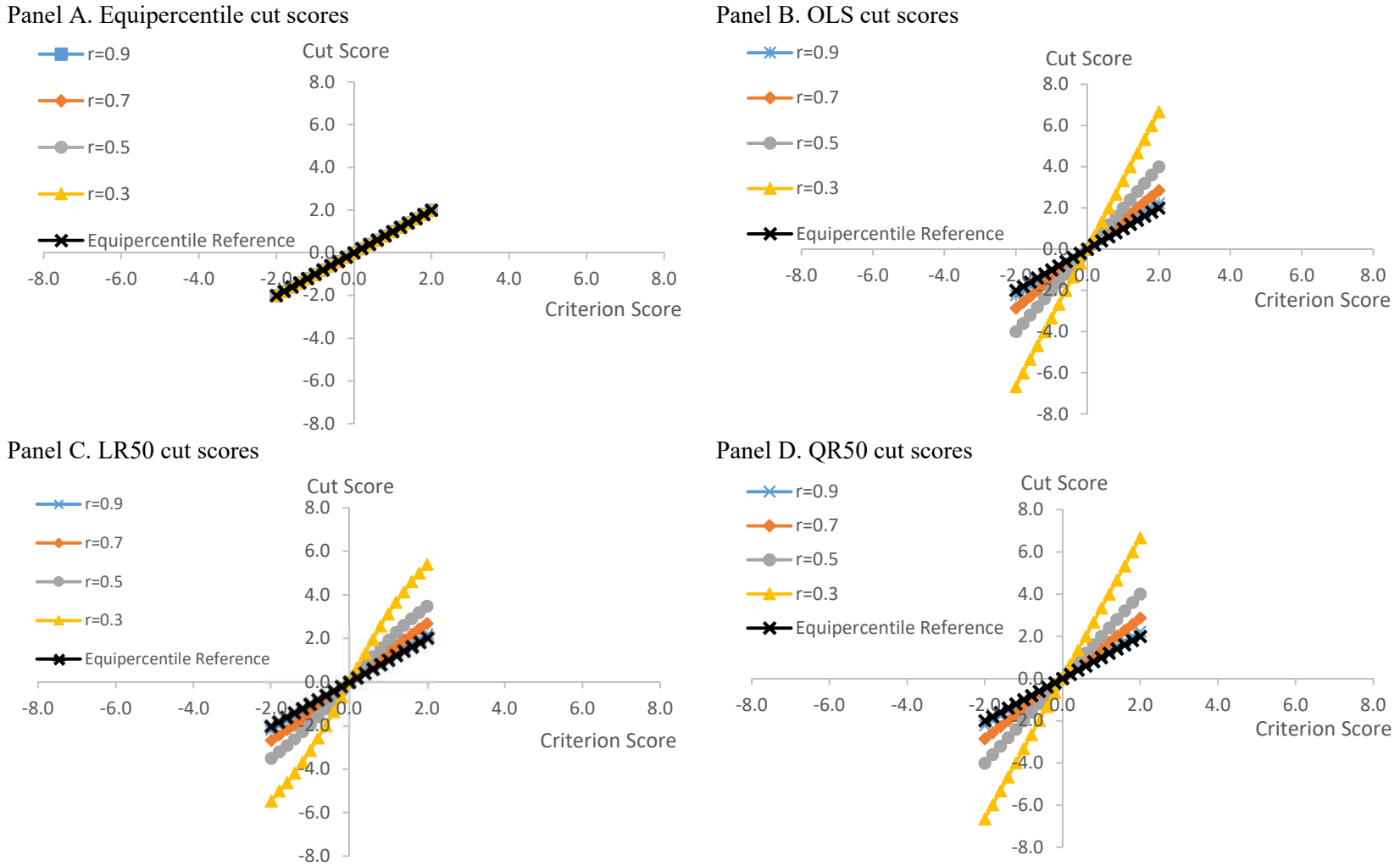
Note: Equi refers to the equipercentile approach. QR50, QR35, and QR25 refers to quantile regression at the median, 35th, and 25th quantile of the outcome test score (corresponding to the criterion score). Based on standard bivariate normal distributions.

Table 8. Probability of scoring above a criterion score on the outcome test, as predicted by cut scores on the focal test, using the logistic regression predictive method, by test correlation

Correlation	Cut Score																		
	-2.0	**-1.6**	**-1.4**	**-1.2**	**-1.0**	**-0.8**	**-0.6**	**-0.4**	**-0.2**	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**1.2**	**1.4**	**1.6**	**2.0**
	Criterion Score = -1.0																		
0.9	0.03	0.14	0.25	0.42	0.60	0.76	0.87	0.93	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.7	0.26	0.42	0.50	0.59	0.67	0.74	0.81	0.85	0.89	0.92	0.94	0.96	0.97	0.98	0.99	0.99	0.99	1.00	1.00
0.5	0.48	0.58	0.63	0.68	0.72	0.76	0.80	0.83	0.86	0.88	0.90	0.92	0.93	0.94	0.95	0.96	0.97	0.97	0.98
0.3	0.66	0.71	0.73	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95
	Criterion Score = -0.4																		
0.9	0.00	0.01	0.03	0.06	0.12	0.21	0.36	0.54	0.71	0.83	0.91	0.96	0.98	0.99	0.99	1.00	1.00	1.00	1.00
0.7	0.08	0.15	0.20	0.26	0.33	0.40	0.49	0.57	0.65	0.72	0.78	0.84	0.88	0.91	0.93	0.95	0.96	0.97	0.99
0.5	0.24	0.32	0.36	0.40	0.45	0.50	0.55	0.59	0.64	0.68	0.72	0.76	0.79	0.82	0.85	0.87	0.89	0.91	0.94
0.3	0.41	0.46	0.49	0.52	0.54	0.57	0.59	0.62	0.64	0.66	0.69	0.71	0.73	0.75	0.77	0.79	0.80	0.82	0.85
	Criterion Score = 0.0																		
0.9	0.00	0.00	0.01	0.01	0.03	0.05	0.10	0.19	0.32	0.50	0.67	0.81	0.90	0.95	0.97	0.99	0.99	1.00	1.00
0.7	0.03	0.07	0.09	0.12	0.16	0.21	0.27	0.34	0.42	0.50	0.58	0.66	0.73	0.79	0.84	0.88	0.91	0.93	0.97
0.5	0.13	0.18	0.21	0.24	0.28	0.32	0.36	0.41	0.45	0.50	0.55	0.59	0.64	0.68	0.72	0.76	0.79	0.82	0.87
0.3	0.27	0.31	0.33	0.35	0.38	0.40	0.42	0.45	0.47	0.50	0.52	0.55	0.58	0.60	0.62	0.65	0.67	0.69	0.73
	Criterion Score = 0.6																		
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.08	0.15	0.27	0.44	0.62	0.77	0.88	0.94	0.97	0.99
0.7	0.01	0.02	0.02	0.03	0.04	0.06	0.08	0.11	0.14	0.19	0.25	0.32	0.39	0.48	0.56	0.64	0.72	0.78	0.88
0.5	0.04	0.06	0.07	0.09	0.10	0.12	0.15	0.17	0.20	0.24	0.28	0.32	0.36	0.41	0.46	0.50	0.55	0.60	0.69
0.3	0.11	0.13	0.14	0.16	0.17	0.19	0.21	0.22	0.24	0.26	0.28	0.31	0.33	0.35	0.38	0.40	0.43	0.45	0.51
	Criterion Score = 1.0																		
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.07	0.13	0.24	0.40	0.59	0.75	0.86	0.97
0.7	0.00	0.00	0.01	0.01	0.01	0.02	0.03	0.04	0.06	0.08	0.11	0.15	0.20	0.26	0.33	0.41	0.50	0.59	0.74
0.5	0.02	0.03	0.03	0.04	0.05	0.06	0.07	0.08	0.10	0.12	0.14	0.17	0.20	0.24	0.28	0.32	0.37	0.42	0.52
0.3	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.16	0.18	0.19	0.21	0.23	0.25	0.27	0.30	0.35

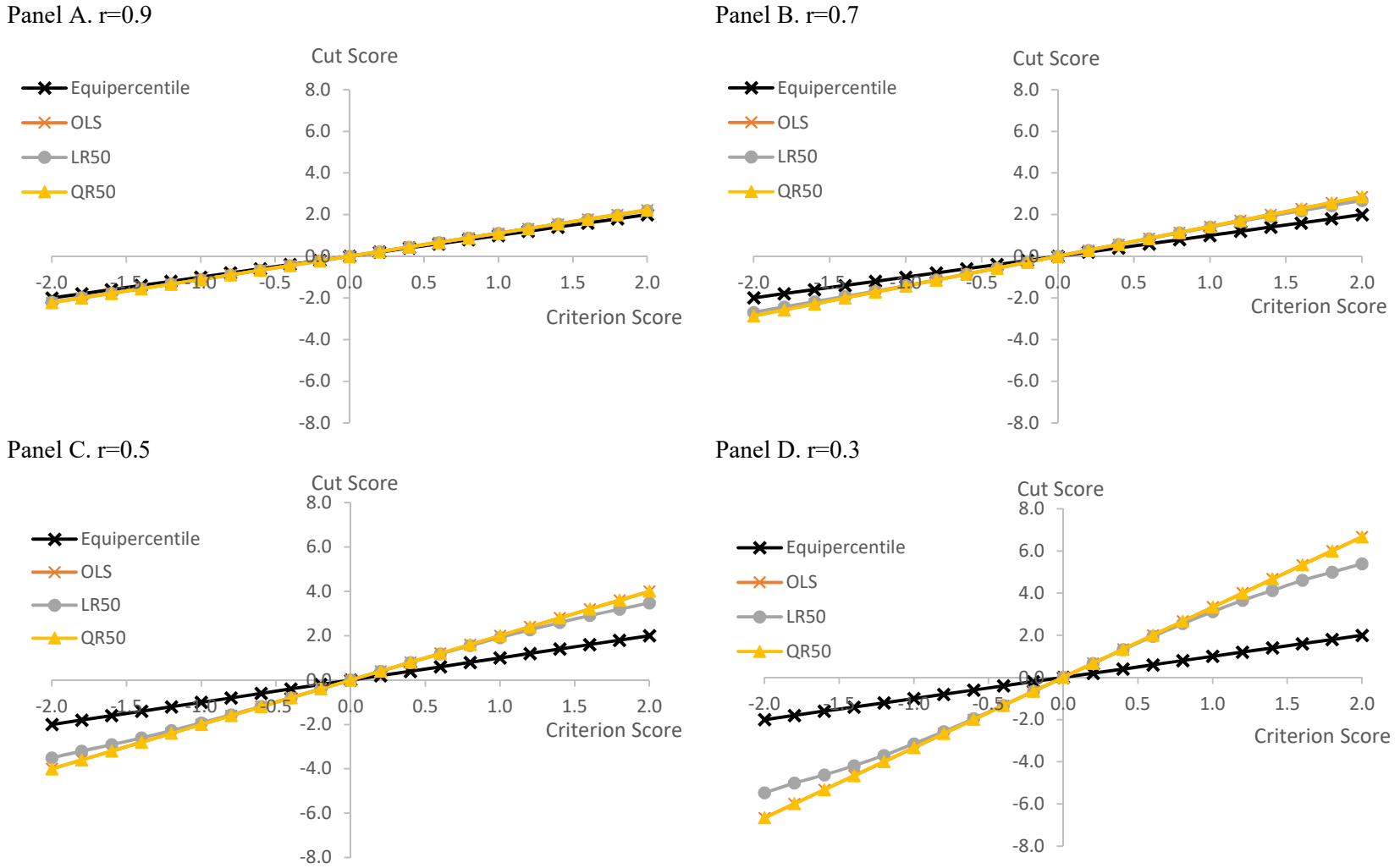
Note: Based on standard bivariate normal distributions.

Figure 1. How predictive cut scores vary with focal-outcome test correlations, by prediction method



Note: LR50 refers to logistic regression with 50% probability of scoring at or above the criterion score. QR50 refers to quantile regression at the median. Based on standard bivariate normal distributions.

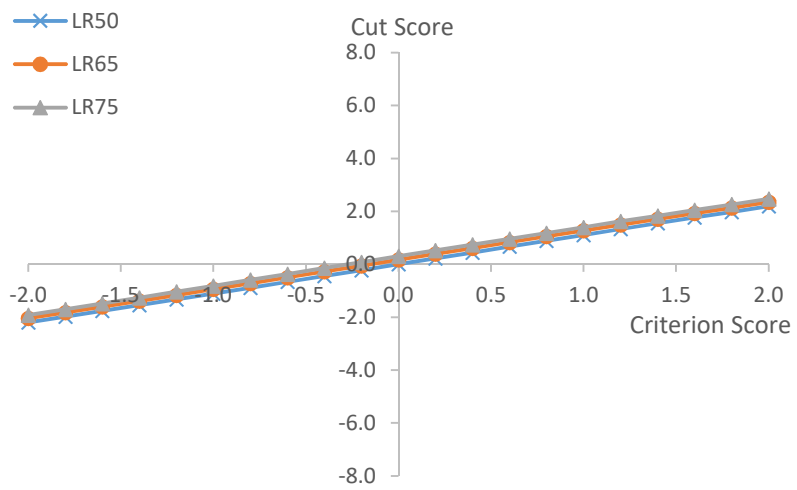
Figure 2. How predictive cut scores vary by prediction method, grouped within focal-test outcome correlations



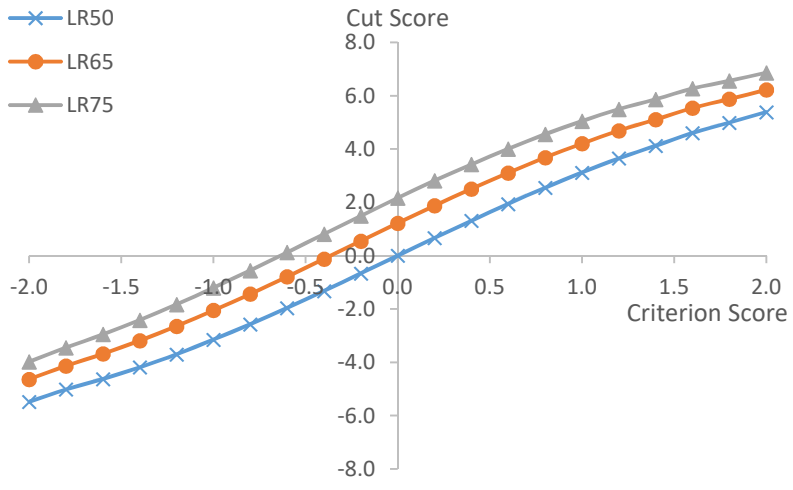
Note: LR50 refers to logistic regression with 50% probability of scoring at or above the criterion score. QR50 refers to quantile regression at the median. Based on standard bivariate normal distributions.

Figure 3. How predictive cut scores vary with probability of scoring at or above the criterion score using logistic regression, by focal-outcome test correlations

Panel A. r=0.9



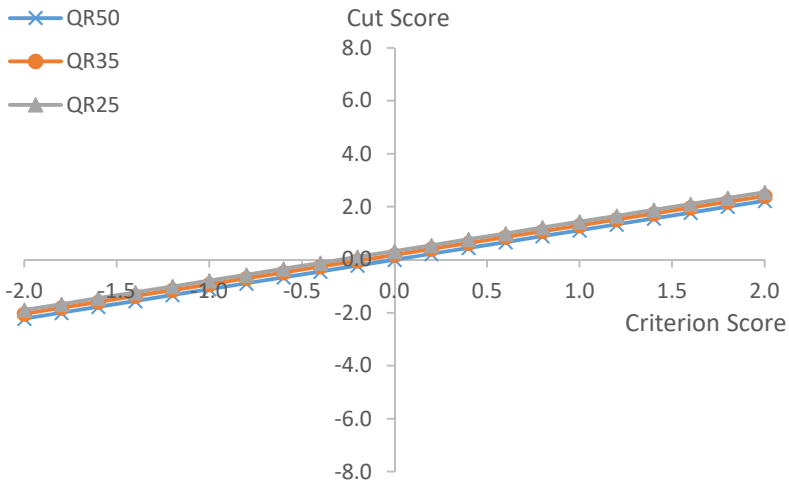
Panel B. r=0.3



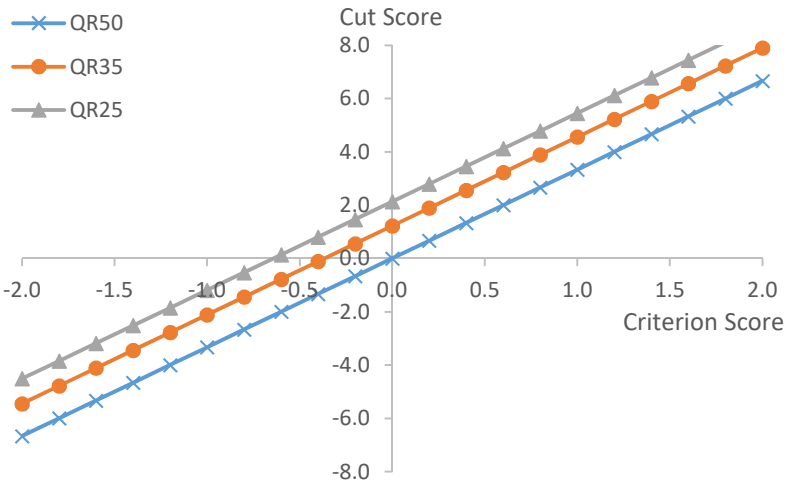
Note: Based on standard bivariate normal distributions. LR50, LR65, LR75 refers to logistic regression with 50%, 65%, and 75% probability of scoring at or above the criterion score.

Figure 4. How predictive cut scores vary with quantile q of quantile regression, by focal-outcome test correlations

Panel A. r=0.9

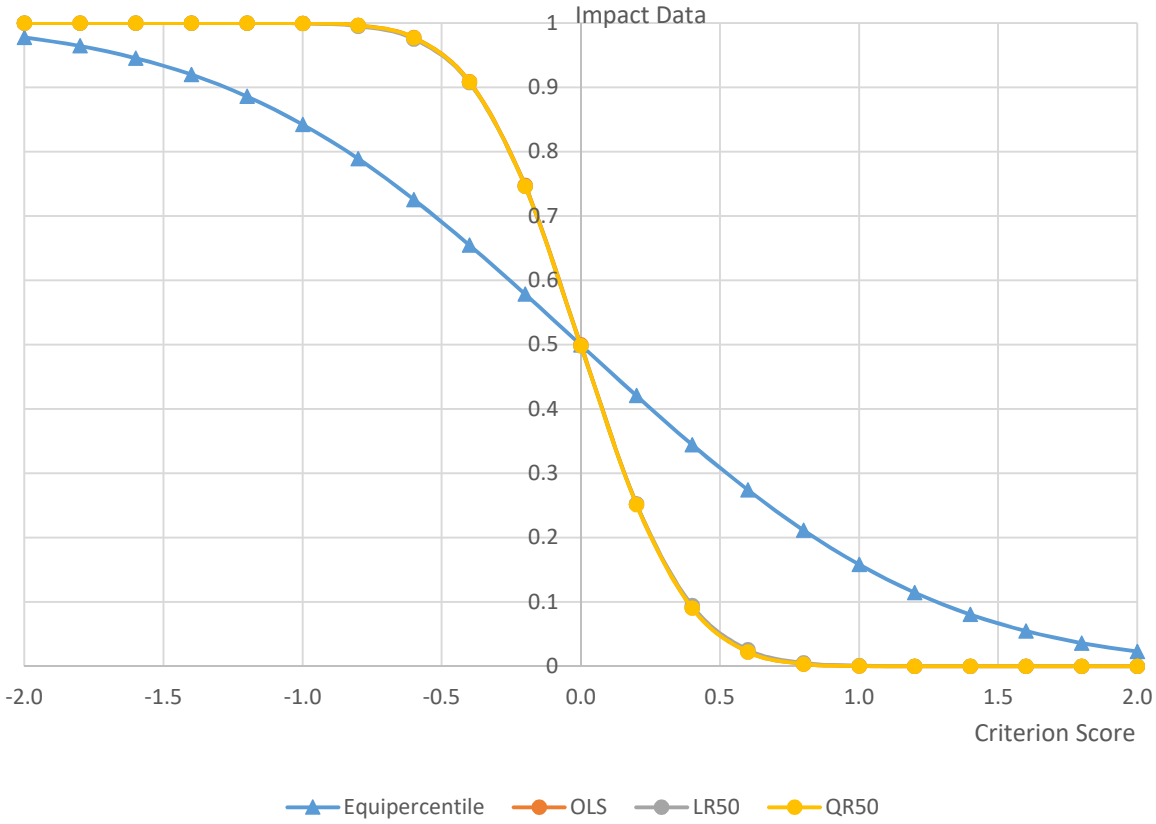


Panel B. r=0.3



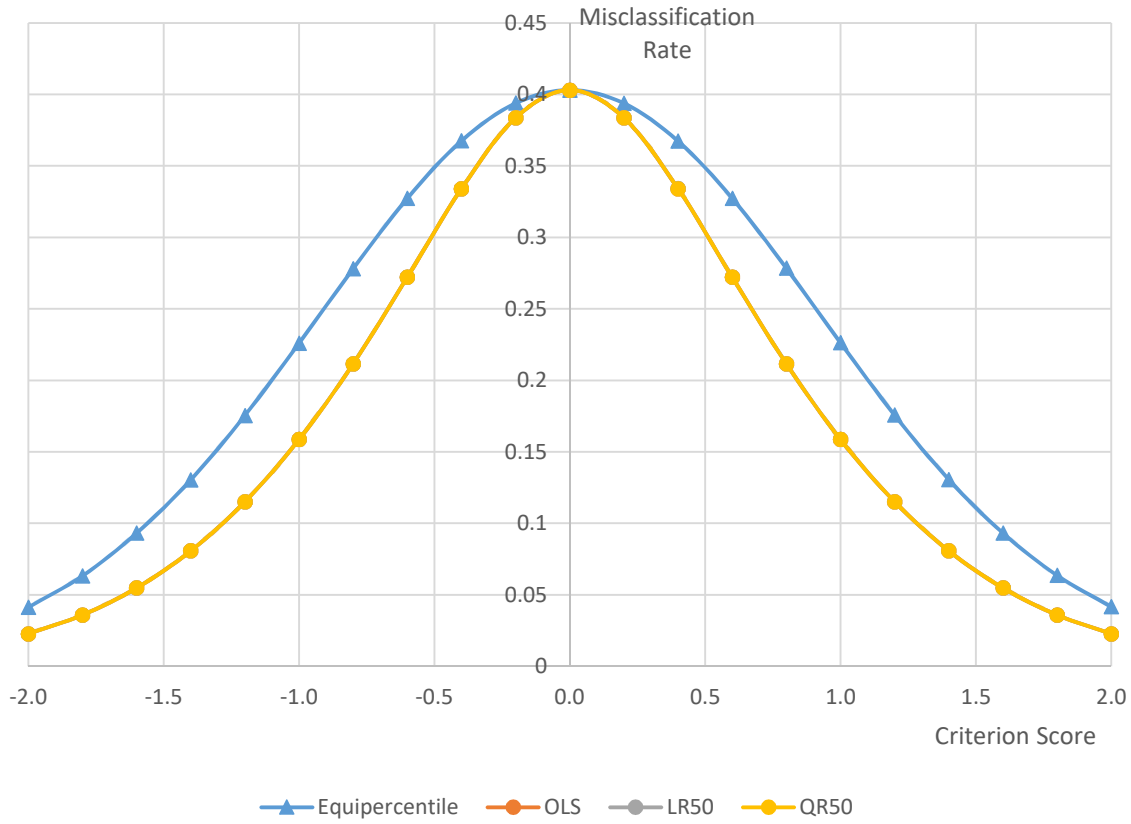
Note: Based on standard bivariate normal distributions. QR50, QR35, and QR25 refers to quantile regression at the median, 35th quantile, and 25th quantile of the outcome test score.

Figure 5. How impact data varies over criterion score, by prediction method for focal-outcome test correlation of r=0.3



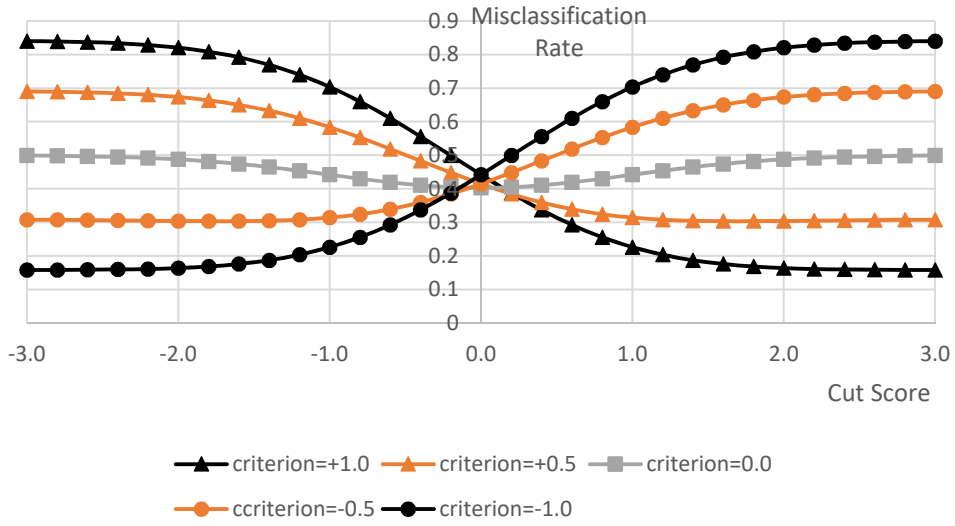
Note: LR50 refers to logistic regression with 50% probability of scoring at or above the criterion score. QR50 refers to quantile regression at the median. Based on standard bivariate normal distributions.

Figure 6. How misclassification rates vary over criterion score, by prediction method for focal-outcome test correlation of r=0.3



Note: LR50 refers to logistic regression with 50% probability of scoring at or above the criterion score. QR50 refers to quantile regression at the median. Based on standard bivariate normal distributions.

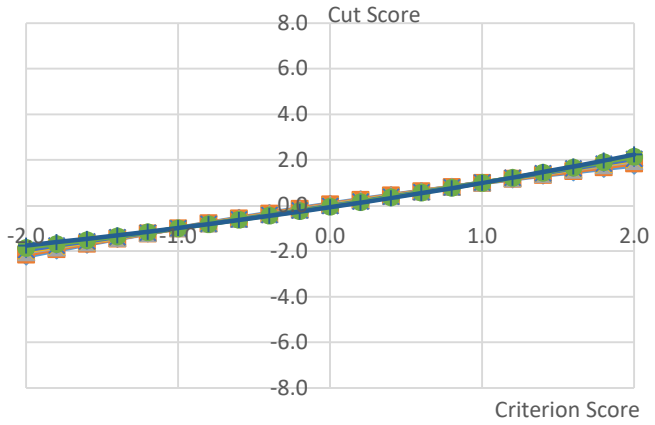
Figure 7. How misclassification rates vary over cut score, by various criterion score, for focal-outcome test correlation of r=0.3



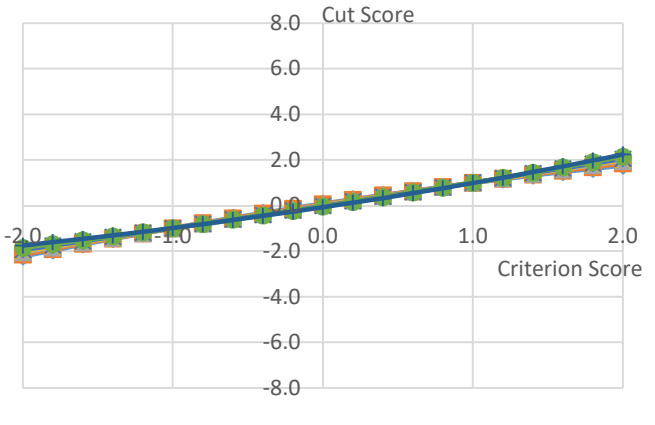
Note: LR50 refers to logistic regression with 50% probability of scoring at or above the criterion score. QR50 refers to quantile regression at the median. Based on standard bivariate normal distributions.

Figure 8. How skewness (by values of γ= ±0.5, ±0.3, and 0.0) affects predictive cut scores, by predictive methods, for focal-outcome test correlations of r=0.3 and r=0.5

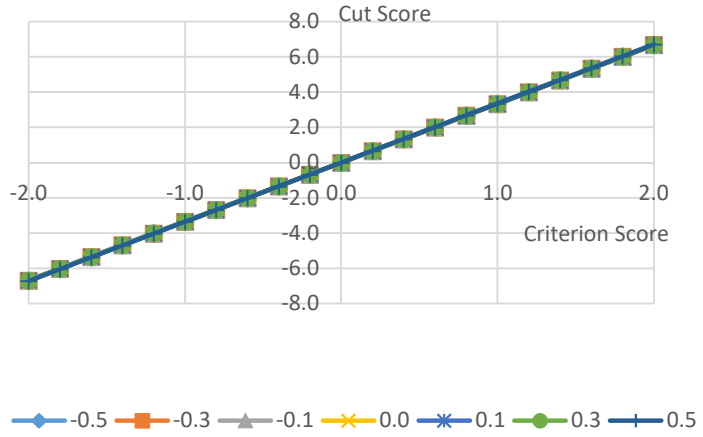
Panel A: Equipercentile cut scores; r=0.3



Panel B: Equipercentile cut scores; r=0.5



Panel C: OLS; r=0.3



Panel D: OLS; r=0.5

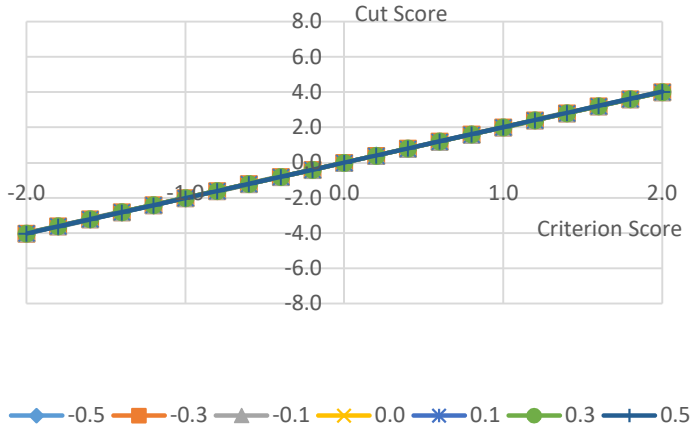
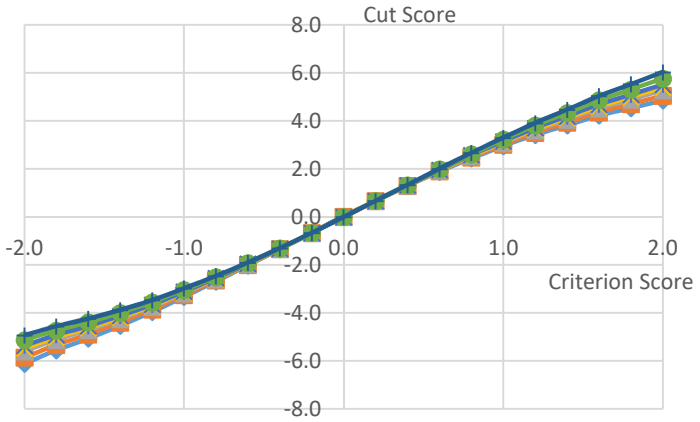
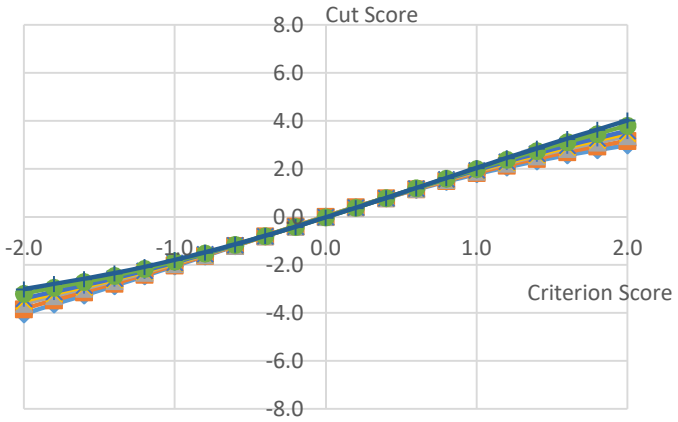


Figure 8 (continued)

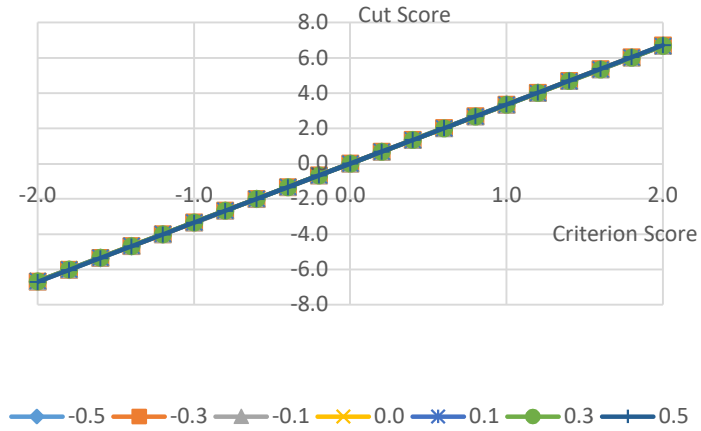
Panel E: LR50; r=0.3



Panel F:LR50; r=0.5



Panel G: QR50; r=0.3



Panel H: QR50; r=0.5

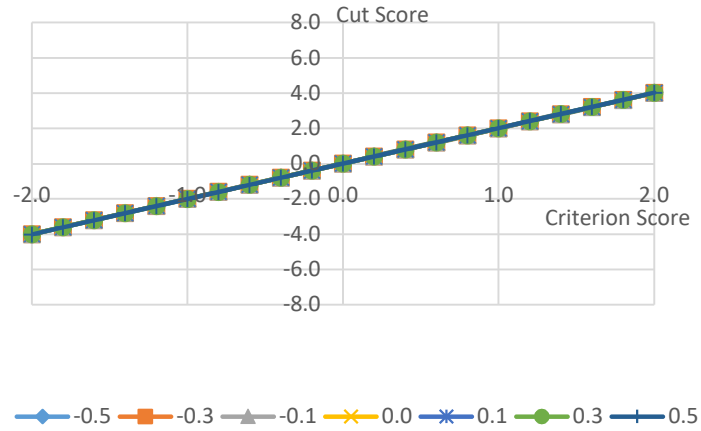
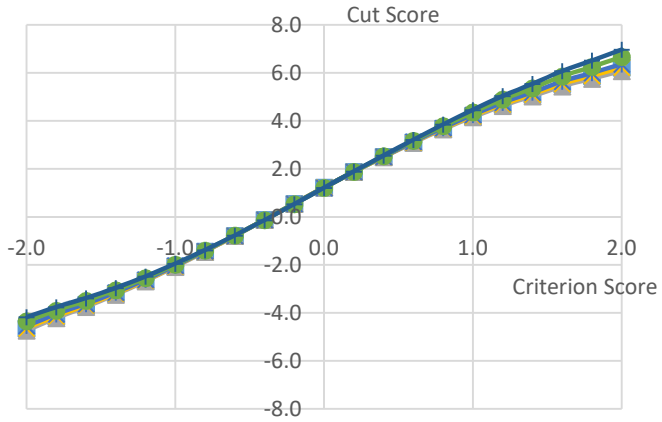
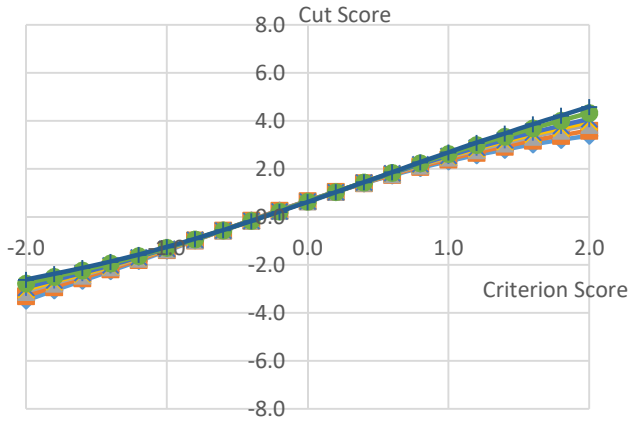


Figure 8 (continued)

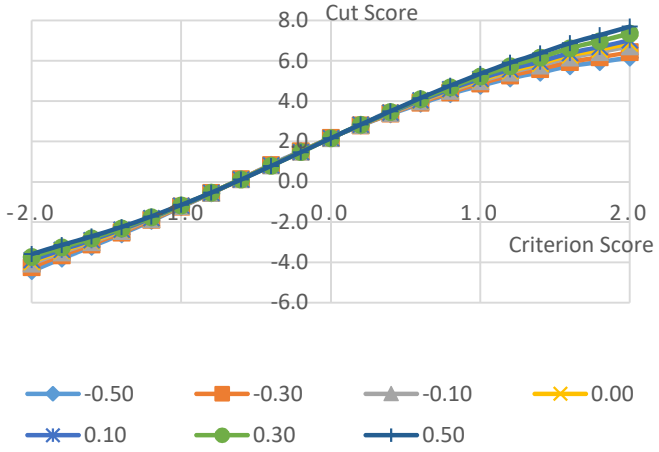
Panel I: LR65; r=0.3



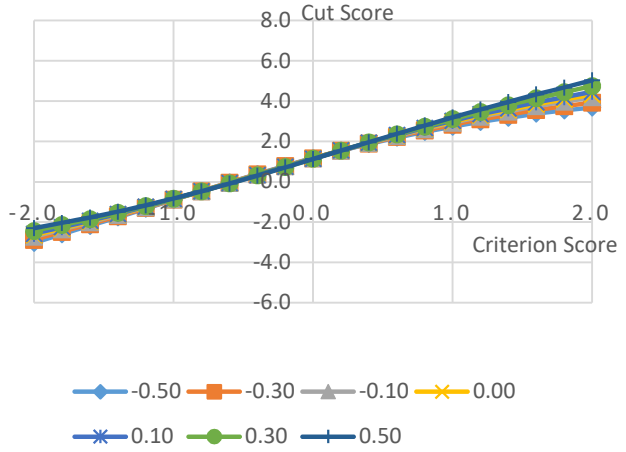
Panel J: LR65; r=0.5



Panel K: LR75; r=0.3



Panel L: LR75; r=0.5



Appendix A

Diagram Depicting Use of Outcome Test Scores to Identify Predictive Cut Scores

Appendix B
Misclassification Errors

Table B1. 2x2 matrix of observed versus actual score relative to identified cut score and actual performance relative to standard

		True scores	
		Below standard	**Above standard**
Observed scores	**Below cut score**	P00	P01
	Above cut score	P10	P11

I base the notation on the discussion in Clauser, Margolis, and Case (2006). Let x and τ be the scores on the observed and true score scales and let x_0 and τ_0 be the respective cut scores. Then,

P00 = Probability ($x < x_0$ and $\tau < \tau_0$)

P01 = Probability ($x < x_0$ and $\tau \geq \tau_0$)

P10 = Probability ($x \geq x_0$ and $\tau < \tau_0$)

P11 = Probability ($x \geq x_0$ and $\tau \geq \tau_0$)

I refer to false positive rate (fp) as the proportion of students who are not proficient but receive passing status on the test (P10). I refer to false negative rate (fn) as the proportion of students who are proficient but received failing status on the test (P01). The misclassification rate is calculated as the sum of the false positive rate and the false negative rate (P10 + P01).

Appendix C
Further Issues with Use of Predictive Standard Setting as a Stand-Alone Standard Setting Method

In this appendix, I discuss a couple of issues that may arise from the dependency of predictive cut scores on the focal-outcome test correlation.

C.1 Implications for the validity of likely interpretations

Construct irrelevant variance. Assuming the regression assumptions are met, the correct interpretations of predictive cut scores are as follows: On average, students who score at cut score X_c^{OLS} are predicted to have a mean outcome score Y_c (OLS cut score); A student who scores at cut score X_c^{LR} is $p\%$ likely to achieve future score Y_c (logistic regression cut score); The q^{th} percentile of outcome score for students who score at cut score X_c^{QR} is expected to be Y_c.

Cut scores are widespread in education and their use are common and predictable (Haertel & Ho, 2012). Once predictive cut scores are set on K-12 tests, performance labels, such as "college-ready"/"not college-ready", or "advanced academic performance"/"satisfactory academic performance"/"unsatisfactory academic performance" (e.g. Texas Education Agency, 2013, p.42), will likely be attached to these cut scores. Even though the underlying statement is probabilistic, cut scores are commonly used to make dichotomous classifications. The resulting classification will suggest that the student is either "college-ready" or "not college-ready", which becomes interpreted as an attribute of the student. The appropriate interpretation of the predictive cut score is that it reflects what students are likely to know and be able to do, only as far as the focal test can predict the outcome (Ho, 2012).

The shift from interpreting the predictive cut score as a probabilistic statement to making a dichotomous classification interpreted as an attribute of the student represents a

"construct shift" (Haertel & Ho, 2012). A "construct shift" happens when the developer construct (probability of achieving a future outcome) is different from the application construct ("college-ready" classification) because the performance labels add new meaning to the existing interpretation. This particular "construct shift" is problematic because a property of the test (focal-outcome test correlation) is confounded with an attribute of a student, which introduces construct irrelevant variance to the interpretation of the performance standard.

Biasedness depending on correlational strength with outcome measure. In this paper, I take the focal-outcome test correlation as a given in the sense that it is not a variable and the standard setting panel has to work within the constraints of the test. Often in the case of standard setting, policymakers and standard setting panels have the task of setting a cut score on a particular test that has been constructed according to test specifications that takes into considerations legal mandates, and the curriculum standards in existence. In this sense, the correlation between the focal test on which cut scores are being set and a given outcome test is fixed and cannot be varied.

However, it is possible that different focal-outcome test correlations may arise when different outcome measures are chosen. For example, focal-outcome test correlations may be higher in studies using concurrent outcomes compared to studies using future outcomes. The further out in time the future outcome is from the time when measurements are taken for the focal test, the weaker the correlations tend to be. Thus, the choice of different types of outcome measures may give rise to differences in focal-outcome test correlations, giving rise to predictive scores that are subjected to different degrees of deviation away from a stringency-only cut score.

Fairness of predictive cut scores. Issues of fairness in the use of predictive cut scores for college-readiness standards can arise in at least two ways.

The first issue of fairness can arise when different jurisdictions set predictive cut scores for college-readiness using different focal tests with different focal-outcome test correlations. Assume two jurisdictions, both of which have high-stakes tests for students and require students to score above the predictive cut score. For discussion purposes, let us assume that both jurisdictions use the same outcome measure to set predictive cut scores.

Assume that due to the specifics of the test construction and specifications, the focal-outcome test has a weaker correlation in jurisdiction A than in jurisdiction B. There will be a group of students who will be unfairly treated than if the test used in the other jurisdiction is available to them. If the criterion score is set above the average of the outcome test distribution, students in jurisdiction A that uses the test with weaker correlation may be penalized by having to reach a higher predictive cut score. If the criterion score is set below the average of the outcome test distribution, some students in jurisdiction A may miss out on opportunities to help them become "college-ready" because the predictive cut score is set more leniently.

The second issue of fairness may arise when the predictive relationship between one test and an external criterion differs among various subgroups, including by gender and ethnicity (Krug, 1966; Seashore, 1961, as cited by Cronbach, 1971). To ensure that a consistent procedure is used to set cut scores for subgroups where the focal-outcome test correlations differ, separate predictive cut scores would have to be estimated for each subgroup. However, if the differential predictive utility of the test arises because of

construct irrelevant factors, such as student and school background factors irrelevant to performance (see Haertel & Ho, 2012), then using the test to set a predictive cut score could constitute unfair use of the test. On the other hand, setting a common predictive cut score as if the predictive utility of the test were the same across all subgroups would constitute unequal application of procedures.

C.2 Judgment required to select inputs for prediction

The appeal of predictive standard setting over judgmental-based standard setting is that it appears more objective and less arbitrary. While this may be true for a given set of inputs to the prediction method, i.e. the data "speaks for itself" when the prediction method, criterion score, and statistic to be modeled are fixed, the choice of the inputs are subject to judgment and decisions of the implementers of the predictive method.

Choice of prediction method. Ordinary least squares regression, logistic regression, and quantile regression are all reasonable prediction methods that can be used to identify predictive cut scores. The results from RQ1bi suggest that at a typical correlation where r=0.3, the difference in the location of predictive cut scores identified by OLS regression and median quantile regression (QR50) on one hand and logistic regression with 50% probability of scoring at or above the criterion score (LR50) may range from negligible to small-sized, for criterion score scores within ±1 standard deviation units of the average outcome. For what is practically considered as a very strong correlation of r=0.5, these differences will range from negligible to small-sized.

Differences in predictive cut scores arise because each method gives rise to slightly different interpretations for the cut score and what can be known about students scoring at different parts of the focal test distribution.

In ordinary least squares (OLS) regression, the predictive OLS cut score has the following interpretation: On average, students who score at the predictive OLS cut score are expected to score at the criterion score on the outcome test. Other than in distributions with standard bivariate normality, the OLS prediction model provides no information about the probability or percentage of students who will score above or below the criterion score at each focal test score level. Depending on the position of the focal test score level with respect to the predictive OLS cut score, we will know the relative position of the conditional average of outcome scores at that score level with respect to the criterion score.

In logistic regression, the predictive logistic regression cut score has the following interpretation: Students scoring at the predictive logistic regression cut score have a p% probability of scoring at or above the criterion score on the outcome test. Students scoring above the predictive logistic regression cut score have a greater than p% probability of scoring above at or above the criterion score while students scoring below the predictive logistic regression cut score have a lower than p% probability of scoring at or above the criterion score. Logistic regression may be popular as a predictive standard setting method because it provides a person-centric interpretation to the probability of an occurring event (scoring above the criterion score, or being "college-ready").

In quantile regression, the predictive quantile regression cut score has the following interpretation: (100-q)% of students scoring at the predictive q^{th} quantile regression cut score will score above the criterion score on the outcome test. At each score level above the predictive quantile regression cut score, (100-q)% of students will score above the predicted outcome score. At each score level above the predictive

quantile regression cut score, (100-q)% of students will score below the predicted outcome score. If ensuring that a certain percentage of students will score above the criterion score is important at the cut score is important, then using quantile regression would be useful.

Because each prediction method models a different statistic, differences arise in the predictive cut score. Depending on where the criterion score is set, whether and how distributions depart from bivariate normality, the differences may range from negligible to being substantially different under correlations typically observed between focal-outcome tests. The choice of a prediction method is thus a non-trivial problem.

Choice of prediction probability or quantile of performance. To use logistic regression or quantile regression, the prediction probability and quantile of performance has to be specified respectively. Equations 2 and 3 suggest that predictive cut scores can be identified for a given criterion score over a distribution of values for probabilities p and quantiles q where $p \in (0, 100)$ and $q \in (0, 100)$ (expressed in percentage) when logistic regression or quantile regression is used.

In the case of logistic regression, the choice of prediction probability has ranged from 50% to 75% for setting college-readiness standards (see ACT, 2004; Kobrin, 2007). This probability has been based on the response probability (RP) criterion often used in judgment-based standard setting. In the traditional standard setting literature, specifically the Bookmark Method and item mapping procedures in which response probability is most commonly used, the selection of a suitable RP value has been widely studied and debated (see Karantonis & Sireci, 2006). Commonly justified RP values have included 50% and 67%, values at which item information is maximized depending on whether

there is guessing and the type of IRT model used (Huynh, 1998; Wang, 2003). Other RP values have included 65% or 74%, but this is based on a slightly different context of item mapping and score anchoring in which the goal is to produce an adequate number of exemplar items (Kolstad et al., 1998, as cited by Zwick, Senturk, & Wang, 2001). Zwick, Senturk, and Wang (2001) found that experts favor 70% to be the minimum percentage of correct responses to consider that students "can do" an item. A National Academies of Sciences study (Hauser, Edley, Koenig, & Elliott, 2005) found that when RP values of .50, .67, and .80 are used to set standards via the Bookmark Method for an assessment of adult literacy, different cut scores were produced. The study also acknowledged that in different fields and contexts, experts may favor different RP values.

In the case of quantile regression, there has been no past precedence for its use in predictive standard setting to the best of our knowledge, nor any literature discussing suitable quantiles of performance to model.

Our results from RQ1bii using simulated data that meet bivariate normality assumptions suggest that at correlations of r=0.5 or below, medium to large differences in predictive cut scores may emerge when logistic regression probabilities are set at 50%, 65%, or 75% for criterion scores ±1 standard deviation units of the average outcome. Similarly, for quantile regression, medium to large differences in predictive cut scores are observed when the quantile of performance is set at 50^{th}, 35^{th}, or 25^{th} quantile. Predictive cut scores identified using the empirical dataset show these patterns as well. The choice of prediction probability or quantile of performance can thus give rise to substantially different cut scores.

More work needs to be done to study whether the response probability used in judgmental-based standard setting and its findings apply to predictive probabilities in logistic regression for predictive standard setting, and what might be appropriate quantiles of performance in quantile regression. Our results show that substantial differences in predictive cut scores arise when different predicted probabilities or quantiles of performance are selected.

Choice of criterion score. The use of predictive standard setting requires judgments about the criterion score that set good enough standards (Beaton, Linn, & Bohrnstedt, 2012). For college admissions tests such as the SAT and ACT, college-readiness benchmarks are provided across a range of criterion scores from B to C (ACT, 2004; Kobrin, 2007). In the case of these college-admissions tests, providing a range of criterion scores give colleges the flexibility to pick and choose the criterion score that fit their needs. However, in K-12 testing, policymakers need to decide the criterion score level that will be pegged to a given performance level, which in turn may have consequential decisions attached for students who meet or do not meet that performance level.

The results in Table 1 show that as expected, the predictive cut score increases as the criterion score increases. As such, the predictive cut score set is subject to judgments about where the criterion score level should be set. Moreover, the amount the predictive cut score increases in proportion to the increase in criterion score is also subject to the strength of correlation between focal test and outcome test.

Equation 2 also shows that there is a joint distribution of (p, X_c) that satisfies the logistic regression equation for achieving a particular criterion score. Similarly, a joint

distribution of (q, X_c) exists for a given criterion score when quantile regression is used. It is thus technically possible to manipulate the values of p or q in order to arrive at a particular cut score, further suggesting the need for judgment in predictive standard setting.

Judgments required for other inputs. So far, I have considered the judgments required for selecting the prediction method, prediction probability or quantile of performance, and the criterion score. There are also at least three other critical inputs for using predictive methods: an appropriate construct to base predictions of "college-readiness" on and a corresponding outcome measure; the focal test score distribution; and the analytic sample. Using an empirical dataset, I illustrate how analytical decisions to deal with non-linearity and non-normality can result in unpredictable differences among empirical cut scores. Suffice to say, judgment is also required in the application of empirical methods.

For the rest of this section, I discuss a number of other issues with the inputs for predictive standard setting.

C.3 Issues with test score distributions

The use of predictive methods is predicated on the availability of a well-defined construct on a well-defined scale on both the focal and outcome tests. However, many K-12 tests do not have continuous normal distributions (Ho & Yu, 2015). As such, alternative model specifications have to be used to deal with non-linearity and non-normality in the data.

In this paper, I illustrate the how skewness in the focal test score distribution may affect equipercentile and predictive cut scores. I find that these departures from normality

in the focal test score distribution have somewhat of an effect on predictive cut scores, but the magnitude of the effect does not affect our broad finding that predictive cut scores deviate from stringency-only cut scores as focal-outcome test correlations weaken, and that choices made to specify the criterion score and the probability of correct prediction or quantile of performance can result in substantial differences in predictive cut scores.

C.4 Issues with outcome construct

Predictive standard setting requires selecting, defining, and measuring a suitable external criterion. College-readiness is most commonly used, but is it suitable when the focal test is a K-12 subject test? In this section, I first discuss college-readiness and its suitability for setting predictive standards for K-12 subjects. Then I discuss issues with the GPA scale before critiquing the suitability of first-year GPA and introductory college course grades which have commonly been used as measures for college-readiness.

External criterion for setting predictive standards. "College-readiness" has become the de facto lingo representing the goal of K-12 education to prepare students for college. While "college-readiness" may be appropriate for general communication, it may not necessarily be the most suitable construct or external criterion to base K-12 predictive standards on.

As a construct, "college-readiness" is multi-faceted. Operationally, it is defined as "the level of preparation a student needs in order to enroll and succeed – without remediation – in a credit-bearing general education course at a postsecondary institution that offers a baccalaureate degree or transfer to a baccalaureate program" (Conley, 2007, p.5). It encompasses not only academic knowledge and skills, but also other factors such as motivation, behavior, cognitive and study skills, and contextual skills and awareness

about college (Conley, 2007; Gaertner & McClarty, 2015; Wiley, Wyatt, & Camera, 2010). Several studies that used first-year GPA as a measure for college-readiness have found that "noncognitive" characteristics such study habits and willingness to seek out support, and involvement in high school activities have a statistically significant relationship with first-year college grades (Pascarella and Terenzini, 1991; Williford, 1996, as cited by Pike and Saupe, 2002).

On the other hand, "academic preparedness" focuses on the academic and cognitive aspects of "college-readiness". The Technical Panel on 12th Grade Preparedness Research convened by the National Assessment Governing Board (NAGB) (2009) states that:

> "[p]reparedness for college refers to the reading and mathematics knowledge and skills necessary to qualify for placement into entry level college credit coursework without the need for remedial coursework in those subjects. … Academic preparedness is separate and different from college readiness because, in addition to academic skills, readiness encompasses behavioral aspects of individual performance related to success, and these additional attributes are not measured by NAEP. Examples of readiness characteristics include persistence, time management, interpersonal skills, and knowledge of the context of college." (p.3)

This is a key reason why the NAEP preparedness research has focused on "academic preparedness" "because this is what grade 12 NAEP is best equipped to measure, but also because academic skills in reading and mathematics constitute an important and foundational dimension of readiness" (NAGB, 2009, p.3).

Another issue relevant to the selection of an external criterion is: when should the construct be measured relative to college? Should it be at the start or after college results are available? Most measures of college-readiness used are obtained after college results are available (Camera, 2013), such as first-year GPA or entry-level course GPA. However, such measures would be affected not only by college-readiness, but also affected by factors during the first semester or first year of college, including the quality of the college and instruction, students' non-academic preparation for and adjustment to college, motivation, behavior, and personal circumstances. As an external criterion for gathering validity evidence about cut scores set by traditional standard setting procedures, using first-year GPA may be useful empirical evidence for checking whether the cut score is set within a reasonable range. However, to base the setting of cut scores on the prediction of college performance would expose the predictive cut score to factors relevant to success in college but not directly relevant to the level of academic knowledge and skills required in a performance standard.

The NAGB (2013) in its definition of preparedness has focused on the knowledge and skills at the entry point to college. Although measuring “college-readiness” at the entry point appears more conceptually appropriate, in practice, such data is typically not available, other than through admissions and placement tests (Camera, 2013). As I will explain further later, admissions and placement tests often set their college-readiness standards based on college grades, which lead back to the same issue that the measure of college-readiness is still affected by college-related factors.

In selecting a suitable external criterion to set predictive standards on, I draw attention back to the purpose of standard setting, which is to set cut scores that represent

what the minimally qualified student knows and is able to do, at some desired level of competence. K-12 tests are similar to NAEP in the sense that they are subject tests that focus on academic knowledge and skills. When "college-readiness" is used as the outcome construct, behavioral, psychological, and contextual factors that are necessary for college success would introduce construct irrelevant factors that K-12 subject tests are not equipped to measure, and would deviate predictive cut scores away from a stringency-only standard that focuses on what students should know and be able to do. "Academic preparedness" measured at entry point to college such as that defined by NAGB (2013) may be more conceptually appropriate as the external criterion because of their focus on the requisite knowledge and skills that students need to be prepared for college-level courses, but before college factors have a chance to come into play.

Prediction versus setting predictive standards. Before concluding this section on the outcome construct for college-readiness, I make a distinction between prediction and predictive standard setting. Both have been widely discussed in the college-readiness literature and both are concerned with setting "benchmarks", but they are not clearly distinguished. Camara (2013) succinctly summarizes a study by Maruyama (2012) which describes the problem: "there is a logical inconsistency in developing college readiness benchmarks from a test score when research consistently demonstrates that multiple factors are the best predictor of college success" (Camara, 2013, p. 19). The distinction lies in the goal and the variable of interest in prediction and predictive standard setting.

In prediction, such as predicting college readiness, the goal is to identify students who are at risk of not being successful in college. The focus is on the outcome of interest, college readiness in this case. The operational goal is to find a set of predictors that best

explain variation in college readiness, to maximize variance explained, and if possible, to maximize the classification accuracy rate. As discussed earlier, academic knowledge and skills is important for predicting college readiness, but so are a number of other factors.

However, in the case of predictive standard setting, the goal is not correct prediction, but identifying a level of performance that meets a certain academic standard, or level of competence. The variable of interest is the focal test score, and setting a cut score on the focal test that indicates an appropriate level of performance. An implicit assumption of an academic performance standard is that students who meet the standard will have sufficient foundation to progress to the next level and benefit from learning in the next level, below which students may lack the requisite skills and knowledge to keep up in the next level. The focus is on the requisite knowledge and skills, as measured by the focal test. What may be construct relevant for predicting college readiness may be construct irrelevant to setting a stringency-only cut score on a K-12 subject test.

Cureton (1951) provides an example to illustrate how a test built to optimize prediction may run counter to what is desired from an instructional standpoint. In Cureton's scenario, the task is to predict students who in real-life German writing, will punctuate most correctly and make the least amount of punctuation mistakes. The predictor test used is a test passage of unpunctuated German. Experimental evidence, as Cureton's scenario goes, found that students who put in the largest amount of correct punctuation, *as well as* the largest amount of incorrect punctuation on the test are the ones who have the desired performance in real-life. Therefore, in the scoring of the predictive test, "some positive credit [has to be given] for incorrect punctuation as well as more credit for correct punctuation" (p.633) even though incorrect punctuation is

undesirable from an instructional perspective. Cronbach (1971) cites a study by Kelly (1966) which found that medical-student performance was better predicted by "interests and other noncognitive variables than by ability measures" (pp.488-489). Both examples illustrate the key purpose of prediction.

An example of a study concerned with prediction of college readiness is found in Gaertner and McClarty (2015), who propose a college-readiness index for middle school students. The college-readiness composite (consisting of SAT, ACT and high school GPA scores) is based on more proximal outcomes for middle school students enroute to college readiness. The college-readiness indicators cover a diverse set of predictors, including academic achievement, motivation, behavior, social engagement, family circumstances, and school characteristics. Together, the indicators explain 69% of the variance in the college-readiness composite, corresponding to about r=0.83 between index and outcome composite. The "benchmarks" set provide a numerical summary of the student's joint status on each of the predictors with respect to the probability of scoring above or below a criterion score on the outcome. Students may reach an academic standard for being college ready, but if they are high in other risk factors for not being college ready, such as having poor study and organization skills, or experiencing financial hardships, then they may not meet the benchmark for being college-ready. This benchmark may include information about students' academic performance, but it also encompasses a variety of factors that have nothing to do with academics. In this sense, the "benchmark" is closer to a risk score rather than a performance standard.

This example also illustrates that even with a diverse set of predictors, the prediction of college-readiness will still be imperfect at best, which is good news for students in the sense that the past (and circumstances) does not determine the future, but bad news for building a predictive model to identify cut scores. The implication is that to set standards for K-12 academic tests, the correlation between the K-12 test and the measure of college-readiness will practically be low due to factors embedded in the college-readiness measure that are construct irrelevant to what is measured by K-12 tests. We can be certain that this low correlation will deviate the predictive cut score away from a cut score that reflects stringency-only standards.

Issues with the GPA scale. The use of prediction methods is predicated on the availability of a well-defined outcome scale. College academic scores, or the GPA scale, are often used as the outcome scale for setting college-readiness standards. However, a number of issues exists with the GPA scale.

Firstly, individual college courses are often graded differently subject to individual instructors' grading policies, even within the same department, or on the same course taught by different instructors (Johnson, 2003). Some instructors grade on a curve, i.e. use a norm-referenced scale. Others grade on students' "absolute performance" with respect to the content taught, where the instructor is the arbiter of the definition of "absolute performance".

There are also concerns that course grades reflect not only standards for students' performance, but may be subject to grade inflation (Rojstaczer & Healy, 2012). In an environment where course evaluations form part of faculty hiring, promotion, and tenure

decisions, instructors may have incentives to award higher grades (Eiszler, 2002; Moore & Trahan, 1998).

GPA scales may also not be comparable across different major fields (Goldman, Schmidt, Hewitt, & Fisher, 1974) or colleges. Some studies conducted within colleges in the 1960s (Aiken, 1963; Hills, 1964) found that even though college admissions became more selective in admitting students with more competitive grades, their college GPA did not rise in tandem. One possible reason put forth was because faculty tend to award grades based on the average performance of the group, so that as the average performance of the group increased, the expectations for average performance also increased (Hills, 1964). On average, selective and non-selective colleges admit students at different parts of the performance spectrum. Once enrolled in the college, these students are graded on a 4.0 point GPA scale. An average of B+ on the GPA scale in a selective college may reflect very different performance from an average of B+ GPA in a non-selective college, especially if the course is graded on a curve. Thus, the GPA scale may not represent a common scale across different colleges.

These are issues peculiar to the choice of GPA as an underlying scale for the outcome measure, which would be relevant whether introductory college courses or first-year GPA are used as an outcome measure for either the academic preparedness or college readiness construct respectively. One needs to assess whether the GPA scale meets the requirement as a "well-defined scale" for college-readiness.

Issues with using introductory course grades. Introductory course grades are subject to the above-mentioned issues with the GPA scale. Additionally, introductory course grades may be subject to sample bias. Academically advanced students with AP

credits may skip introductory courses and proceed to more advanced college courses. As such, introductory course grades may be biased downwards and affect the predictive cut score set.

Issues with using first-year GPA. First-year GPA as a measure can also depend on the combination of courses that an incoming freshman takes.

As discussed earlier, more academically advanced students may take more advanced college courses in their first year, while less academically advanced students may take introductory college courses. This will weaken the predictive power of the K-12 focal test for first-year GPA. It also introduces academic advancedness as a construct-irrelevant factor to using first-year GPA as a measure for college-readiness.

The mix of first-year courses taken by students across majors may also differ. A first-year social science major may take more courses in areas such as writing, and the social sciences, while a first-year engineering major may take more courses in areas such as math and sciences. In using college admissions tests to predict college-readiness, one common solution is to ignore differences across majors and course sequences, and directly aggregate course grades taken in the first year into the first-year GPA. The problem when first-year GPA is used as the outcome test to set cut scores on K-12 subject focal tests, a mix of subject grades is used to set the standard for a specific subject, such as math or ELA. This may be a hidden problem when a college admissions test is used as an intermediate outcome test to set cut scores on a K-12 subject focal test, but the ultimate criterion for the standard on the college admissions test is based on first-year GPA.

In summary, there is wide variation in which instructors, departments and colleges set standards for the GPA scale, and wide variation in which the scope, sequence, and difficulty of courses taken in the first year of college differs across students, majors, and how advanced academically they are. These factors contribute to issues with first-year GPA as a well-defined scale, which calls into question its suitability as an outcome measure that forms the basis for setting K-12 academic college-readiness standards.

C.5 Issues with the analytic sample

A number of issues also exist with constructing a representative analytic sample for setting college-readiness standards. Judgments will have to be made to address these issues. The issues may also cause predictive cut scores to be different from that if a representative sample were available.

Selection bias. To set predictive standards, the analytic sample has to be representative of the population taking the focal test. Selection bias can occur because students at the bottom of the K-12 performance distribution would likely not enroll in college, and hence are excluded from the analytic sample when college academic performance is used as the outcome measure. This will be a problem regardless whether the prediction is based on existing data for college performance such as GPA, or whether the focal test is only administered to college students.

Range restriction versus comparability of GPA scales. In order to set predictive cut scores that reflects performance across the student population, the analytic sample also requires a representative sample of colleges. In particular, restriction of range

may occur when specific colleges accepting students from a narrow range of performance are included in the sample.

There may be a trade-off between the representativeness of colleges and the comparability of the GPA scale. When a more narrow-range of college types which admit students from specific score ranges are included in the sample, the GPA scale may reflect a similar range of college performance, making the ensuing GPA scale more comparable across colleges. However, this may give rise to a restriction of range problem. As discussed earlier, where correlations are attenuated, predictive cut scores are skewed away from a stringency only standard. Where the sample of colleges are more representative of the college student population, a wider range of performance is measured on the same GPA scale due to the selectivity of colleges, resulting in a GPA scale that may not be comparable across colleges.

Changing distribution of student performance. College readiness is not an immutable quality. The notion of using tests to predict college readiness, especially at the lower grades, is so that students and educators can take some course of action to change their future outcome for the better over their predicted outcome. Ideally, students identified as not being college-ready in the earlier grades would be identified for intervention. If the interventions are effective, these students should gain proficiency towards college-readiness. When such interventions are effective on a systems level, we would expect the distribution of test scores to change. A time extrapolation principle of predictive testing (Cronbach, 1971, p.485) arises:

> A study that predicts success by a statistical formula has clearest significance when the formula is developed in the locale of the proposed application and the

> situation is sufficiently stable that the findings are representative of what will happen in succeeding years. Only if the supply of applicants and the curriculum remain much the same in character are the findings likely to remain directly applicable.

Practically, predictive cut scores for various grades would be set in year *y* based on prior years' data. The predictive cut scores would be set for various grades based on a population of students who did not receive intervention on the basis of being identified as "not college-ready". In year *y*+1, the year immediately after predictive cut scores have been set and used to identify students at risk of not being college-ready, students would have received 1 year of intervention. This would change the distribution of student performance on the predictor test with respect to the distribution in the previous year, and the predictive cut score set in previous years may no longer apply. Going forward for each successive year, we would have cohorts of students who received increasingly more years of intervention. For predicting college readiness from grades 3 to 12, it would take at least 9 years to reach steady state. Furthermore, the eventual distribution of student performance on the outcome test could also potentially change, if the interventions work well enough to influence the college-readiness of students.

In theory, the relationship, especially the correlation, between focal test scores and outcome test scores have to be closely monitored for changes. The standards set by predictive methods would have to adjusted accordingly, presenting a substantial data challenge.

C.6 Impact of GPA scale and analytic sample issues on empirical cut scores

The above issues with GPA scale and analytic sample affect all empirical methods that use a criterion score to set cut scores on the focal test, including the equipercentile method. I illustrate this when equipercentile linking is used.

In the first example, I consider a case when the outcome score scale is shifted downwards from a "true" outcome score scale. For simplicity, assume the analytic sample is representative of the population, and the outcome scale used in the analytic sample is similar in all aspects to that used in the population, except that it is shifted downwards from the outcome test score scale in the population. The equipercentile cut score obtained using the analytic sample would then be biased downwards from the college-ready criterion score measured by the "true" outcome score scale used in the population.

In the second example, I consider the case of selection bias. If selection bias exists such that the distribution of outcome test scores in the analytic sample is biased downwards from the population distribution, the equipercentile cut score obtained using the analytic sample would also be biased downwards from the college-ready criterion score in the population.

Appendix D
Evidence-Based Standard Setting: Suitability of Predictive Standard Setting to Identify Neighborhoods of Potential Cut Scores

Having discussed the issues with predictive standard setting as a stand-alone standard setting method, I discuss evidence-based standard setting in this appendix. An implicit assumption of evidence-based standard setting is that using different sources of evidence and types of empirical methods would give rise to different cut scores. However, the convergence of evidence around a particular cut score region would suggest where the "true" performance standard likely lies.

The strength of evidence-based standard setting is that it lays out a systematic way for standard setting panelists to consider the external validity evidence in the process of setting standards. However, central to the validity argument of this standard setting method is the quality of the evidence. I start first by introducing the types of evidence that evidence-based standard setting typically relies on, before critiquing the quality of the evidence, and issues with evidence-based standard setting.

Common types of evidence in EBSS. Evidence-based standard setting commonly uses empirical evidence from concurrent studies, predictive studies and longitudinal studies (see McClarty et al., 2013).

Concurrent studies use external criterion evidence that is gathered from students at around the same time as the focal test. When focal tests are end-of-course tests, the external criterion evidence may include other tests that high school students typically take, such as different tests within the same content area, high school course grades; or college admissions test such as the SAT or ACT.

Predictive studies and longitudinal studies use external criterion evidence gathered from students at a future time from the focal test. Predictive studies, as it is referred to in McClarty et al. (2013) involves administering the focal test to students before they start a college-level course, and collecting the college grades at the end of the college-level course. Longitudinal studies referred to in McClarty et al., (2013) involve collecting grade-to-grade test scores. I term these grade-to-grade test scores "intermediate outcome tests". The criterion score in the "intermediate outcome test" is typically an evidence-based cut score identified by an earlier predictive standard setting process in which the ultimate criterion is a specified level of performance in a college outcome measure (e.g. see Texas Education Agency, 2013).

Empirical method used. Regardless of the type of external criterion used, all cut scores derived from regression-based methods will suffer from the similar problem encountered in predictive standard setting, in that identified cut scores deviate away from the stringency-only cut score when focal-outcome test correlations are less than unity. Having many regression-based cut scores from various studies converge in one neighborhood does not imply that cut scores are converging around a "correct" cut score.

Although equipercentile-based cut scores avoid confounding the correlational strength of the focal-outcome test with the stringency of the standard, equipercentile-based cut scores can be subject to selection bias when the analytic sample is not representative of the student population. However, when the analytic sample is representative of colleges, an issue of the comparability of

GPA scales may arise, which may in turn affect the level where equipercentile-based cut scores are set at.

Definition of what students know and are able to do. For both regression-based and equipercentile-based cut scores, the description of what students know and are able to do comes after the cut score is identified. This description is test-dependent and potentially could differ from test to test, even if the correlations are similar. In this sense, the performance standard is circumscribed by the test as a measurement tool, and not an "absolute" criterion based on the knowledge and skills needed to be college-ready.

External criterion used. There are various factors to consider in assessing the quality of evidence when various external criteria are used.

***Concurrent versus future outcome*.** Concurrent outcomes often have a higher correlation with the focal test than future outcomes. The identified cut scores based on concurrent outcomes would thus be less susceptible to deviation from a stringency-only cut score.

College versus non-college outcomes. To predict college-readiness standards, it appears to make sense to base the external criterion on college outcomes. However, a number of issues may arise from the analytic sample used. Selection bias may be a problem because students who are not college-ready will not enroll in college, and thus their college results would be missing. Range restriction may also occur when the colleges sampled are not representative of the college population. Since the direction of bias is known, it is possible to apply a correction formula to address both of these issues.

However, when a representative sample of colleges is used, the comparability of the GPA scale across colleges may become an issue.

Construct irrelevant factors may also be introduced when college outcomes are used. The quality of college, social and personal factors related to college may all affect a student's success in college. Construct irrelevant factors would thus affect both predictive and equipercentile cut scores.

To avoid the issues associated with use of college outcomes, it is theoretically possible to select constructs similar to non-college outcomes but which are not college outcomes as the external criterion. However, as Kane (2001) noted in the context of using other criteria as external validity evidence, the other criteria are often open to question themselves.

One common alternative to college grades are results for admissions tests, such as the SAT or the ACT. Part of the logic of using admissions test results is that the data is collected prior to college entry, and hence, better measure student preparedness for college (Gaertner & McClarty, 2015). Yet the targeted criterion scores that admissions tests use to indicate college-readiness often rely on college grades. The level of performance considered to be college-ready is thus also subject to college factors that affect eventual college performance.

First-year GPA versus course grades. First-year GPA may be attractive as an overall measure of college success. However, it combines results from a *mix of subjects* to set predictive cut scores for K-12 *subject tests*. The subjects other than that being tested in the focal test would introduce another source of construct irrelevance to empirical-based cut scores.

Intermediate versus ultimate outcome. I discuss here the special case of using an intermediate criterion set on intermediate outcomes. Often, grade-to-grade tests or college admissions tests are used as an intermediate outcome, which in turn predicts college grades as an ultimate outcome. The correlations between the focal test and the intermediate test outcomes are typically much higher than if college grades were used as the outcome. However, these apparently high focal-intermediate outcome test correlations belie the low correlation typically seen between the intermediate test and the ultimate outcome of interest, college grades. If the ultimate outcome has a low correlation with the intermediate test, then the cut scores set on the intermediate test would be too stringent or too lenient, which in turn would affect cut scores set on the focal test. Therefore, to assess the strength of evidence, one needs to identify the ultimate criterion that predictive cut scores are based on.

Independent or families of evidence? One advantage of evidence-based standard setting is that it collects information from a large base of evidence. However, one issue is whether the various sources of evidence are independent from one another, or they represent form families of evidence subject to similar issues and biases?

Take for example, if many of the studies use regression-based methods rather than equipercentile method, those studies would all be susceptible to issues that affect regression-based methods. If those predictive cut scores happen to converge at a location different from the equipercentile method, does that

strengthen the argument to recommend a cut score around the region of convergence?

Consider the case when empirical standard setting is used to set cut scores on an end-of-course focal test. Let's say two studies are conducted, one a concurrent study using the SAT as the outcome, another a predictive study using first-year GPA as the outcome. If the cut score set on the SAT ultimately uses first-year GPA as the outcome, should the two studies be considered as two independent pieces of evidence, or from the same family of evidence?

In evidence-based standard setting, special attention has to be paid to what is termed “salience” in behavioral psychology. Information that stands out is more likely to influence our thinking and action. When panelists are presented with a large number of cut scores in a particular neighborhood, panelists may subconsciously be conditioned to think that the region with the largest number of cut scores is the region where the eventual cut score should lie, when in fact, they need to also weigh the quality of the information that each identified cut score rests on.

References

Introduction

American Diploma Project. (2004). Ready or not: Creating a high school diploma that counts. Washington, DC: American Diploma Project, Achieve. Retrieved from https://www.achieve.org/publications/ready-or-not-creating-high-school-diploma-counts

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, *11*(1-2), 1-18.

Jerald, C. D. (2008). Benchmarking for success: Ensuring US students receive a world-class education. Washington, DC: National Governors Association, Council of Chief State School Officers, and Achieve, Inc..

Koretz, D. (2013). Commentary on E. Haertel, "How is testing supposed to improve schooling?" *Measurement: Interdisciplinary Research and Perspectives, 11*(1-2), 40-43.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing: Preliminary evidence about generalization across tests. In R. L. Linn (Chair), *The effects of high stakes testing*. Symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis*, *32*(1), 5-23.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development (2010). ESEA Blueprint for Reform, Washington, DC.

Essay 1

ACT (2004). Crisis at the core: Preparing all students for college and work. Iowa City, IA: ACT.

Aiken Jr, L. R. (1963). The grading behavior of a college faculty. *Educational and Psychological Measurement*, *23*(2), 319-322.

Beaton, A. E., Linn, R. L., & Bohrnstedt, G. W. (2012). *Alternative approaches to setting performance standards for the National Assessment of Educational Progress (NAEP)*. Washington, DC: American Institutes for Research.

Camara, W. (2013). Defining and measuring college and career readiness: A validation framework. *Educational Measurement: Issues and Practice*, *32*(4), 16-27.

Cizek, G. J. (2012). An introduction to contemporary standard setting. In G. J. Cizek, *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3-14). New York, NY: Routledge.

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701-731). Westport, CT: Praeger Publishers.

Common Core State Standards Initiative. (2010). *Common core state standards for mathematics*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

Conley, D. T. (2007). *Redefining college readiness*. Eugene, OR: Educational Policy Improvement Center.

Cronbach, L. J. (1971). Test validity. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 621-694). Washington, DC: American Council on Education.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*(4), 483-501.

Gaertner, M. N., & McClarty, K. L. (2015). Performance, perseverance, and the full picture of college readiness. *Educational Measurement: Issues and Practice*, *34*(2), 20-33.

Goldman, R. D., Schmidt, D. E., Hewitt, B. N., & Fisher, R. (1974). Grading practices in different major fields. *American Educational Research Journal*, *11*(4), 343-357.

Haertel, E. H., Beimers, J. N., & Miles, J. A. (2012). The briefing book method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 283-299). New York, NY: Routledge.

Haertel, E., & Ho, A. (2016). Fairness using derived scores. In N. J. Dorans, & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 217-238). New York, NY: Routledge.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger Publishers.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G.J. Cizek, *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47-76). New York: Routledge.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345-359.

Hauser, R. M., Edley Jr, C. F., Koenig, J. A., & Elliott, S. W. (2005). *Measuring Literacy: Performance Levels for Adults*. Washington, DC: National Academies Press.

Hills, J. R. (1964). The effect of admissions policy on college grading standards. *Journal of Educational Measurement*, *1*(2), 115-118.

Ho, A. D. (2012). Off track: Problems with "on track" inferences in empirical and predictive standard setting. Retrieved from http://scholar.harvard.edu/files/andrewho/files/off_track_-_andrew_ho_working_paper.pdf

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365-388.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, *23*(1), 35-56.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York, NY: Springer-Verlag.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3) 425-461.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4-12.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard. Washington, DC: American Institutes for Research.

Kobrin, J. L. (2007). Determining SAT benchmarks for college readiness. College Board Research Notes RN-30. New York, NY: College Board. Retrieved from: http://research.collegeboard.org/publications/content/2012/05/determining-sat-benchmarks-college-readiness

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average*. Research Report No. 2008-5. New York, NY: College Board.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, *15*(4), 143-156.

Lord, F. M. (1955). Equating test scores – A maximum likelihood solution. *Psychometrika*, *20*(3), 193-200.

Maruyama, G. (2012). Assessing college readiness: Should we be satisfied with ACT or other threshold scores? *Educational Researcher*, *41*(7), 252-261.

Mattern, K., Burrus, J., Camara, W., O'Connor, R., Hansen, M. A., Gambrell, J., ... & Bobek, B. (2014). Broadening the Definition of College and Career Readiness: A Holistic Approach. ACT Research Report Series, 2014 (5). Iowa City, IA: ACT, Inc.

McClarty, K. L., Murphy, D., Keng, L., Turhan, A., Tong, Y. (2012, April). Putting ducks in a row: Methods for empirical alignment of performance standards. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Vancouver, Canada.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher*, *42*(2), 78-88.

Moore, M., & Trahan, R. (1998). Tenure status and grading practices. *Sociological Perspectives*, *41*(4), 775-781.

National Assessment Governing Board (2009). Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report.

Phillips, G. W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods and innovations* (2nd ed., pp. 323-345). New York, NY: Routledge.

Phillips, G. W. (2014). International Benchmarking: State and National Education Performance Standards. Washington, DC: American Institutes for Research.

Pike, G. R., & Saupe, J. L. (2002). Does high school matter? An analysis of three methods of predicting first-year grades. *Research in Higher Education*, *43*(2), 187-207.

Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, *40*(2), 158-189.

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record*, *114*(7), 1-23.

Shaw, E. J. (2015). An SAT Validity Primer. New York, NY: College Board. Retrieved from https://files.eric.ed.gov/fulltext/ED558085.pdf

Texas Education Agency (2013). State of Texas Assessments of Academic Readiness (STAAR$_{TM}$) Assessments. Standard Setting Technical Report. Retrieved from https://tea.texas.gov/student.assessment/reports/

U.S. Department of Education, Office of Planning, Evaluation and Policy Development (2010). ESEA Blueprint for Reform, Washington, DC.

Wang, N. (2003). Use of the Rasch IRT Model in Standard Setting: An Item-Mapping Method. *Journal of Educational Measurement*, *40*(3), 231-253.

Wiley, A., Wyatt, J., & Camara, W. J. (2010). The Development of a Multidimensional College Readiness Index. Research Report 2010-3. New York, NY: College Board.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15-25.

Essay 2

High in Standards, Lenient in Stakes: The Consequences of Scoring Barely Below the Passing Score of High School Exit Exams in Minnesota

High in Standards, Lenient in Stakes: The Consequences of Scoring Barely Below the Passing Score of High School Exit Exams in Minnesota

Introduction

In 2010-11, about 65% of public school students lived in 25 states with high school exit exam passing requirements in place (Center for Education Policy, 2011). One key argument for high school exit exams is that they can motivate student and system efforts to promote learning (Dee & Jacob, 2008; Holme, Richards, Jimerson, & Cohen, 2010). However, a handful of papers have found unintended, causal consequences for students who barely fail the exam, mainly for students from disadvantaged backgrounds, such as higher high school drop-out probabilities (Ou, 2010), lower on-time high school graduation (Papay, Murnane, & Willett, 2010; Martorell, 2004) and lower college attendance (Martorell, 2004). Papay, Murnane, and Willett (2014) also find that barely failing an exit exam in Massachusetts reduces the probability of college enrollment even several years after the test.

The role of high school exit exams as part of standards-based educational reform has been extensively discussed within the literature (Amrein & Berliner, 2003; Dee & Jacob, 2008; Greene & Winters, 2004; Holme, Richards, Jimerson, & Cohen, 2010; Grodsky, Warren, & Kalogrides, 2009; Shuster, 2012; Warren, Grodsky, & Lee, 2008). Past research on the causal effects of high school exit exams fall into two main categories. The first type of research examines how the presence of exit exams impact overall student outcomes (e.g. Dee & Jacob, 2008; Jacob, 2001; Reardon & Kurlaender,

2009). The second type of research examines how exit exams impact students on the margins. My study falls under the second category.

Examining the impact of barely failing or barely passing high school exit exams for students on the margins is important from a policy perspective. High school exit exams hold significance for students especially in states where they are required for high school graduation. High school graduation is increasingly an important gateway into higher education (Bailey & Dynarski, 2011; Greene & Forster, 2003). In turn, postsecondary education is increasingly a determinant of labor market success (Autor, 2014).

This study examines whether scoring barely below or barely above the passing score of the Graduation-Required Assessment for Diploma (GRAD) in Minnesota has any impacts on students' high school and college outcomes. For the high school class of 2009 and earlier, students were required to pass a Basic Skills Test in order to be eligible for a high school diploma. For the high school class of 2010, Minnesota switched to the more rigorous GRAD with higher passing standards. However, for the math test, the passing standard was set very high such that a very high percentage of students could not pass it. To address the high failure rate, Minnesota waived the passing requirement on the math exit exam for high school diploma. Within this policy context where the passing standard is high, but in which the stakes are lowered, I examine the impacts of scoring barely below or above the passing score on high school and college outcomes. I use longitudinal data from the Minnesota Statewide Longitudinal Education Data System (SLEDS) for this study.

Collecting evidence on the consequences resulting from the use of tests for their proposed score interpretations is integral to the development and evaluation of tests (Kane, 2006). In 1999, at a time when high-stakes testing was gaining prominence, the National Research Council set up the Committee on Appropriate Test Use at the request of Congress. The Committee published a study that looked into the proper and fair use of high-stakes tests. At the time of writing, the committee wrote that "very little is known about the specific consequences of passing or failing a high school graduation exam, but a good deal is known about whether and how earning a high school diploma affects a student's future life chances (National Research Council, 1999, p. 176)."

A number of studies in the 2010s have used a regression discontinuity (RD) design to look at the causal impacts of passing or failing high school exit exams for students on the margins, and have focused mainly on high school outcomes such as academic course-taking, dropping out, and graduation (Ahn, 2014; Ou, 2010; Papay, Murnane, & Willett, 2010; Reardon, Arshan, Atteberry, & Kurlaender, 2009). It is left to the reader to infer the societal impacts as a consequence of not obtaining a high school diploma (see National Research Council, 1999). To my knowledge, two other studies have looked at the causal impacts of barely failing or barely passing a high school exit exam on longer-term outcomes such as employment (Martorell, 2004) and college enrollment (Papay, Murnane, & Willett, 2014). In this study, I seek to extend the literature by studying the consequences of scoring barely below or above the passing score of the math GRAD and reading GRAD in Minnesota on students' college outcomes, an increasingly important predictor of future success.

However, the study context is unique and different from past studies in the following ways. First, like the states in the other studies, the passing score for the math high school exit exam in Minnesota was set and announced prior to the first cohort of students sitting for the test. However, in Minnesota, the requirement to pass the math exit exam in order to be eligible for high school graduation was waived after the first cohort of students sat for it. Subsequent cohorts prepared for and took the exams knowing that they do not necessarily have to pass it in order to graduate. Second, the state law requires students who fail the exit exam to undergo remediation provided by the district. To my knowledge, this is unlike the states in other studies where opportunities for remediation depend on the districts that students attend.

Thus, in my study, the treatment group is slightly different from that found in other studies. Students who score barely below the passing score on the math GRAD would not be denied a high school diploma because of a failing score, but instead would receive district-prescribed remediation. Any differences in outcomes that I observe between this group of students and those who score barely above the passing score would be the net impact of scoring below the passing score on the math GRAD and the remediation.

In the next section, I discuss the various ways in which failing a high school exit exam may give rise to consequences for students who barely pass or fail. For studies that delve into the mechanisms by which these consequences arise, see for example, Reardon, Arshan, Atteberry, and Kurlaender (2009).

Background

High school exit exams and potential mechanisms

In general, states use high school exit exams to determine whether students have acquired the necessary knowledge and skills in key academic areas before they leave high school (Ferrara & DeMauro, 2006). Students are required to pass exit exams in specific subjects, most often in ELA or reading, and math, and sometimes various science subjects and social studies (see Center on Education Policy, 2011), before they are eligible to receive a high school diploma.

Students may use the subject pass/fail label on an exam to inform whether they have mastered the high school academic standards of that subject. This may affect decisions pertaining to high school outcomes. Some students who fail the exam may be motivated to work harder in order to pass the retest. Some students may be discouraged because of the failing label. Parents with a child who fail the exam may be motivated to invest more resources to help the student succeed on a second attempt, or parents may discourage students from staying on in school. On the other hand, students who pass the exam may experience an encouragement effect which motivates them to work harder in school and to meet the other requirements for high school graduation.

When the high school diploma is awarded based on the pass/fail status in required subjects, students who are awarded the high school diploma may interpret the diploma as indicative of having mastered the high school curriculum. Students whose diplomas are withheld may infer the opposite, even for those who may have barely failed.

Because the high school diploma is a minimum academic requirement for some colleges, students may also interpret their performance on exit exams and their pass/fail label as information on their preparedness for college. For students who meet all the exit exam passing requirements in addition to other requirements and are awarded the high

school diploma, this piece of certification may be a stamp of approval for them to move on to the next phase of education. Students who meet only the high school exit exam requirements in part, or not at all, may use the information to gauge their likelihood of obtaining a high school diploma and the likelihood of admissions into certain colleges. They may also use the results to gauge their suitability for various fields of study. The exit exam results may inform whether they apply to college at all, and the types of college and field of study that they eventually apply to and enroll in, all of which may have consequences for their college outcomes.

Last but not least, students' performance on exit exams may be used to determine their course offerings or tracks in the final year of high school. Especially in states where exit exams are taken before grade 12, districts and schools may use the exit exam results as a sorting mechanism for placing students into different tracks in their final high school year (see Ahn, 2014, for an example in North Carolina). Students who pass the exit exam may be deemed to have met the minimum academic standards for that subject in high school and be allowed to take more challenging courses or electives in their final high school year, which in turn may make their college applications more competitive. Students who fail the exit exam, however, need to retake the exit exam until they pass. The time used to prepare for exam retakes may be well-spent in the sense that the students are acquiring academic knowledge and skills important for success in college. However, especially for the small group of students who barely fail the exam, the time used to meet a graduation requirement may come at the cost of missing out on courses that may make them more competitive for college admissions.

Hence, the combination of the timing of high school exit exams in conjunction with the use of exit exam results to determine courses taken in the final high school year may result in consequences with respect to students' competitiveness for college admissions. Students who pass the exit exam subject may be placed on college preparatory tracks, which then becomes a self-fulfilling prophecy that students who pass the exit exam are "college-ready".

Remediation and potential mechanisms

The district-prescribed remediation constitutes another part of the "treatment" of barely failing the exit exam, which is distinct from failing the exam, but may also shape the consequences of barely passing or barely failing the exit exam. There are many educational settings in which students' test scores relative to a cut score are used to determine assignment to remedial programs (e.g. in elementary and middle schools, see Jacob & Lefgren, 2004; in middle schools, see Dougherty, 2012; Taylor, 2014), and the empirical evidence generally points to null or some positive effects (see Taylor, 2014). One key feature of the remedial "treatment" in this study is that despite being state mandated for students who fail the exit exam, it is "district-prescribed". Features of the remediation may vary greatly between one district to the next.

Remediation may result in improved test scores in a particular subject as a result of increased instructional time in that area (Dougherty, 2012; Taylor, 2014). To the extent that it is effective, remediation may have a human capital effect by enhancing the knowledge and skills of the students who take it compared to the students who barely pass and are not required (or even ineligible) to take it. It may help students who barely fail at the first attempt to pass at subsequent attempts, and possibly even better prepare

them for post-secondary education. If remediation includes components on college counseling, then students may be aided in their college application efforts.

However, remediation may also have negative consequences. It may be an inefficient use of students' time and distract them from subjects which may require more attention. It may also give rise to a labeling effect, especially if schoolmates and teachers can see which students attend remediation and thus infer who has failed the exit exam. However, depending on the student, labeling may provide a motivation to work harder, or a discouragement due to the stigma.

In short, the mechanisms through which barely passing or barely failing an exit exam can exert its effect in the short-term and longer-term through a complex mix of positive and negative influences, but these mechanisms may be difficult to disentangle. Instead, this study seeks to replicate past studies on whether there are any net impacts of these mechanisms, as a result of barely passing or barely failing a high school exit exam in Minnesota on high school dropout and withdrawal, and high school graduation.

Potential contributions

My study design is much like other previous studies: I study the impacts of barely failing the exit exam by subject. Hence, this study like a replication of previous studies. However, the context and the longer-term outcomes studied can enrich the literature in the following ways.

Firstly, the context that I look at is unique in that a pass/fail score for the math high school exit exam was initially set, but the requirement to pass the exam in order to be eligible for a high school diploma was later waived. The reason for the waiver was because the math passing standard was set very high such that the passing rate was

relatively low compared to previous years (Center on Education Policy, 2010). Furthermore, students were required to attend district-prescribed remediation to be eligible to receive the high school diploma despite failing the exit exam. If setting a passing score serves as a means to indicate a (high) performance standard required of students, while waiving the passing requirement removes a barrier towards high school graduation, and remediation provides greater support to help students pass the exam, then the combination of these three components may serve to mitigate some of the negative consequences that may arise due to the presence of a cut score. I call this policy context in which students take the Minnesota GRAD one which is "lenient in rules, high in standards".

To be able to study whether such a policy mitigates the negative consequences of barely failing a high school exit exam would require a counterfactual in which a passing score on a high school exit exam were set, and students who fail would be ineligible for a high school diploma. Unfortunately, such a counterfactual is not available because the math exit exam in Minnesota was newly introduced at the start of the study period[3].

It may be possible that without the passing waiver and district remediation, there may be severe negative consequences; while the change in policy might mitigate some of these consequences, we may still observe some negative consequences. It may also be the case where before the change in policy, there may be little, if not no consequences to begin with; and there remains no detected consequences after the waiver. In any case, the results of the study do not allow me to differentiate among the different scenarios. In

[3] The year the passing requirement for the math exit exam was waived was also the year when it replaced a basic skills test. Hence, a pre/post comparison of impacts would confound test change with the waiver of passing requirement. Minnesota was also undergoing other changes in assessment policies between the last cohort eligible for the waiver, and the first un-exempted cohort.

other words, I cannot answer the question of: what is the impact of the policy waiver on scoring barely below or above the passing score on the exam. Instead, the question I ask is a "status" question – with the particular policy context in place, are there any negative consequences? i.e., within such a "high on standards, lenient in rules" policy context, are there consequences of scoring barely below the passing score?

For more completeness and for similarity to the other studies, I also look at the impacts of scoring barely below or above the passing score on the reading GRAD in Minnesota[4]. Unlike for the math GRAD, there is no waiver of the passing requirement for the reading GRAD – students who fail will be ineligible for high school graduation. However, like math, students who fail the reading exam are also required to take district-remediation.

Secondly, I seek to extend the literature by looking at the impact of longer-term outcomes, including college enrollment and college graduation, which have increasing importance for the post-high school trajectories of students. Unfortunately, the post-secondary dataset has non-random missing data – it only includes college enrollment and graduation records for high school *graduates*, rather than for all students who take the exit exam. I discuss this data limitation when interpreting the findings.

I do not, however, look at the impact of imposing high school exit exams in general, or the impact of setting tougher standards for high school exit exams (see Clark & See, 2011; Reardon & Kurlaender, 2009). Finally, from this study, I am unable to distinguish whether the observed impacts of barely passing or failing the exit exam arise

[4] Although students are also required to pass a writing test, I did not include the writing test as part of this study because data was available for only one cohort of students, and students were rated on a score of 1 to 6, which may render it unsuitable for regression discontinuity analyses.

due to negative consequences of scoring barely below, or positive consequences of scoring barely above the passing score on the test, or both.

The Minnesota Context

Passing requirement for high school exit exam

Minnesota has had a relatively long history of high school exit exams. Since 1997, Minnesota has required students to take and pass the "Basic Skills Tests" (BST) in math, reading, and writing to ensure that they have a minimum level of knowledge and skills before graduating from high school (Minnesota Department of Education, 2010a). Beginning in 2000, diplomas were withheld for students who did not pass the three BSTs (Center on Education Policy, 2010).

In 2005, the Minnesota legislature enacted the Omnibus K-12 and Early Childhood Act of 2005 that had the Basic Skills Tests replaced (Minnesota Statutes, 2005). Students enrolled in grade 8 from the 2005-2006 school year and onwards must pass the Graduation-Required Assessments for Diploma (GRAD) in reading and math or obtain an achievement level equivalent to or greater than proficient on the Minnesota Comprehensive Assessments-Series II (MCA-II). Students are also required to pass the GRAD in writing. The writing, reading, and math tests are taken in spring of grades 9, 10, and 11 respectively. This statute was revised in 2007 to include options for retests to meet the graduation-testing requirement (Minnesota Session Laws, 2007). See Appendix A for details on the diploma requirements and Appendix B for policy timeline.

While the state uses either the GRAD or MCA-II to determine eligibility for high school graduation, there are some key differences between the two. The purpose of the GRAD is to measure the writing, reading and math proficiency of high school students

(Minnesota Department of Education, 2010a) while the reading and math MCA-II are used for state accountability purposes to meet NCLB requirements. The MCA-II results are also used to compare the performance of districts across the state and to provide feedback on curriculum and instruction (Minnesota Department of Education, 2010b). Both the GRAD and MCA-II benchmarks originate from the same set of academic standards. Some of the benchmarks in the GRAD and MCA-II overlap, while other benchmarks are unique to only the GRAD or the MCA-II respectively.

The administration of the math GRAD is embedded within the MCA-II assessment, i.e. students take both assessments in one seating. The math GRAD consists of 40 items in multiple-choice format while the math MCA-II consists of 55 items, 45 items in multiple-choice format and 10 items in gridded response format. Of these, 25 items are common to both the GRAD and MCA-II. In the data, I observe a correlation of about 0.94 between the GRAD and MCA-II.

For both math and reading, GRAD is the key assessment for considering eligibility for high school graduation. However, students who pass the MCA-II but fail the GRAD are also considered eligible for high school graduation[5]. Students who fail the GRAD on the first attempt are allowed to retake the GRAD component.

Waiver of the passing requirement on math GRAD

Amid concerns that failure rates on the math GRAD might be higher than anticipated, a legislative task force was established in the late 2008 and early 2009 to review the GRAD passing policy and its implication for high school graduation (Center

[5] Based on conversations with Minnesota Department of Education officials, the MCA-II is deemed to be based on more rigorous standards than the GRAD, hence this exception for students who passed the MCA-II but failed the GRAD.

on Education Policy, 2010). Options for a short-term remedy and long-term direction were discussed. In May 2009, the governor signed a law (Minnesota Session Laws, 2009) that waived the passing requirement on the math exam. Students in the high school classes of 2010-2014 are not required to pass the math GRAD or obtain a proficient score on the math MCA-II in order to be eligible for a high school diploma.

Under this updated Education Bill, students in the high school graduation classes of 2010-2014 would meet the state math graduation requirement by meeting state and local coursework and credits requirements, and by scoring at or above the passing score for the MCA-II or math GRAD component (Minnesota Statutes, 2009; Minnesota Department of Education, 2010). Students who could not meet the passing requirement on both the math MCA-II and the math GRAD would still be eligible to receive a high school diploma by: (i) participating in district-prescribed academic remediation in math, and (ii) fully participating in at least two retests of the math GRAD or until they pass the math GRAD[6]. Based on conversations with Minnesota Department of Education officials and district personnel, the retest requirement was also waived[7] (SLEDS coordinator, personal communication, 6 September, 2018; district personnel, personal communication, 1 October, 2018).

Between cohort variation of treatment. The timing of these changes to the passing requirement for math GRAD gives rise to variation in the treatment from cohort

[6] However, students will still receive a "not pass" notation on their transcript if they are unable to pass the test by the high school graduation date. See Appendix A for an example of the transcript notation when pass waiver is in place.

[7] I was unable to determine when the retest requirement was waived. Part of the challenge was identifying personnel present at the time when the policy was effected. Review of documents from the Department of Education suggests that the retest requirement was in place for all five cohorts of students. However, the district personnel whom I spoke with recalled that the policy went from "three attempts" (including the initial attempt and up to two retests), to "any attempt" (retests not required).

to cohort. For the cohort of students whose initial math GRAD attempt was in 2009, the students prepared and sat for the exam under the context that they had to pass the exam in order to be eligible for a high school diploma. Within a month, the law to waive the passing requirement was signed. Therefore, the cohort of students whose initial math GRAD attempt was in 2009 differs from subsequent cohorts in their understanding of the passing requirement when they prepared and sat for the exam, and for a short period of time after they sat for the exam. This gives rise to the possibility that students may withdraw or drop out from high school due to their perceived performance on the exam, before they learn about the waiver.

For the cohorts of students whose initial math GRAD attempt was in 2010 and after, the students prepared and sat for the exam knowing that they could fail the exam, but still go on to graduate from high school provided that they meet the other requirements. However, even amongst these cohorts, there is potential for other differences. Based on anecdotal accounts, it appears that the retest policy has become more lenient over time ("three attempts" to "any attempts"). To the extent that the retesting requirement imposes additional effort and psychological barriers on students, the leniency over time might lower barriers towards high school graduation. Thus over time, we might expect little to no net impact of consequences for students who barely fail versus those who barely pass. I thus conduct analyses by cohort to address the different passing requirement from cohort to cohort.

Differences in passing requirement between Minnesota and other states. The key difference between Minnesota and most other states in past studies, including California, Massachusetts, New Jersey, North Carolina, and Texas, is whether the passing

requirement for the high school exam is upheld. For most states, students are required to obtain a passing score on the high school exit exam in order to graduate. In Minnesota, there is a passing score for the math exit exam, but it is not necessary for students to score above it in order to graduate from high school. On this dimension, the reading test in Minnesota is similar to the other states since the passing requirement is upheld.

Another difference is that in Minnesota, the state requires that districts provide remediation to students who fail the exit exam, although districts have the discretion to determine what that remediation might be. This is the case both for reading and math[8]. In most other states, such as Massachusetts, and California, it does not appear that such a rule exists on a state-wide basis[9]. Therefore, this constitutes a key difference for students on the margins in Minnesota compared to those in other states[10]. In Minnesota, students who barely fail the math exit exam not only score below the passing score, but also receive district-prescribed remediation whereas students who barely pass do not. In the other states studied in past regression discontinuity studies, students who barely fail the math exit exam may or may not receive remediation, depending on the districts that they are in[11].

---

[8] The rules may differ slightly. For reading, the rule stipulates that districts provide remediation for students who fail, and that students "have a minimum of six weeks for remediation before the next testing opportunity … Minnesota Rule 3501 (2009)" (excerpt from state document provided in personal communications with MDE coordinator, November 6, 2018). It appears that there is latitude to interpret the six weeks as pertaining to remediation, or simply the duration from one test to the next. For math, students are to take "district-prescribed remediation" but the rules did not stipulate the duration.

[9] Based on the description available from the published studies. An internet search of the policies in place in Massachusetts and California at the time of the regression discontinuity studies did not yield documents of policies in place during that time period.

[10] There are also other differences, such as the location of the passing score, which may affect the profile of students on the margins. I discuss this in the discussion section.

[11] Students in both Minnesota and other states have retest options when they fail the exit exam, but the number of attempts varies across states as well.

Finally, another key difference is that the subject exit exams in Minnesota are taken in different years, each a year apart. For the other states, the exams are taken in the same year, and sometimes as part of the same test (e.g. Texas). To address this difference, I analyze the discontinuities for math and reading separately by cohorts which took the respective tests for the first time. The estimates obtained represent the effects of scoring barely below versus barely above the passing score of the respective test for all students who attempted it for the first time.

Research Questions

Within the context of Minnesota in which a pass/fail score was initially set on a high school exit exam, but the requirement to pass the exam in order to be eligible for a high school diploma was later waived for one of the subjects, I ask the following questions:

RQ1: What are the effects of scoring barely below versus barely above the passing score on the math GRAD, in the context where the passing rule for the math exam for high school diploma is waived, and students who fail are required to take remediation, on students': (a) high school, (b) college enrollment, and (c) college graduation outcomes?

RQ2: What are the effects of scoring barely below versus barely above the passing score on the reading GRAD on students': (a) high school, (b) college enrollment, and (c) college graduation outcomes?

Data

The data used in this study comes from the Minnesota Statewide Longitudinal Education Data System (SLEDS) that matches public K-12 student data with

postsecondary education data from the National Student Clearinghouse (NSC). This database allows us to track public school students longitudinally through high school and college. The database consists of test scores on the Minnesota GRAD, as well as student records from the National Student Clearinghouse (NSC), which allows us to track post-secondary college enrollment in private and public colleges across the United States. One caveat about the NSC data is that it only consists of records for high school graduates. I discuss this data limitation in the interpretation of findings.

My primary analyses for math focus on students who took the math GRAD for the first time in 11th grade in the spring of 2009 to 2011[12], corresponding to the high school graduation class of 2010 to 2012. These students make up the first three cohorts which sat for the GRAD, for which the waiver of the math exam passing requirement applies to. My reading analyses focus on students in these cohorts, but who took the reading GRAD for the first time in 10th grade in the spring of 2008 to 2010. I did not constrain the data to include students who have both reading and math scores so that the discontinuities reflect the impact on the full sample of students who took the tests for the first time in the indicated years[13].

Of the 170, 070 students who took the math exit exam for the first time in spring of 2009 to 2011, I exclude 17,686 students classified as special education students during

[12] I use a cohort naming notation based on the subject and year in which students first attempt the exam, e.g. the 2009 math GRAD cohort first took the math exit exam in spring 2009; they are the expected high school class of 2010. The 2008 reading GRAD cohort took the reading exit exam in spring 2008 and are also the expected high school class of 2010. See Appendix B Figure 1 for academic milestones by math cohorts.

[13] For the reading analyses, this allows me to include all students who attempted the reading GRAD in the spring of the indicated years, and track students who withdraw or drop out at any time after the test and who may not persist to take the math GRAD. For the math analyses, this allows me to include all students who attempted the math GRAD in the spring of the indicated years, regardless of prior actions (such as grade retention) due to their writing and reading results taken in earlier years.

the academic year in which they took the math exit exam. Students could be required to take the GRAD or an alternative assessment, depending on their Individual Education Plan (IEP). Since the SLEDS database does not have data on the required test specified in the IEP, I exclude these students from the analyses. Eventually I retain 152,317 students with non-missing values for the outcome variables for my math analyses.

Of the 178,307 students who took the reading exit exam for the first time in spring of 2008 to 2010, I exclude 16,794 students classified as special education students during the academic year in which they took the reading exit exam. Eventually I retain 161,452 students with non-missing values for the outcome variables for my reading analyses.

Outcome measures

I estimate the effect of scoring barely below versus above the GRAD (math or reading) on the probability of various high school, college enrollment and college graduation outcomes. Each of these outcomes are coded as a dichotomous variable.

The high school outcomes include on-time high school graduation expected for the cohort at the time of the test[14]. I create a variable to indicate if students ever withdraw or drop out[15, 16] from high school within 1 year of taking the math GRAD in 11th grade, which corresponds to the timeframe to the anticipated high school graduation date; or within 1 year or 2 years of taking the reading exam in 10th grade, in which the anticipated

[14] This corresponds to 2010 for the 2008 reading cohort and 2009 math cohort, up till 2012 for the 2010 reading cohort and 2011 math cohort.

[15] In the SLEDS database, "withdrew" refers to students who left school with the intention to reenroll, while "dropout" refers to students who left without intending to reenroll. However, in the dataset, I observe students with "withdrew" records who did not reenroll and students with "dropout" records who reenroll. Hence, I combine the two variables to indicate that students ever left school.

[16] Using the "withdraw or dropout" variable, I include students who ever withdraw or drop out of high school, even if they reenroll again later on. This captures treatment effects on the decision to withdraw or drop out on high school, and delineates it from more distal effects, or future decisions that may influence them to reenroll in school again.

high school graduation date is within 2 years after students take the reading exam. The estimates for the high school graduation outcome and withdraw/dropout outcome may not necessarily be identical because some students who withdraw or dropout from high school may return to high school later on, and there are some students who persist till 12th grade but never graduate from high school.

The college outcomes include enrollment in, as well as graduation from a 2-year college or 4-year college. I also look at overall college (2-year or 4-year) enrollment and graduation. Across cohorts, I track college enrollment within two years of expected high school graduation and college graduation within four years of expected high school graduation. This allows up to 4 years for students to complete 2-year colleges, and includes only graduation within 4 years at 4-year colleges (provided that they enroll in college on-time)[17]. The college data includes records for both private and public universities and colleges.

Forcing variable: GRAD scale scores

In this paper, I estimate the impact of scoring barely below versus barely above the passing score on the exit exam using a standard sharp regression discontinuity design, with the GRAD score as the forcing variable, for math and reading respectively. I center the forcing variable on the highest failing score. All students with a positive value of the re-centered forcing variable have a "pass" status on the exit exam. This means that they have satisfied the passing requirements for the high school exit exam with respect to the GRAD, and no further action is required from them in this respect.

[17] This is a rather high standard for college graduation, but it uses up to the last available year of data at the time of data request (2016) for the last cohort studied (high school class of 2012).

The estimates obtained from this strategy represent reduced form estimates of the impact of barely failing versus barely passing the exit exam. All students who score at or above the GRAD passing score will receive a "pass" status regardless of their MCA-II score. Some students with a value of zero or below on the re-centered forcing variable may have a "pass" status if they pass the MCA-II, and no further action is required on their part with respect to the GRAD. Other students who fail both the GRAD and MCA-II will have a "not pass" status. This group of students will be required to take district remediation. For math, these students will still be eligible to graduate from high school if they meet the retest requirements for their cohort, amongst other requirements. For reading, these group of students will not be eligible to graduate from high school if they cannot pass the reading GRAD on subsequent retest attempts.

Because the probability of passing the exit exam is not strictly zero for students who score below the GRAD passing score, the regression discontinuity estimates obtained using the standard sharp RD design is considered a reduced form estimate. At the end of the Analytic Strategy section, I explain further reasons why the estimates from this paper are reduced form estimates of barely failing the exit exam.

Analytic Strategy

I use a regression discontinuity (RD) strategy to identify the effects of scoring barely below versus barely above the passing score of the math and reading high school exit exam in Minnesota. The RD design relies on a few key conditions. One is an exogenously determined cut score that assigns students to one treatment condition or another. Due to the exogeneity of the cut score, it is as if students are randomly assigned to a treatment or a control condition, in our case, scoring above or below the passing

score. Another is that the design assumes that potential outcomes vary continuously around the cut score. If these conditions hold, the RD design would allow us to observe potential outcomes for both the treatment and control group, and to identify the causal effects of the treatment (fail) versus control (pass) condition.

I conduct a sharp regression discontinuity analysis that uses the GRAD scale score (for math or reading) as the forcing variable. For all of the analyses, I fit identical models using the relevant variables as below:

$$Y_i = \beta_0 + \beta_1 Score + \beta_2(BelowPass) + \beta_3(BelowPass \text{ x } Score) + \varepsilon_i$$

where *Y* is an indicator variable for the high school or college outcome of interest; *Score* is the forcing variable, the GRAD scale score centered on the highest failing score in math or in reading, for student *i*. *BelowPass* is an indicator coded with a value of 0 if the student scores at or above the GRAD passing score and 1 if below. β_2 provides the parameter estimate of interest, the causal effect of scoring barely below compared to scoring barely above the relevant exit exam, on the probability of the outcome (high school graduation, college enrollment, or college graduation). Since the forcing variable is a discrete variable, I cluster standard errors by the forcing variable as recommended by Lee and Card (2008). For math, I conduct analyses separately by cohort to examine if there are differences in impacts across cohorts, due to differences in knowledge of the waiver at the time of taking the exam, and in retesting policy over time. I also conduct analyses for reading by cohort to allow for the possibility that differences in passing requirement for math may also influence the impacts for reading.

In the RD design, we are concerned with estimating effects at the boundary for a single point, rather than maximizing fit across the whole range of data. I conduct local

linear regression which has been found to have better boundary properties (Hahn, Todd, & Van der Klaauw, 2001; Imbens & Lemieux, 2008). I fit local linear regression models using a rectangular kernel and using only observations within a narrow bandwidth (h), around the pass/fail cut score. For reporting, I use the cross-validation procedure described by Ludwig and Miller (2007) and Imbens and Lemieux (2008)[18] to identify the optimal bandwidth, separately for each outcome, subject, and cohort.

Internal validity checks

The internal validity of the RD design requires that the cut score be exogenously determined, and assignment to treatment is independent of the potential outcomes. The processes in place in Minnesota suggest that this assumption may hold. The performance standard setting process for determining the passing score was conducted independently of the scoring (or test-taking) process, by a standard setting panel consisting of educators and community members using a well-established standard setting method (item-mapping method) (Minnesota Department of Education, 2011). It seems unlikely that an individual teacher or group of teachers can influence the location of the cut score.

The cut score is set based on a raw score on the GRAD, while the cut score on the MCA-II is an equated scale score to place it on the same scale as the GRAD score scale. Having a cut score based on a raw score may raise concerns about potential for

[18] The optimal bandwidth is one that minimizes the mean squared prediction error $CV_Y(h) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$ for the δ quantile of the empirical distribution of the forcing variable. I follow the practice guide (Imbens & Lemieux, 2008) to discard observations from the tails since observations far from the cut score may introduce bias to the boundary estimate at the expense of improved precision. I discard observations more than 12 scale score points away from the cut score to retain scale scores approximately within 25th percentile and 75th percentile band of scores. I fit regression models separately for each side of the cut score for $X_i < c - h$ or $X_i > c + h$ where c is the cut-score. I implement the leave-one-out cross-validation with modification for RD as such: for a given bandwidth, I leave out observations at the boundary and use the rest of the data to predict the outcome at the boundary, progressively narrowing the bins over which the data is used to predict the boundary.

manipulation of the raw scores. However, the performance standard setting for math was conducted after the first cohort of students sat for the test, so it is not possible for students to manipulate their scores if the cut was unknown at the time[19]. Furthermore, the tests are machine-scored, so it seems unlikely that the answer sheets and scores may be manipulated. Additionally, the GRAD and MCA-II items are interspersed throughout the test. It seems unlikely that students would track which test the items belonged to and manipulate their answers to specific items, to fall on one side of the passing score or another.

I also examine the distribution of math GRAD scale scores and reading GRAD scale scores by cohort. The histograms suggest a smooth underlying distribution in the scale score. Using the Frandsen test (2017), I do not find evidence of potential manipulation of the forcing variable at the cut score.

I also plot graphs of the density of pre-treatment (demographic) covariates on the forcing variable for reading and math separately. Beyond statistical noise, the plots do not seem to suggest discontinuities in the pre-treatment covariate density close to the cut score. See Appendix C for further details.

Reduced form estimates of barely failing the exit exam

I estimate the effect of scoring barely below versus barely above the passing score of the relevant subject test, which provides reduced form estimates of the effect of barely failing versus barely passing the exit exam.

[19] The performance standard setting process was based mainly on content considerations. Although it is possible that the panel may adjust the cut score based on actual student performance on the first administration of the test, this will be possible only for the first cohort of students and not subsequent cohorts.

This is because the pass/fail status of students is determined by both the scores of the GRAD and the MCA-II. Since some students may fail the GRAD but pass the MCA-II, they may still gain a "pass" status on the exit exam (see Appendix D for further details). These students who score below the GRAD passing score will not be required to take further actions with respect to the GRAD, while other students who score below the passing score but have a "not pass" status will be required to take remediation and to meet the respective retest requirements for the math and reading exams. To the extent that scoring barely below (which increases the likelihood of failing the exam) the passing score has a stronger negative impact than any positive benefits from taking district-prescribed remediation, my discontinuity estimates may actually be conservative estimates of the impact of barely failing the exit exam.

In another way, the estimates also represent intent-to-treat effects of scoring barely below the passing score in conjunction with the other accompanying treatment (remediation and retest) versus scoring barely above the passing score. While students either score above or below the GRAD passing score, the other part of the treatment consists of remediation. The SLEDS database does not contain records of students who actually took and satisfied the district-prescribed remediation requirement. I also look at the impact at the first attempt. Hence, some students who score below the passing score on the first attempt may eventually pass the test on subsequent retest attempts and receive a "pass" status.

Descriptive Statistics

Table 1 shows the descriptive statistics and mean outcomes for the students included in the analyses by math cohorts, and pooled across reading cohorts. About 80%

of students in the overall sample are white. Students eligible for free and reduced-price lunch or eligible to receive Title 1 services (low income) during the academic year when they took the math exam constitute about a quarter of the sample; about 6% are enrolled in a school with Title 1 schoolwide program during the academic year of the math exam. About 4% of the students were assessed as limited English proficiency.

The passing rates for the math GRAD and the reading GRAD are relatively low. About 37% score below the math GRAD passing score, and have a "not pass" status on the math GRAD at the first attempt. About 22% score below the reading GRAD passing score and about 18% have a "not pass" status on the test at the first attempt.

I also show the descriptive statistics for the outcomes of interests in the overall sample. Over 80% of the students graduate from high school in the overall sample, and about 10% of students withdraw or drop out from high school after taking the GRAD. About 70% of the students in each cohort enroll in 2-year or 4-year colleges within 2 years of their anticipated high school graduation year. About 50% of the students enroll in 4-year colleges.

Results

In this section, I present the results for separate discontinuities for the math GRAD and the reading GRAD. For the math GRAD discontinuity, I estimate results by cohort (one discontinuity each for the cohorts which attempted the math GRAD for the first time in the spring of 2009 to 2011) to examine if there are differences in impact due to differences in exposure to the announcement of the waiver relative to time of taking the test, and possibly different retest requirements. Although there are no differences in terms of the rules for reading GRAD with regards to high school graduation, I also

analyze the results separately by cohort in parallel to math. Although my regression discontinuity analyses are for students at the margins of the passing score on their first attempt, in the discussion, I sometimes refer to them as the first cohort subjected to the GRAD (2008 reading cohort and 2009 math cohort) through the third cohort (2010 reading cohort and 2011 math cohort) to indicate that these are roughly the same group of students even though they take the reading and math exam in different grade-year. This helps us to track how each cohort of students are encountering the math passing policy as they take their math and reading exams.

The tables of results are organized as follows: The impacts using the math GRAD score as the forcing variable are shown in Table 2 (high school outcomes) and Table 3 (college outcomes). The impacts using the reading GRAD score as the forcing variable are shown in Table 4 (high school outcomes) and Table 5 (college outcomes). The heading for each row shows the outcome. Within each table, I present estimates obtained using the optimal bandwidth[20], followed by estimates obtained using a smaller (about half the width of the optimal) bandwidth and a larger (about two times the width of the optimal) bandwidth.

The figures are organized as follows: Figures 1 and 2 each present the graphical relationship between the probability of the relevant outcome and the forcing variable – GRAD scale scores centered on the highest failing score for math and reading respectively. For each outcome, plots are shown for the optimal bandwidth. I show the plots for the 2011 math cohort and the 2008 reading cohort[21].

[20] I determine the optimal bandwidth separately for each outcome and cohort, using the cross-validation procedure suggested by Ludwig and Miller (2007) and Imbens and Lemieux (2008).
[21] To avoid repetition, I show plots for only one cohort for each subject, for which the results are more prominent.

All the results represent estimates of the effect of scoring barely below versus barely above the passing score of the relevant subject exam, which reflect reduced form estimates for the effect of failing that subject exam. See earlier section on Analytic Strategy (reduced form estimates sub-section) for more details. For simplicity, I discuss the results in terms of negative consequences for those who score barely below the passing score. In reality, I cannot tell if this is the case, or if it is due to a positive boost for students who score barely above the passing score, or both.

Results by Subject: Math

Math: High school outcomes

Table 2 shows the estimates of scoring barely below the math passing score compared to scoring barely above, on the high school outcomes, by cohort. Broadly, the results suggest that there are no statistically significant impacts for the first math cohort (2009) to experience the waiver, but there are statistically significant impacts on high school outcomes for the subsequent (2010 and 2011) math cohorts to experience the waiver.

For the 2009 math cohort, students took the math GRAD with the understanding that passing it would be required for high school graduation, only to learn a few months later that this passing requirement would be waived. For this first cohort of students which received different signals about the passing requirement before and after taking the exam, scoring barely below the math passing score does not appear to have any statistically significant impact on the probability of graduating from high school on-time, or on the probability of ever withdrawing or dropping out from high school within 1 year of the first attempt.

For the subsequent math cohorts, we might expect impacts on high school outcomes, if any, to be less negative, since the policies appear to lower the barrier towards high school graduation. As anecdotal accounts suggest, the retesting rules appear to become more lenient over time, going from "three attempts" to "any attempt". Over time, we might also expect districts to learn from experience and to improve in their district remediation programs. Overall, we might expect less negative impacts, if any, across cohorts.

However, for the 2010 and 2011 math cohorts which took the math GRAD knowing that they do not have to pass the exam in order to graduate, I find that scoring barely below the math passing score *lowers* the probability of on-time high school graduation by about 0.7 percentage points for the 2010 math cohort and by about 1.2 percentage points for the 2011 math cohort. I also find that scoring barely below the math passing score *increases* the probability of ever withdrawing or dropping out from high school within 1 year of the first attempt, by about 0.5 percentage points (2010 math cohort) and by about 1.1 percentage points (2011 math cohort). These results appear fairly robust to the choice of bandwidth.

Figure 1 Panels A and B present the results in Table 2 for the 2011 math cohort graphically, showing the relationship between the probability of on-time high school graduation (Panel A) as well as the probability of ever withdrawing or dropping out from high school within one year of the first attempt (Panel B) respectively, and the forcing variable – math GRAD scale scores centered on the highest failing score. Although the results for these two outcomes are statistically significant, the discontinuity is barely visible.

Math: College enrollment outcomes

The results shown in Table 3 suggest that there may be some impact of scoring barely below the math passing score on 4-year college enrollment, and overall (2-year or 4-year) college enrollment across all three cohorts, but the results appear most prominent for the 2011 math cohort.

For the 2011 math cohort, scoring barely below the math GRAD passing score appears to lower the probability of 4-year college enrollment by about 4.4 percentage points, with the findings being robust to the choice of bandwidth. Due to the earlier finding that there may be a negative impact of scoring barely below the math passing score on high school graduation, it is possible that these results may be biased by missing college records for high school non-graduates rather than represent an actual treatment effect.

If scoring barely below the exam passing score has a negative impact on high school graduation, then there will be a greater proportion of high school non-graduates among those who score barely below the passing score compared to among those who score barely above. This will translate to a higher rate of missing data among those who score barely below the passing score relative to those who score barely above. This will downward bias the probability of observing any college records among those who score barely below the passing score even in the absence of any impact of scoring barely below the passing score on college enrollment or graduation.

Hence, if I observe a lower probability of college enrollment or college graduation for those who score barely below the passing score compared to those who score barely above, I will not be able to differentiate whether it is due to an actual impact

on college outcomes, or due to a bias imparted by missing college records for high school non-graduates. On the other hand, if I observe a lower probability of college enrollment or college graduation that is greater in magnitude than the impact on high school graduation for those who score barely below the passing score relative to those who score above, that would suggest that the impact is driven by an underlying mechanism other than the bias imparted by missing college records for high school non-graduates. I then assume that the underlying mechanism is the impact of scoring barely below the passing score on the college outcome in question. To take into account statistical imprecision, I compare not only the point estimates, but the 95% confidence interval of the impact on high school graduation to the college outcome of interest, within the same bandwidth for the college outcome[22].

When I compare the 95% confidence interval of the impacts for college enrollment to that of high school graduation for the 2011 math cohort estimated within 11 scale score points, I find that the 95% confidence interval of the probability of 4-year college enrollment in percentage points (-6.6, -2.3) does not overlap with that for high school graduation (-2.0, 0.3). This non-overlap is consistent across bandwidth sensitivity checks for the impact on 4-year college enrollment when compared to the impact on high school graduation. This gives some assurance that there may be an impact of scoring barely below the math GRAD passing score versus scoring barely above, on the probability of 4-year college enrollment for the 2011 math cohort, apart from any bias imparted by missing college records for high school non-graduates.

[22] For example, if I am looking at the impact on 4-year college enrollment estimated within a bandwidth of 11 scale score points, I evaluate the potential bias imparted by high school non-graduates by looking at the impact on high school graduation also within a bandwidth of 11 scale score points.

Still for the 2011 math cohort, although there appears to be a negative impact on the probability of overall (2-year or 4-year) college enrollment, the 95% confidence interval (-3.8, -1.6) overlaps with that for high school graduation (-2.1. 0.0) estimated at 10 scale score points. Thus, I cannot differentiate whether the observed difference in probability represents an actual impact on overall college enrollment, or whether it is due to bias arising from missing college records for high school non-graduates.

Turning to the other two math cohorts, I find that for the 2009 math cohort, there may be an impact of scoring barely below the math passing score on 4-year college enrollment (-3.3 percentage points) and overall (2-year or 4-year) college enrollment (-2.2 percentage points). However, the 95% confidence interval of both these college impacts overlap with the 95% confidence interval of the impact of high school graduation. Hence, I cannot discount the possibility that this may be due to a bias in missing college data for high school non-graduates. For the 2010 math cohort, I do not observe any statistically significant impacts on college enrollment.

Figure 1 Panels C to E show the above results for the 2011 math cohort graphically. A discontinuity is visible for overall (2-year or 4-year) college enrollment (Panel C) and 4-year college enrollment (Panel E).

Math: College graduation outcomes

Finally, I also look at longer-term outcomes on college graduation to examine if the negative impacts persist over time.

The impact of scoring barely below the math GRAD passing score versus scoring barely above appears to fade out for the college graduation outcomes examined. Figure 1 Panels F to H graphically show the relationship between the three college graduation

outcomes and math GRAD scale scores for the 2011 math cohort. No visible discontinuity is seen.

Due to attrition at high school and college enrollment, we might expect to see at least the magnitudes of those impacts reflected in the point estimates for college graduation. However, the magnitudes of the estimates for college graduation are generally quite small. This might suggest that there is little negative impact on college graduation for students who score barely below the passing score, but it may also imply greater attrition for those who score barely above the passing score. It is also possible that because all the college graduation outcomes are based on graduation within 4 years of the anticipated high school graduation date, the general null findings may just reflect that there is insufficient study duration to detect impacts, especially for 4-year college graduation.

Summary: Impact for math

In summary, I generally find some evidence of the negative impacts of scoring barely below versus barely above the math GRAD passing score on various high school and college enrollment outcomes. However, the negative impact findings are cohort-dependent.

It appears that there may be no impact of scoring barely below the passing score on high school outcomes for the first (2009) math cohort, but the impact grows stronger in magnitude for the 2010 and 2011 math cohorts. This is observed even though the barriers towards high school graduation imposed by the math passing requirement may be lowered for each successive cohort. We also observe the strongest evidence of impact on 4-year college enrollment for the 2011 math cohort.

Results by Subject: Reading

For math, the treatment-control contrast is between students who score barely below the math GRAD passing score – students who would still be eligible to graduate from high school provided that they take district-prescribed remediation, and sit for at least two retests or until they pass, whichever comes first; compared to students who score barely above – who on the basis of the math GRAD score are eligible to graduate from high school[23]. For reading, the treatment-control contrast is also for students who score barely below versus above the passing score. Similar to math, students who fail the reading GRAD are also required to take remediation. But unlike for math, students who fail the reading GRAD would not be eligible to graduate from high school unless they pass a retest. In other words, the barriers towards high school graduation imposed by the passing requirement for reading are higher than that for math.

The results for the impacts of reading are shown in the Table 4. In contrast to the results for math, it appears that the impacts for reading are more prominent for the first cohort of students to sit for the GRAD (2008 reading cohort), with some impact on ever withdrawing or dropping out from high school for the second (2009 reading) cohort, and no statistically significant impacts on high school outcomes for the third (2010 reading) cohort of students.

For the 2008 reading cohort, scoring barely below the reading passing score appears to lower the probability of graduating from high school on time by 1.1 percentage points, but the results are not statistically significant,. However, I observe that scoring barely below the reading passing score also increases the probability of ever

[23] This is subject to meeting all other non-math GRAD related high school graduation requirements.

withdrawing or dropping out from high school within 2 years of the first attempt by about 1.0 percentage points[24]. This latter result is statistically significant and robust to the choice of bandwidth. Figure 2 shows the graphical relationship for the 2008 reading cohort between on-time high school graduation (Panel A), as well as ever withdrawing or dropping out of high school within 2 years of the first reading attempt (Panel B), and reading GRAD scores centered on the highest failing score. The discontinuity is barely visible.

For the 2009 reading cohort, I do not find any impact of scoring barely below the reading passing score on on-time high school graduation. However, for the 2009 reading cohort, I do find that scoring barely below the reading passing score has an impact on ever withdrawing or dropping out of high school within 2 years of the first attempt, which although is not statistically significant, is similar in magnitude to the impact found for the 2008 reading cohort.

Reading: College enrollment and graduation outcomes

The estimates shown in Table 5 suggest that scoring barely below the reading passing score may have some impact on college enrollment outcomes for the 2008 and 2009 reading cohort. However, I find overlaps in the 95% confidence intervals of these impacts on college enrollment outcomes with the 95% confidence interval of impact for high school graduation. Hence, it is possible that the observed differences in probabilities may arise due to bias in missing college data for high school non-graduates.

[24] This includes students who have ever withdrawn or dropped out of high school, but may decide to return to school later. In additional analyses not shown, where I look at withdrawing or dropping out of high school without ever graduating from high school, I find that scoring barely below the reading passing score increases the probability by 0.7 percentage points (standard error of .08 percentage points, bandwidth=±3 scale score points). This lends some evidence that there may be some impact of scoring barely below the reading passing score on on-time high school graduation, but that the impact is imprecisely estimated.

Figure 2 Panels C to H show the relationships between the college enrollment and graduation outcomes respectively and the reading GRAD scores centered on the highest failing score. A slight discontinuity in impact is observed for overall (2-year or 4-year) college enrollment (Panel C) and 4-year college enrollment (Panel E).

Discussion

The impacts of passing or failing the high school exit exam on high school outcomes and college outcomes are of great interest to educators and policymakers and certainly for students on the margins of passing or failing. At a time when there are calls to set higher educational standards to prepare students to be college-ready, states may be under pressure to raise passing standards on their high school exit exams. One implication of raising standards is that a higher proportion of students may fail the exams. My study is situated in a context where the passing standards have been raised, but the passing requirement was later waived for math and where students who fail the math or reading exam have to take district-prescribed remediation. I term such a context one that is "high in standards, lenient in stakes".

Within this context, my study provides a closer look at whether there are consequences of scoring barely below versus barely above the passing score on an exit exam when passing requirements are in place for reading, and whether there are consequences for students on the margin when the passing requirements are waived for math.

The results of this study suggest several interesting patterns, and provides a more nuanced view of their impacts than previous studies suggest. Overall, the findings of the impact for the math and reading exam suggest that the impact of scoring barely below the

passing score of an exit exam may depend on the subject and the passing requirement in place. In addition, over time, the consequences for students on the margins may also change from cohort to cohort.

The students who initially attempt the reading GRAD in 2008 and the math GRAD in 2009 represent the first cohort of students to take the new high school exit exam after Minnesota switched from the Basic Skills Test. It appears that for this first cohort of students, the impacts of the reading exam, for which the passing requirement is upheld, are more pronounced than that for math, for which the passing requirement is waived. Scoring barely below the reading passing score increases the probability of ever withdrawing or dropping out from high school by the anticipated high school graduation year by 1.0 percentage points (statistically significant) and the probability of on-time high school graduation by 1.1 percentage points (not statistically significant). The corresponding impacts for math are smaller in magnitude and not statistically significant. These results may reflect the role of the passing requirement, whether it is upheld or waived, in students' decision to withdraw or drop out from high school, or it may reflect the influence of the subject, or both.

One arising puzzle is: what drives the relative influence of the reading or math exam in withdrawal or dropping out from high school? One thing to note about the 2008 reading cohort and 2009 math cohort is that they are roughly the same group of students, but some students have already dropped out after taking the reading GRAD in 10th grade and before spring of 2009 when they might have been taking the math GRAD. I find that scoring barely below versus barely above the reading GRAD passing score for the 2008 reading cohort increases the probability of withdrawing or dropping out from high school

within 1 year of the first attempt by about 0.6 percentage points (S.E. = 0.4 percentage points). Hence, some students may already have decided to withdraw or drop out from high school after they take the reading exam without waiting to take the math exam. This may also explain why the impacts on high school outcomes for the math exit exam for this cohort are more subdued, because those who might have withdrew or dropped out have already done so after they took the reading exam in grade 10.

Another related question is the timing of the impacts on decisions to withdraw or drop out from high school. The results suggest that for some students, the decision to withdraw or drop out from high school comes within one year of taking the exam while for other students, the decision arises closer to the anticipated high school graduation date. For the reading exam which students take in 10^{th} grade, we see a sizable impact (+0.6 percentage points) on the probability of ever withdrawing or dropping out from high school within one year of the initial attempt, but this impact is not statistically significant. By the second year of the initial attempt, the impact of scoring barely below the passing score grows to about +0.10 percentage points and is statistically significant. For math, even though students take the exam under conditions in which the passing requirement is in place, the impact of scoring barely below the passing score on ever withdrawing or dropping out of high school within one year of the initial attempt is practically negligible. From this timeframe, we can infer that there is little impact on this outcome within the short timeframe when students take the math exam in April 2009 and before they learn about the waiver (sometime after May 2009).

The impacts of the remedial policy are less clear. Since the remedial policy is in place for both reading and math, the differences in impacts for reading and math appear

to arise due to the differences in passing requirement for each subject. On the other hand, it may also be the case that without the remedial policy in place, the negative impacts of scoring barely below the passing score could be more severe for each of the subjects.

Observationally, the stronger influence of reading relative to math for the first cohort of students appear to undergo shifts for the second cohort of students. By the third cohort of students, the relative influence of math appear stronger than that for reading.

For the second cohort of students in the study (reading cohort 2009 and math cohort 2010), the impact of reading for students on the margins appears to be about the same or stronger (depending on bandwidth) than that for the 2008 reading cohort for high school outcomes. For math, we start to observe some negative impact of scoring barely below the math passing score on on-time high school graduation (-0.7 percentage points) and ever withdrawing or dropping out from high school prior to the anticipated high school graduation date (+0.5 percentage points).

By the third cohort of students in this study (reading cohort 2010 and math cohort 2011), it appears that the influence of the reading exam has subsided and the influence of the math exam has grown stronger. By the anticipated high school graduation date for this cohort, we do not observe any statistically significant impact of scoring barely below the reading passing score on on-time high school graduation, or students ever withdrawing or dropping out of high school. There may be a few reasons for this. It may be because over time, students come to realize that the retest opportunities allow them to eventually pass the reading exam and go on to graduate from high school[25]. It is also

[25] Unfortunately, the SLEDS database does not contain retest results for us to test the hypothesis whether students eventually receive a pass on the reading GRAD by the time of their anticipated high school graduation date.

possible that the district-remediation is working in tandem with the retesting policy to help students pass the retests.

In contrast, even though the passing requirement for the math exit exam has been waived, scoring barely below the math passing score appears to have negative impacts on high school outcomes for this third cohort of students. It does not seem that barriers towards high school graduation, such as retesting, or district-prescribed remediation are causing these impacts, since these barriers are also present for the reading exit exam.

Comparing the impact of the math to reading exam suggests that students are using the signal from the math exam rather than the reading exam to influence whether they ever withdraw or drop out of high school. This happens more so for the third cohort than for the first cohort of students in this study. Among the evidence regarding the impacts on college enrollment, we also find the strongest evidence that scoring barely below the math passing score reduces the probability of enrolling in 4-year colleges for this third cohort of students.

In Appendix E, I show the results from additional analyses where I limit the regression discontinuity analysis to include only students who pass their reading test. These analyses seek to answer the question: would we still see an impact of scoring barely below the math passing score if students' reading test results pose no barrier to high school graduation?

The results in Appendix E Table 1 suggest that among students in the 2011 math cohort who pass the reading GRAD, negative impacts are still observed for high school and college enrollment outcomes for students who score barely below the math passing score versus those who score barely above. For these students, scoring barely below the

math passing score *reduces* the probability of on-time high school graduation by about 0.7 percentage points and 5-year graduation by about 1.0 percentage points, and *increases* the probability of ever withdrawing or dropping out of high school within 1 year of the initial attempt by about 1.1 percentage points. Furthermore, for these students who pass their reading GRAD, scoring barely below the math passing score *reduces* the probability of 4-year college enrollment by about 4.7 percentage points.

Taken together, these results suggest that students in the high school class of 2011 are taking the signal from scoring below versus above the passing score of the math exam more seriously than the other cohorts.

One possibility is that the passing standard for the math exam is set very high, such that students on the margins may be relatively high-achieving students who would rather avoid a "not pass" status on their high school transcript. These students may instead choose to transfer to a non-public high school where they are not subject to the state passing requirements. In further analyses not shown, in which I constrain the reasons for withdrawing or dropping out of high school to include only transfer to a non-public high school or moving out of state, I did not find any impact. It does not seem that students are trying to find a substitute for their high school education in order to avoid a "not pass" status on their high school diploma[26]. Perhaps students who score barely below the math passing score are perceiving that the passing score is set to high standards, and the discouragement is driving them to withdraw or drop out from high school, and even not to apply to 4-year colleges.

[26] If these were relatively high-achieving students, it would seem unlikely that they withdraw or drop out of high school in order to pursue a GED in exchange for receiving a high school diploma with a "not pass" notation for their math. In additional analyses where I looked at pursuing a GED or an alternative diploma as the outcome, I also did not find any impact of scoring barely below versus above the math passing score.

Remediation provided by the district might also influence the impact on ever withdrawing or dropping out of high school and subsequently high school graduation. Based on anecdotal accounts from district staff, students who fail the math or reading exit exam are highly visible to others within school, because they have to attend remedial classes. While students who pass the math exit exam can go on to attend other electives in 12th grade, students who fail math need to attend remediation. Thus, based on the classes that students attend, students and teachers can differentiate those who pass from those who did not. Perhaps this negative visibility might contribute to demotivation and eventually result in some students withdrawing or dropping out of school. Students who have to take remediation instead of math electives may also be less competitive in their 4-year college application, resulting in a lower probability of enrollment in 4-year colleges. However, this does not explain why we might observe different patterns of the impact on withdrawing or dropping out of high school across cohorts even though students who score barely below the passing score for both subjects across all years have to attend district remediation.

Still, it is also be possible that the district remediation may be effective in helping students learn the materials on the exit exam better, so that students who persist eventually pass the retest. If this is the case for reading, it may explain why we do not see any statistically significant negative impact on high school graduation for reading despite the passing requirement being upheld.

On the whole, the finding that there are negative consequences on withdrawing or dropping out of high school is quite substantial. For students at the pass/fail margin for math, the probability of withdrawing or dropping out is about 10% while the magnitude

of the observed impact is around +0.5 percentage points to +1.1 percentage points. For reading, the probability of withdrawing or dropping out of high school is about 20% for students at the pass/fail margins while the magnitude of the impact of scoring barely below the pass score is around +1.0 percentage points.

Comparison with past studies – high school outcomes

I compare my findings to the findings from past studies to put them into perspective. Appendix F Table 1 summarizes the research design and findings of past studies that use regression discontinuity analysis to look at the impact of barely passing or barely failing an exit exam.

The results found in this study are most similar to that in New Jersey where Ou (2010) found that barely failing the math exit exam in New Jersey increases the probability of dropping out from high school by about 1.1 percentage points and 0.5 percentage points for Language Arts Literacy. The results are different from the substantive conclusion reached in Massachusetts but the point estimates are similar. Papay, Murnane, and Willett (2010) found that barely *passing* the 10th grade exit exam on the first attempt *increases* the probability of on-time high school graduation by about 1.7 percentage points for math and by 0.5 percentage points for ELA[27], but the study failed to reject the null hypothesis that there is no impact.

In terms of sample and cohort studied, my study comes the closest to that in New Jersey. Both this study on Minnesota and that on New Jersey uses data from all students in the state, with at least three high school cohorts. Although the study on Massachusetts also uses data from all students within the state, it only does so for one cohort of students

[27] The results of the study in Massachusetts are reported in terms of the positive impact from barely passing the exit exam. Here, I report the results according to this original reference point.

subject to the exit exam. From the math analyses in Minnesota, it appears that we might draw different conclusions about the impacts of scoring barely below or barely above the math passing score on high school graduation if I had used data for only one cohort and sampled a different cohort for the analyses. If we look at only findings for the 2009 math cohort and 2008 reading cohort, the results would be more similar to the findings from Massachusetts. The results in Minnesota also differs from the study on California, which uses data from multiple cohorts, but only used student data for a couple of large districts in the state. It is an open but verifiable question whether the findings across these studies would be more similar had the cohort sampling and study population (data from entire state or from only some districts) been similar.

It is also possible that the differences in results may arise because the treatment effects are different in Minnesota compared to that in other studies. However, the direction of the observed impacts do not fit if this were the case. In most of the other states studied, the implications of failing the exit exam are quite severe – high school diplomas would be withheld if students do not pass the test or subsequent retests. In Minnesota, students who do not pass the reading exam are required to pass the exit exam or subsequent retests to be eligible to graduate from high school. Those who do not pass are required to take remediation, which is supposed to help them to eventually pass the exam. Students who do not pass the math exam even have the passing requirement waived. If the waiver of the passing requirement for math and the remediation requirement for both math and reading are rules that represent lower barrier towards high school graduation or supportive measures to help students overcome the barriers, we might least expect negative consequences within such a policy context, as compared to

the policy context in other states. However, we see impacts on both high school graduation and withdrawing or dropping out of high school within Minnesota, that are more prominent than that found in states like Massachusetts or California.

Last but not least, the differences that arise in the results across studies may be due to the location of the passing thresholds. Reardon et al. (2010) put forth a hypothesis that the impact of barely failing or barely passing an exit exam may depend on the location of the passing threshold. When the passing rate on the exam is relatively high, students who fail it may be relatively low-achieving students for whom there may be other constraints towards graduation rather than the barrier imposed by the exit exam itself. In this scenario, we might expect to observe little impacts of barely failing versus barely passing the exam.

If this is the case, it may partially explain why we observe negative impacts of withdrawing or dropping out of high school for students who score barely below versus barely above the passing score of the reading test in Minnesota. The passing rate for reading in Minnesota is relatively low compared to the other states (81% in Minnesota, compared to 87% in New Jersey, 85% in Texas), indicating that the passing standard is quite high. The students who fail reading may not be particularly low-achieving, and failing the exam may be the main barrier towards high school graduation. This may lead to withdrawing or dropping out from high school if students do not eventually pass the reading exam on subsequent retests.

However, it does not quite explain the results for the math 2011 cohort, since not passing the math exam imposes a much lower barrier towards high school graduation when the passing requirement is waived. It may be possible that even when the passing

requirement is waived for math, students are still taking the signal from passing or failing the test seriously. If the passing score is perceived to be of "high standards", students may view the information regarding their mastery of the high school math curriculum as meaningful, which may influence them to act on this signal with respect to their high school and college outcomes.

Discussion on College Outcomes

Summary of past studies – college outcomes

The study that comes closest to my study is from Massachusetts (Papay, Murnane, & Willett, 2014) that looks at the impacts of barely passing or barely failing the math MCAS or the reading MCAS on college enrollment (2-year or 4-year college, public or private, from the NSC dataset), within two years of the cohort's high school graduation. The study, which looked at students who attempted the MCAS for the first time in 2004, 2005, or 2006, found that barely passing the ELA exam increases the probability that students attend college by 4.5 percentage points, while barely passing the math exam increases the probability by 2.8 percentage points.

In Texas, Martorell (2004) found that barely failing the "last-chance" test reduces the probability of attending college within 5 years of high school graduation by about 5.7 percentage points. From the results on 4-year colleges, Martorell (2004) inferred that this result was mainly driven by an impact on attendance in 2-year colleges. Note that in Texas, the estimates are for the last-chance test before students are due to graduate. Furthermore, the results are for *public* colleges within *Texas*, whereas our results are for public and private colleges across the United States.

Comparison with past studies – college outcomes

Although the direction of impacts in my study is consistent with that found in Massachusetts (Papay, Murnane, & Willett, 2014), I cannot tell whether this is due to bias imparted by missing data for high school non-graduates, or whether it is indeed due to an impact on college enrollment. On the surface, the magnitudes of the impact appear consistent for math and for the 2008 reading cohort.

In light of the findings from Massachusetts where positive impacts are found for barely passing the exit exam, and the impacts are greater for reading than for math, it is somewhat surprising that in my study, I generally find larger impacts on college enrollment for math rather than for reading. This is particularly puzzling since students who fail math in Minnesota can still be eligible for high school graduation. One possibility might be due to the remediation policy and their opportunity costs in Minnesota. It is not very clear what students may be missing out if they attend reading remediation. But for students who fail the math exit exam, math remediation may come at the expense of class time to take other math electives, especially if remediation takes place during the timetable slot for math. Students may lose out in their college application competitiveness if math electives are more valued. Unfortunately, my study does not allow me to differentiate if the lower probability of enrollment for those who score barely below versus barely above the math passing score is due to lower rates in college application, or lower rates of college acceptance.

The findings in Texas (Martorell, 2004) suggest that the negative impacts found on college enrollment was mainly due to impact on 2-year college enrollment. However, the impacts were for barely passing or failing a test that required passing each of the math, reading, and writing sections. It may be possible that the passing threshold

hypothesis put forth by Reardon et al. (2010) discussed earlier might play a part. The passing rate for the TAAS in the Texas sample is about 85%, while the passing rate for math in my sample is about 63% and that for reading is about 81%. The students at the pass/fail margin in Texas may thus be relatively low-achieving students compared to their peers in the state, whereas the students at the pass/fail margin in Minnesota may include relatively high-achieving students. If this is the case, perhaps the college of choice for students at the margins of the passing score in Minnesota are 4-year colleges rather than 2-year colleges, and hence this is where we observe greater impacts.

Conclusion

Examining the consequences of barely failing an exit exam is an important topic for policymakers and students alike. This study replicates past studies on this topic by looking at the impact of scoring barely below versus barely above the passing score of the math and reading exit exam in Minnesota. My study is different from that of other states in that the treatment for students who score below the passing score of the exit exam in Minnesota have to take remediation, and in the case of math, may still be eligible to graduate from high school due to a waiver of the passing requirement.

I find that within this policy context, there may be consequences on on-time high school graduation as well as withdrawing or dropping out of high school for students who score barely below versus barely above the passing score of the math exam, even though students can still be eligible to graduate from high school. For the reading exam in which students who fail are not eligible for high school graduation, I also find a negative impact on withdrawing or dropping out of high school.

The findings from my study are coherent with some earlier studies but not others. Other than the different passing requirements for the high school exit exam, the study design – including the cohorts studied and the student population studied – may play a role. The relative difficulty of the passing standard and the profile of students (relatively high-achieving or low-achieving) affected at the pass/fail margins may also influence the findings. Together, the results of my study relative to past studies suggest that whether there are impacts of barely passing or barely failing an exit exam may not be a settled question, and the study design as well as policy contexts may play a role. At a time when high school exit exams are still in place for a number of states, it is still a worthwhile question to ask whether there are consequences for students at the pass/fail margins.

Furthermore, at a time when there are increasing calls to set higher standards to prepare students to be college and future ready, states may move towards setting exams with higher bars for passing. But to do so may risk increasing the proportion of students who fail the exam. Setting lower standards may keep the passing rate at an "acceptable" level, but this may not signal the higher standards desired. As we saw earlier, when the passing standard is set at a level where higher-achieving students may be affected at the margins, college enrollment, particularly 4-year college enrollment, may also be impacted. If so, motivating the student population towards higher standards may potentially impose a price for those at the margins.

The particular context in Minnesota, where the passing requirement on the math exit exam was waived due to a high proportion of students failing it, provides an example of a policy response to a problem that has increasing relevance today. The findings in this

study also suggest a need to think about whether such consequences for students at the passing/failing margins need to be addressed, and if so, how.

It may also be worthwhile to ask whether students can be motivated to reach higher standards, be rewarded for their efforts when they try, but not have barriers set up should they fail. One key question that this study raises but cannot answer is, what is the impact of waiving the passing requirement and providing remediation on the consequences for students at the pass/fail margin? This study raises questions for future studies: Would administering a high school exit exam with high passing standards improve overall student performance? Or would waiving the passing requirement on the exam cause overall student performance to drop due to students working less hard towards a difficult standard that does not matter anymore? Finally, would waiving the passing requirement and providing remediation support mitigate negative consequences, if any, for students at the margins?

Tables

Table 1. Descriptive statistics for demographic covariates, high school outcomes, and college outcomes

	Math Cohort			Reading
	2009	**2010**	**2011**	**Pooled Across Cohorts**
Number of observations	51143	50971	50203	161452
Treatment characteristics				
Proportion …				
fail reading GRAD				0.19
score below reading GRAD passing score				0.22
fail math GRAD	0.37	0.37	0.36	
score below math GRAD passing score	0.38	0.37	0.36	
Demographic variables				
Male	0.49	0.49	0.49	0.49
Female	0.51	0.51	0.51	0.51
White	0.84	0.83	0.82	0.82
African-American	0.06	0.06	0.07	0.07
Hispanic	0.03	0.03	0.04	0.04
Asian or Pacific Islander	0.05	0.06	0.06	0.06
Native American/Alaskan Native	0.01	0.01	0.01	0.01
Low income	0.25	0.28	0.29	0.27
Enrolled in Title 1 school	0.06	0.06	0.06	0.06
Limited English proficiency	0.04	0.04	0.04	0.04
High School Outcomes				
Proportion …				
graduated from high school	0.88	0.88	0.88	0.84
withdrew from high school	0.08	0.08	0.09	0.11
dropped out from high school	0.03	0.02	0.02	0.03
withdrew or dropped out from high school	0.09	0.09	0.10	0.12
College Outcomes				
Proportion …				
enrolled in college (2- or 4-year)	0.72	0.72	0.72	0.68
enrolled in 2-year college	0.32	0.32	0.32	0.31
enrolled in 4-year college	0.51	0.51	0.51	0.48
graduated from college (2- or 4-year)	0.37	0.37	0.39	0.36
graduated from 2-year college	0.10	0.10	0.11	0.10
graduated from 4-year college	0.28	0.28	0.29	0.27

Table 2. Estimated impacts on high school outcomes of scoring barely below the math GRAD passing score versus scoring above at the first attempt, by math GRAD cohorts

	2009			**2010**			**2011**		
	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2
	(1)	(2)	(3)	(4)	(5)	6)	(7)	(8)	(9)
Probability of …	*h* = ±7	*h* = ±4	*h* = ±14	*h* = ±4	*h* = ±3	*h* = ±8	*h* = ±6	*h* = ±3	*h* = ±12
graduating from high school (on-time)	-0.007	0.002	-0.006	-0.007*	-0.011***	-0.005	-0.012*	-0.010**	-0.010
	(0.005)	(0.002)	(0.009)	(0.003)	(0.001)	(0.006)	(0.004)	(0.002)	(0.007)
n	23277	13984	37841	14835	10643	27515	17799	10702	33797
	h = ±7	*h* = ±4	*h* = ±14	*h* = ±5	*h* = ±4	*h* = ±10	*h* = ±4	*h* = ±3	*h* = ±8
ever withdrawing or dropping out	-0.002	0.000	-0.002	0.005†	0.007**	0.000	0.011**	0.010**	0.009*
from high school (within 1 year)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)
n	23277	13984	37841	19123	14835	32590	12939	10702	22538

†*p* < .10. **p* < .05. ***p* < .01. *** < .001.

Note: Each cell is a separate regression discontinuity impact estimated using standard sharp regression discontinuity method for the reported bandwidth. Optimal bandwidths are determined separately for each outcome and cohort, using the cross-validation procedure suggested by Ludwig and Miller (2007) and Imbens and Lemieux (2008). The smaller and larger bandwidth checks for sensitivity are approximately half or two times the size of the optimal width respectively. Standard errors shown in parentheses are clustered at discrete values of the math GRAD score.

Table 3. Estimated impacts on college outcomes of scoring barely below the math GRAD passing score versus scoring above at the first attempt, by math GRAD cohorts

	2009			**2010**			**2011**		
	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Probability of …	*h* = ±11	*h* = ±5	*h* = ±22	*h* = ±10	*h* = ±5	*h* = ±20	*h* = ±10	*h* = ±6	*h* = ±20
enrolling in 2-year or 4-year college	-0.022*	-0.011	-0.037**	-0.012	-0.010	-0.031**	-0.027***	-0.037***	-0.048***
	(0.009)	(0.007)	(0.012)	(0.007)	(0.006)	(0.010)	(0.006)	(0.005)	(0.009)
n	32949	15707	45573	32590	14835	45308	28984	17799	45093
	h = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±5	*h* = ±20	*h* = ±11	*h* = ±6	*h* = ±20
enrolling in 2-year college	0.006	0.007	0.025	-0.010	-0.006	0.010	0.005	0.009	0.006
	(0.011)	(0.007)	(0.015)	(0.013)	(0.003)	(0.015)	(0.015)	(0.012)	(0.013)
n	32949	15707	45573	33513	14835	45308	29942	17799	45093
	h = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±5	*h* = ±20	*h* = ±11	*h* = ±6	*h* = ±20
enrolling in 4-year college	-0.033*	-0.011	-0.080***	-0.013	-0.009	-0.064**	-0.044***	-0.056***	-0.088***
	(0.013)	(0.010)	(0.020)	(0.013)	(0.010)	(0.020)	(0.011)	(0.008)	(0.015)
n	32949	15707	45573	33513	14835	45308	29942	17799	45093
	h = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±5	*h* = ±20	*h* = ±11	*h* = ±6	*h* = ±20
graduating from 2-year or 4-year college	-0.010	-0.016	-0.035**	-0.009	-0.023*	-0.034*	-0.008	-0.001	-0.028*
	(0.007)	(0.011)	(0.010)	(0.015)	(0.009)	(0.013)	(0.009)	(0.009)	(0.010)
n	32949	15707	45573	33513	14835	45308	29942	17799	45093
	h = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±5	*h* = ±20	*h* = ±11	*h* = ±6	*h* = ±20
graduating from 2-year college	0.000	0.000	-0.003	-0.008	-0.008	-0.009	0.003	0.005	-0.005
	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)	(0.011)	(0.012)	(0.008)
n	32949	15707	45573	33513	14835	45308	29942	17799	45093
	h = ±5	*h* = ±3	*h* = ±10	*h* = ±6	*h* = ±5	*h* = ±12	*h* = ±8	*h* = ±4	*h* = ±16
graduating from 4-year college	-0.016†	0.002	-0.013*	-0.010	-0.020***	-0.012	-0.014	-0.013***	-0.016
	(0.007)	(0.001)	(0.006)	(0.013)	(0.003)	(0.012)	(0.008)	(0.001)	(0.009)
n	15707	11325	30014	20425	14835	36488	22538	12939	38233

†*p* < .10. **p* < .05. ***p* < .01. *** < .001.

Note: Each cell is a separate regression discontinuity impact estimated using standard sharp regression discontinuity method for the reported bandwidth. Optimal bandwidths are determined separately for each outcome and cohort, using the cross-validation procedure suggested by Ludwig and Miller (2007) and Imbens and Lemieux (2008). The smaller and larger bandwidth checks for sensitivity are approximately half or two times the size of the optimal width respectively. Standard errors shown in parentheses are clustered at discrete values of the math GRAD score.

Table 4. Estimated impacts on high school outcomes of scoring barely below the reading GRAD passing score versus scoring above at the first attempt, by reading GRAD cohorts

	2008			**2009**			**2010**		
	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Probability of …	*h* = ±3	*h* = ±4	*h* = ±6	*h* = ±4	*h* = ±3	*h* = ±8	*h* = ±6	*h* = ±3	*h* = ±11
graduating from high school (on-time)	-0.011	-0.011	-0.014	-0.002	-0.001	-0.018†	-0.011	0.001	-0.022†
	(0.010)	(0.010)	(0.011)	(0.005)	(0.002)	(0.010)	(0.014)	(0.015)	(0.011)
n	12310	13243	21430	13573	9941	22436	18630	11038	29196
	h = ±4	*h* = ±3	*h* = ±8	*h* = ±4	*h* = ±3	*h* = ±8	*h* = ±6	*h* = ±3	*h* = ±11
ever withdrawing or dropping out	0.006	0.004	0.011*	0.011	0.020**	0.012†	0.000	-0.006	0.001
from high school (within 1 year)	(0.004)	(0.006)	(0.004)	(0.006)	(0.004)	(0.006)	(0.004)	(0.005)	(0.003)
n	13243	12310	26258	13573	9941	22436	18630	11038	29196
	h = ±3	*h* = ±4	*h* = ±6	*h* = ±11	*h* = ±6	*h* = ±22	*h* = ±8	*h* = ±4	*h* = ±16
ever withdrawing or dropping out	0.010***	0.013**	0.007**	0.012	0.023*	0.013	0.006	-0.004	0.010†
from high school (within 2 years)	(0.001)	(0.003)	(0.002)	(0.009)	(0.009)	(0.010)	(0.006)	(0.002)	(0.005)
n	12310	13243	21430	32667	17524	44924	23585	11925	39913

†*p* < .10. **p* < *.05.* ***p* < *.01.* **** < *.001.*

Note: Each cell is a separate regression discontinuity impact estimated using standard sharp regression discontinuity method for the reported bandwidth. Optimal bandwidths are determined separately for each outcome and cohort, using the cross-validation procedure suggested by Ludwig and Miller (2007) and Imbens and Lemieux (2008). The smaller and larger bandwidth checks for sensitivity are approximately half or two times the size of the optimal width respectively. Where the optimal bandwidth is the smallest discrete bandwidth available, the sensitivity check is performed for the next larger possible bandwidth and reported in lighter shade. Standard errors shown in parentheses are clustered at discrete values of the reading GRAD score.

Table 5. Estimated impacts on college outcomes of scoring barely below the reading GRAD passing score versus scoring above at the first attempt, by reading GRAD cohorts

	2008			**2009**			**2010**		
	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Probability of …	*h* = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±6	*h* = ±22
enrolling in 2-year or 4-year college	-0.041***	-0.025†	-0.075***	-0.023	-0.011	-0.044**	-0.021	0.010	-0.060**
	0.010	0.012	0.014	0.013	0.014	0.013	0.016	0.014	0.018
n	32438	17148	45415	32667	13573	44924	29196	18630	49526
	h = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±6	*h* = ±22
enrolling in 2-year college	-0.017	0.000	-0.016	-0.021	-0.016	-0.011	-0.003	0.008	-0.003
	(0.012)	(0.009)	(0.011)	(0.013)	(0.018)	(0.011)	(0.008)	(0.010)	(0.008)
n	32438	17148	45415	32667	13573	44924	29196	18630	49526
	h = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±6	*h* = ±22
enrolling in 4-year college	-0.033*	-0.009	-0.085***	-0.011	-0.004	-0.060**	-0.030	0.011	-0.097**
	(0.014)	(0.012)	(0.021)	(0.013)	(0.006)	(0.019)	(0.020)	(0.015)	(0.027)
n	32438	17148	45415	32667	13573	44924	29196	18630	49526
	h = ±7	*h* = ±4	*h* = ±14	*h* = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±6	*h* = ±22
graduating from 2-year or 4-year college	-0.007	-0.007	-0.024†	-0.012	-0.006	-0.029*	-0.018*	-0.008	-0.042***
	(0.015)	(0.020)	(0.012)	(0.013)	(0.009)	(0.012)	(0.008)	(0.009)	(0.010)
n	22081	13243	38543	32667	13573	44924	29196	18630	49526
	h = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±5	*h* = ±22	*h* = ±11	*h* = ±6	*h* = ±22
graduating from 2-year college	-0.016†	0.001	-0.021*	-0.015*	-0.009	-0.017***	-0.009	-0.002	-0.018**
	(0.008)	(0.011)	(0.007)	(0.005)	(0.006)	(0.004)	(0.006)	(0.006)	(0.005)
n	32438	17148	45415	32667	13573	44924	29196	18630	49526
	h = ±5	*h* = ±4	*h* = ±11	*h* = ±4	*h* = ±3	*h* = ±8	*h* = ±4	*h* = ±3	*h* = ±8
graduating from 4-year college	0.003	0.012	-0.003	-0.002	-0.001	-0.006	0.007	0.008	-0.004
	(0.012)	(0.018)	(0.011)	(0.004)	(0.004)	(0.008)	(0.007)	(0.006)	(0.008)
n	17148	13243	32438	13573	9941	22436	11925	11038	23585

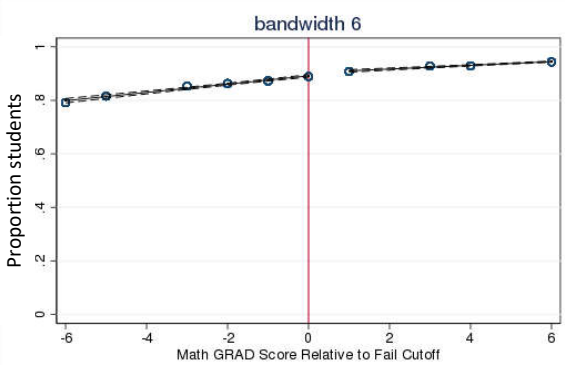
†*p* < .10. **p* < .05. ***p* < .01. *** < .001.

Note: Each cell is a separate regression discontinuity impact estimated using standard sharp regression discontinuity method for the reported bandwidth. Optimal bandwidths are determined separately for each outcome and cohort, using the cross-validation procedure suggested by Ludwig and Miller (2007) and Imbens and Lemieux (2008). The smaller and larger bandwidth checks for sensitivity are approximately half or two times the size of the optimal width respectively. Standard errors shown in parentheses are clustered at discrete values of the reading GRAD score.

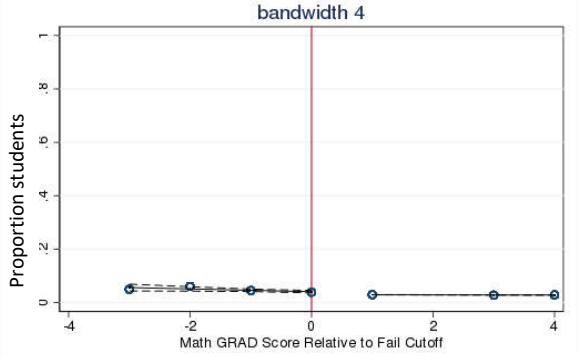
Figures

Figure 1. Proportion of students with a value of 1 on the dichotomous indicator for the respective high school, college enrollment, and college graduation outcomes on math GRAD score for 2011 math cohort

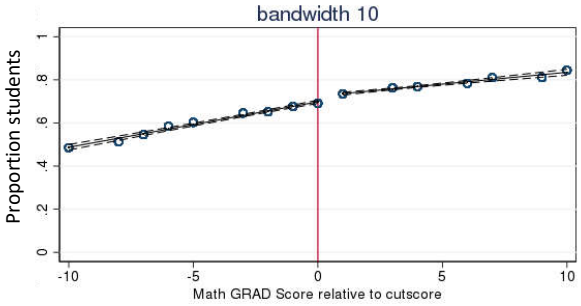
Panel A. On-time high school graduation. *Statistically significant effect*



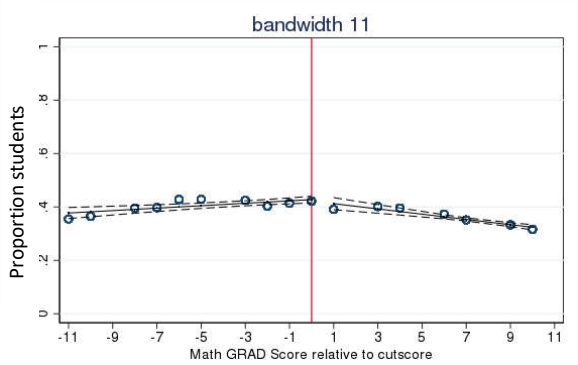
Panel B. Ever withdrew or dropped out from high school within 1 year of first attempt. *Statistically significant effect*



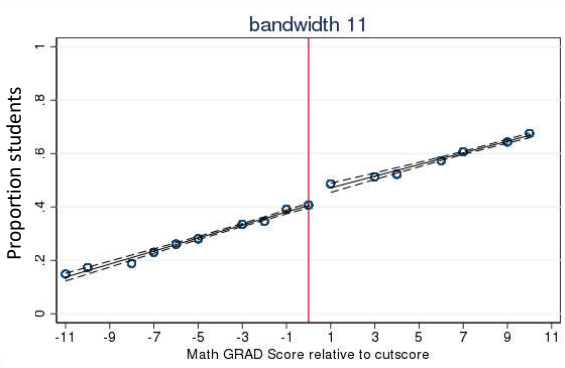
Panel C. Enrollment in 2-year or 4-year college. *Statistically significant effect*



Panel D. Enrollment in 2-year college. *No statistically significant effect*



Panel E. Enrollment in 4-year college. *Statistically significant effect*



Panel F. Graduation from 2-year or 4-year college. *No statistically significant effect*

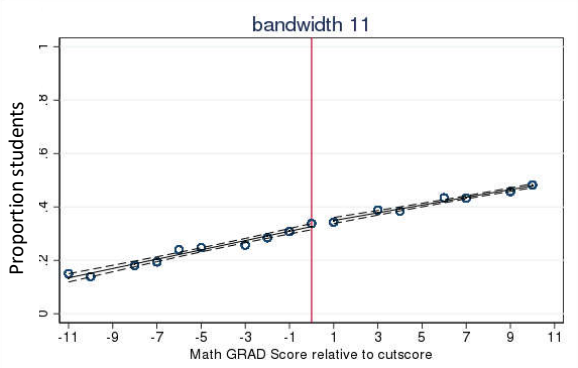
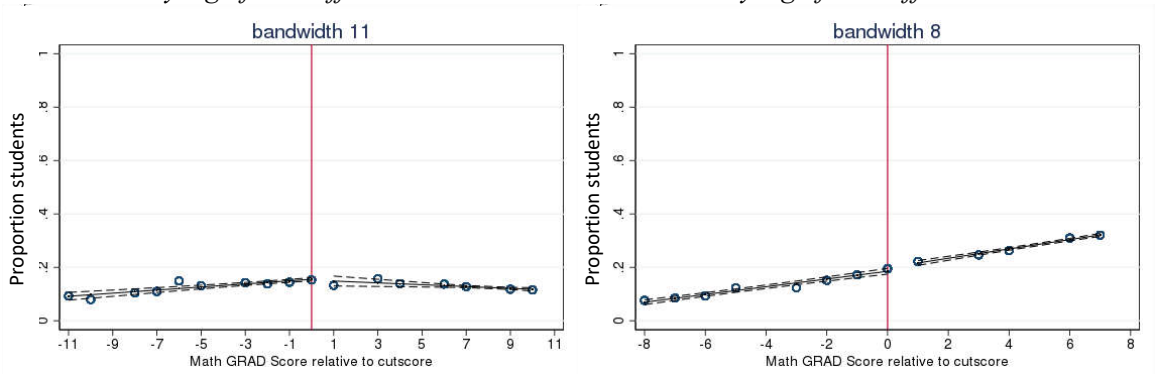


Figure 1 (continued)

Panel G. Graduation from 2-year college. *No statistically significant effect*

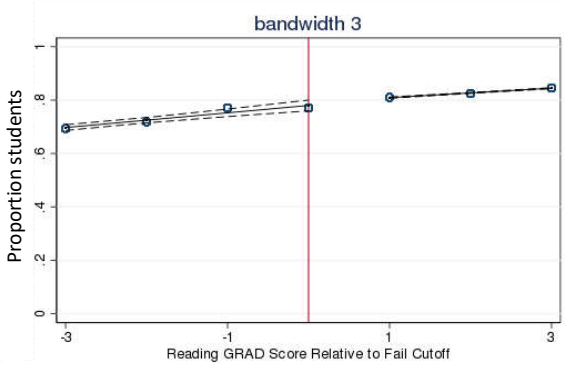
Panel H. Graduation from 4-year college. *No statistically significant effect*



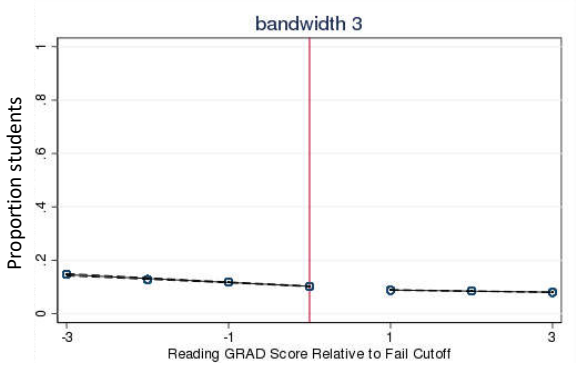
Note: Each circle represents the proportion of students (y-axis) with a value of 1 on the dichotomous indicator as described in the panel title, for each discrete scale score on the math GRAD (x-axis), i.e. the smallest bin possible. The vertical line represents the highest failing score on the math GRAD. Local linear fitted lines (solid line) are estimated using data for all students using a rectangular kernel and the displayed bandwidth. The displayed bandwidth is the optimal bandwidth determined using the cross-validation procedure described in text. Dotted lines represent the 95% confidence interval.

Figure 2. Proportion of students with a value of 1 on the dichotomous indicator for the respective high school, college enrollment, and college graduation outcomes on reading GRAD score for 2008 reading GRAD cohort

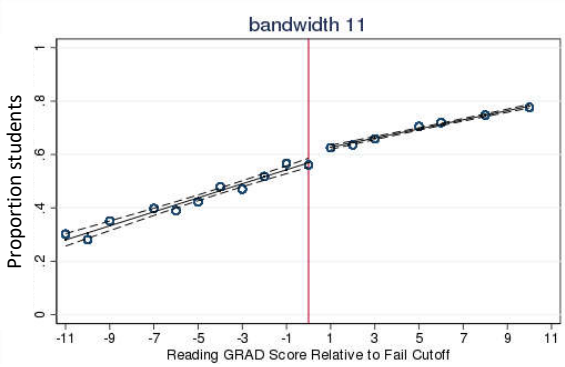
Panel A. On-time high school graduation. *No statistically significant effect*



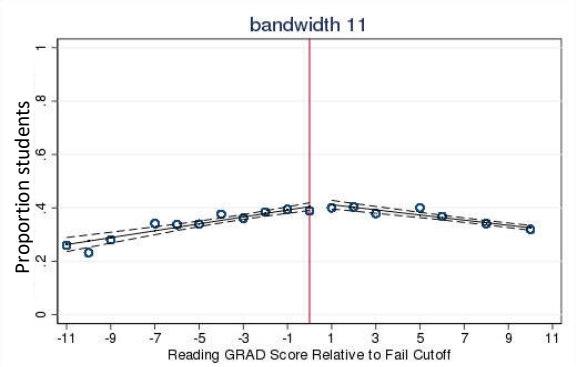
Panel B. Ever withdrew or dropped out from high school within 2 years of first attempt. *Statistically significant effect*



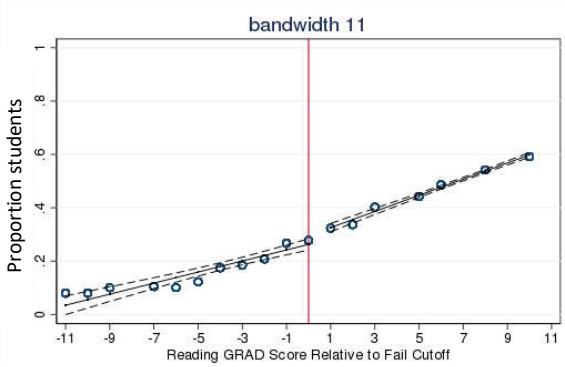
Panel C. Enrollment in 2-year or 4-year college. *Statistically significant effect*



Panel D. Enrollment in 2-year college. *No statistically significant effect*



Panel E. Enrollment in 4-year college. *Statistically significant effect*



Panel F. Graduation from 2-year or 4-year college. *No statistically significant effect*

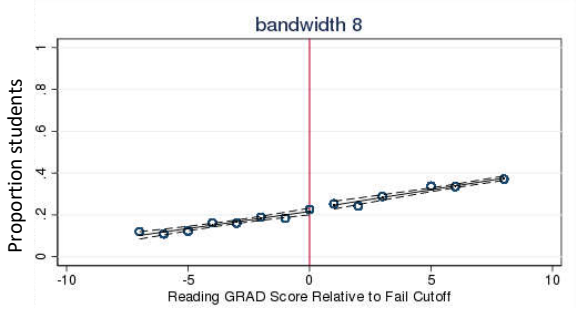
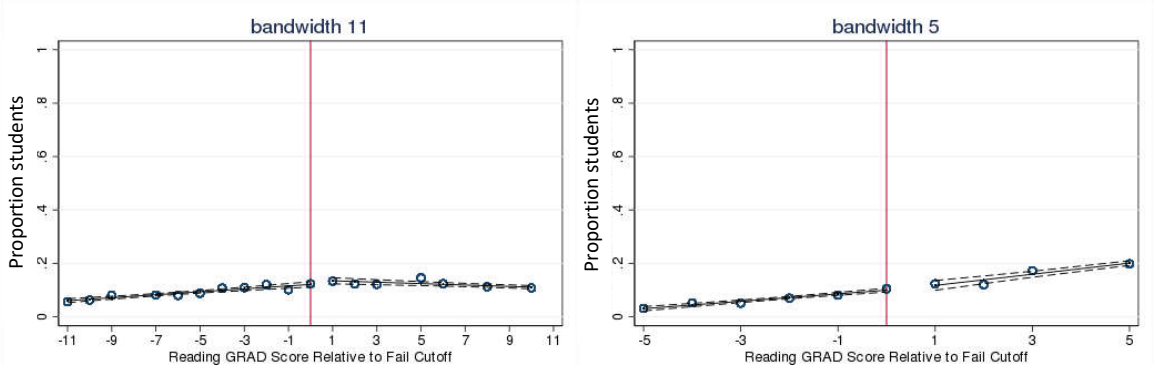


Figure 2 (continued)

Panel G. Graduation from 2-year college. *No statistically significant effect*

Panel H. Graduation from 4-year college. *No statistically significant effect*



Note: Each circle represents the proportion of students (y-axis) with a value of 1 on the dichotomous indicator as described in the panel title, for each discrete scale score on the reading GRAD (x-axis), i.e. the smallest bin possible. The vertical line represents the highest failing score on the reading GRAD. Local linear fitted lines (solid line) are estimated using data for all students using a rectangular kernel and the displayed bandwidth. The displayed bandwidth is the optimal bandwidth determined using the cross-validation procedure described in text. Dotted lines represent the 95% confidence interval.

Appendices

Appendix A
Role of GRAD for Graduation from Public High Schools in Minnesota

High school class of 2010 to 2014

To graduate from a Minnesota public high school, students must meet the state's course credit and testing requirements and any additional local requirements established by the school district (Larson, 2010).

The role of Graduation-Required Assessment for Diploma (GRAD) for graduation from a public high school in Minnesota is as follows:

> Minnesota Statutes 2006 120B.30 Statewide Testing and Reporting System…
> (b) For students enrolled in grade 8 in the 2005-2006 school year and later, only the following options shall fulfill students' state graduation test requirements:
> (1) for reading and mathematics:
> (i) obtaining an achievement level equivalent to or greater than proficient as determined through a standard setting process on the Minnesota comprehensive assessments in grade 10 for reading and grade 11 for mathematics or achieving a passing score as determined through a standard setting process on the graduation-required assessment for diploma in grade 10 for reading and grade 11 for mathematics or subsequent retests;
> …
> (2) for writing:
> (i) achieving a passing score on the graduation-required assessment for diploma;
> (Minnesota Statutes 2006, 2006, 120B.30)

Based on conversations with Minnesota Department of Education officials, after students set for the GRAD in spring 2009, the state announced that the 2009 Legislature waived the passing requirement on math GRAD. The statute was amended to include the following:

> (d) Students enrolled in grade 8 in any school year from the 2005-2006 school year to the 2009-2010 school year who do not pass the mathematics graduation-required assessment for diploma under paragraph (b) are eligible to receive a high school diploma with a passing state notation if they:
> (1) complete with a passing score or grade all state and local coursework and credits required for graduation by the school board granting the students their diploma;
> (2) participate in district-prescribed academic remediation in mathematics; and

> (3) fully participate in at least two retests of the mathematics GRAD test or until they pass the mathematics GRAD test, whichever comes first. A school, district, or charter school must place a student's highest assessment score for each of the following assessments on the student's high school transcript:
> - the mathematics Minnesota Comprehensive Assessment,
> - reading Minnesota Comprehensive Assessment, and
> - writing Graduation-Required Assessment for Diploma,
> and when applicable,
> - the mathematics Graduation-Required Assessment for Diploma and
> - reading Graduation-Required Assessment for Diploma.
> (Minnesota Statutes 2009, 2009, 120B.30)

Some student groups were exempted from the above requirements. Larson (2010) summarized:

> Students with limited English proficiency who first enroll in a Minnesota public school in grade 9 or above need not pass the GRAD tests to graduate. … Students with IEPs and significant cognitive disabilities can take the Minnesota Test of Academic Skills (MTAS) instead of the GRAD reading and math tests.

The MCA and GRAD in Minnesota

According to the Research Department of the Minnesota House of Representatives (Larson, 2010):

> The GRAD reading and math test items that students must pass to graduate in Minnesota are embedded in the reading and math Minnesota Comprehensive Assessments-Series II (MCA-II). Students' GRAD test scores and MCA-II test scores are reported separately. The state and districts use students' GRAD test scores to determine whether students graduate.

Based on conversations with Minnesota Department of Education officials, and what is observed in the data, in cases where students did not pass the GRAD but scored above the MCA-II passing score, students are also considered to have passed the GRAD. In cases where students passed the GRAD but did not score above the MCA-II passing score, students are still considered to have passed the GRAD.

Example of transcript notation when the pass waiver is in place

Here is an example of a student's transcript notation as she progressed in a typical fashion through her high school. These notations would be used no matter the assessment used (MCA, GRAD, MTELL, MTAS, Writing Alternate Assessment).

Example of the notation in the spring of grade 12. Jane passed the reading GRAD in the summer after grade 11 and writing GRAD in grade 9, but has not yet passed the math GRAD despite several retest attempts. Under the current state statute, Jane may still graduate if she fulfilled the mathematics requirements for the alternate pathway to graduation. However, her transcript will still reflect a "Not pass" status for mathematics.

Jane Doe – Spring of Grade 12	
Grad assessment status mathematics	Not pass
Grad assessment status reading	Pass
Grad assessment status writing	Pass

Source: Excerpt from state document provided in personal communications with MDE coordinator, November 6, 2018.

References

Revisor of Statutes, State of Minnesota (2007). 2007 Minnesota Session Laws Chapter 146 Versions. Retrieved from https://www.revisor.mn.gov/laws?id=146&doctype=Chapter&year=2007&type=0

Revisor of Statutes, State of Minnesota (2009). Minnesota Statues 2009. Retrieved from https://www.revisor.mn.gov/statutes/?id=120b.30

Larson, L. (2010). Minnesota's High School Graduation Requirements (2010). House Research Department. Retrieved from http://www.house.leg.state.mn.us/hrd/pubs/ss/sshsgrad.pdf

Appendix B
Timeline of Policy Announcement and Implementation by Cohorts

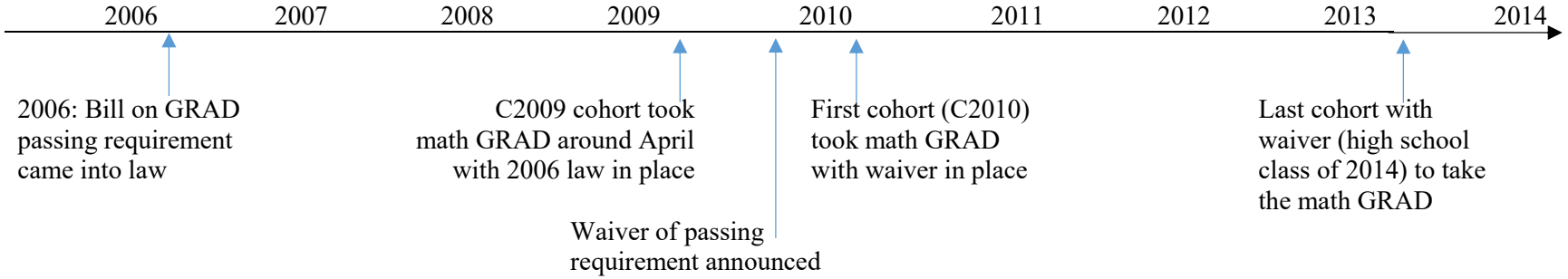
Table B1. Grade (high school and post-high school) and academic milestone by math GRAD cohort (year in which students took math GRAD for the first time) and academic year

Grade[1]	**Milestone**	**Academic Year ending in Spring**									
		2007	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
G9	Writing GRAD	C2009	C2010	C2011							
G10	Reading GRAD		C2009	C2010	C2011						
G11	Math GRAD			C2009	C2010	C2011					
G12	High school graduation				C2009	C2010	C2011				
PHS1	College enrollment					C2009	C2010	C2011			
PHS2	College graduation (within 2 years)						C2009	C2010	C2011		
PHS3								C2009	C2010	C2011	
PHS4	College graduation (within 4 years)								C2009	C2010	C2011

Waiver of passing requirement for math GRAD announced

[1] Grade: G refers to high school grade; PHS refers to post-high school year.

Figure B1. Timeline of Policy Announcement and Implementation



Appendix C
Regression Discontinuity Internal Validity Check

Identification in the regression discontinuity design requires that the cut score is exogenously determined. One potential violation of this assumption is if there is potential manipulation in the forcing variable, resulting in discontinuity in the density of the forcing variable around the cut score. Formal tests for this potential violation include the one more commonly used and developed by McCrary (2008) for continuous forcing variables, and one more recently developed by Frandsen (2017) for discrete forcing variables.

Frandsen (2017) suggests that the McCrary test works well when the forcing variable is continuous, but is inconsistent when the forcing variable is discrete. The McCrary test relies on the number of observed support points near the cutoff growing to infinity as the sample size increases, which is the case for a continuous forcing variable but not a discrete one. The McCrary estimator for testing continuity of the forcing variable is robust when the bandwidth to binsize ratio $h/b > 10$ (McCrary, 2008). The Frandsen test uses the fact that if the discrete forcing variable is based on an underlying continuous variable with a continuous density, then the observed frequency at the threshold has a known approximate conditional distribution. The test then uses only support points immediately adjacent to the cut score. Frandsen shows that if the discrete forcing variable has an underlying continuous distribution, then conditional on the forcing variable taking on a value at the cut score or at the immediate adjacent support point, the probability of being exactly at the cut score is approximately 1/3. This forms the null hypothesis for no manipulation for the Frandsen test.

I examine the histograms of the forcing variable separately for math and reading, by cohort, and conduct the Frandsen test (2017) for each subject-cohort. The histograms suggest that the forcing variable is an approximately smoothly increasing function around the cutoff. Using the Frandsen test, I fail to reject the null hypothesis that there is no manipulation. Furthermore, I also plot graphs of the density of pre-treatment (demographic) covariates on the forcing variable. Beyond statistical noise, the plots do not seem to suggest discontinuities in the pre-treatment covariate density close to the cut score. The analyses are available upon request.

Appendix D
Use of GRAD and MCA-II Score to Determine Pass/Not Pass Status

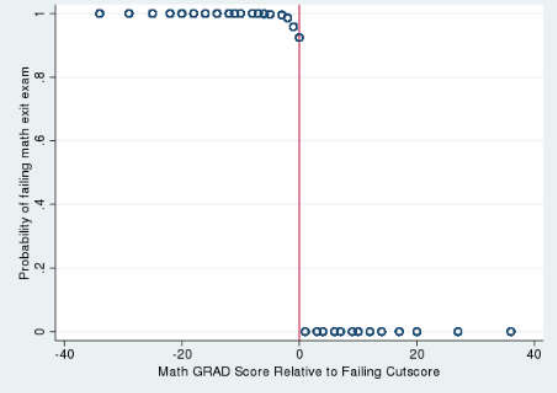
The "pass" / "not pass" status on the exit exams in Minnesota is determined by scores on two related tests, the GRAD and MCA-II. Students sit for a single administration of the math exam during spring of their 11th grade. The component originally designed as the high school graduation requirement, GRAD, which students have to pass, is a set of items that partially overlaps with items for the MCA-II component. The MCA-II is used by the state and districts for accountability purposes. The GRAD items are interspersed with the MCA-II items so that students do not know which test(s) the items apply to.

However, about 0.60% of the students failed the GRAD component but passed the MCA-II component. These students also received a "pass" status on their high school exit exam, which gives rise to the non-zero passing probabilities below the passing score (see Figure 1 Panel A of this appendix) because the MCA-II passing score is deemed more rigorous than that for the GRAD component. For reading, about 3.75% of the students failed the GRAD component but passed the MCA-II component, giving rise to the non-zero failing probability below the passing score (Figure 1 Panel B).

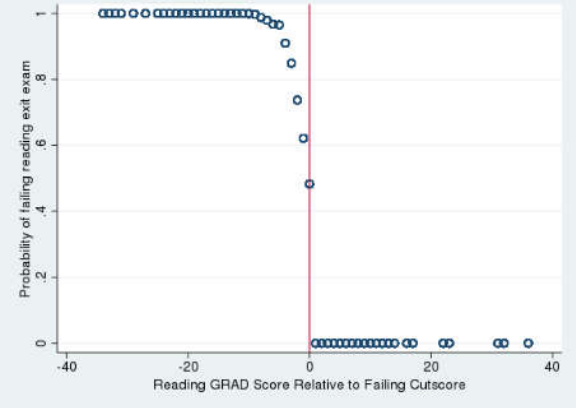
I conduct standard sharp regression discontinuity analyses by using students who score barely below the passing score as the treatment group versus those barely above the passing score as the control group. This definition of treatment provides reduced form estimates of barely failing versus barely passing the exit exam.

Figure D1. Probability of receiving a "not pass" status on GRAD scale score for math (Panel A) and reading (Panel B) at first attempt, pooled across math cohorts and reading cohorts respectively

Panel A: Math



Panel B: Reading



Appendix E
Additional Regression Discontinuity Analyses for Students who Passed Reading GRAD

Table E1. Estimated impacts on selected high school and college enrollment outcomes of scoring barely below the math GRAD passing score versus scoring above at the first attempt, among those who passed the reading GRAD at the first attempt, for math GRAD cohort 2011

	2011		
	Optimal *h*	(Optimal *h*)/2	(Optimal *h*)x2
	(1)	(2)	(3)
Probability of …	*h* = ±6	*h* = ±3	*h* = ±12
graduating from high school (on-time)	-0.007*	-0.007***	-0.010*
	(0.002)	(0.001)	(0.004)
n	15027	9028	28501
	h = ±7	*h* = ±4	*h* = ±14
graduating from high school (within 5 years)	-0.010***	-0.007***	-0.013***
	(0.002)	(0.001)	(0.003)
n	18203	11067	31740
	h = ±4	*h* = ±3	*h* = ±8
ever withdrawing or dropping out	0.011**	0.010**	0.010**
from high school (within 1 year)	(0.003)	(0.003)	(0.003)
n	11067	9028	18831
	h = ±10	*h* = ±5	*h* = ±22
enrolling in 2-year or 4-year college	-0.025**	-0.033***	-0.047***
	(0.007)	(0.006)	(0.008)
n	24675	12058	38121
	h = ±10	*h* = ±5	*h* = ±22
enrolling in 2-year college	0.006	0.029*	0.002
	(0.016)	(0.009)	(0.012)
n	24675	12058	38121
	h = ±11	*h* = ±5	*h* = ±22
enrolling in 4-year college	-0.047***	-0.067***	-0.080***
	(0.011)	(0.006)	(0.014)
n	28501	12058	38121

†*p* < .10. **p* < .05. ***p* < .01. **** < .001.*

Note: Each cell is a separate regression discontinuity impact estimated using standard sharp regression discontinuity method for the reported bandwidth. Optimal bandwidths are determined separately for each outcome, using the cross-validation procedure suggested by Ludwig and Miller (2007) and Imbens and Lemieux (2008). The smaller and larger bandwidth checks for sensitivity are approximately half or two times the size of the optimal width respectively. Standard errors shown in parentheses are clustered at discrete values of the math GRAD score.

Appendix F
Summary of Past Studies

Table F1. Summary of study design, context, and findings in Minnesota compared from past regression discontinuity studies conducted in Massachusetts, New Jersey, California, and Texas on the effect of barely failing versus barely passing exit exams

State	Study Design	Context	Math Findings	Reading Findings
Minnesota	• By cohort analyses across 3 cohorts • Analyses conducted for all students in state (n_{math} = 152,317) ($n_{reading}$ = 161,452)	• First reading attempt in 10th grade in 2008 to 2010; • First math attempt in 11th grade in 2009 to 2011 • Passing Rate (math): 63% • Passing Rate (reading): 81%	• On-time high school graduation by math cohort: 2009: -0.007 (0.005) 2010: -0.007* (0.003) 2011: -0.012* (0.004) • Ever withdraw or dropout from high school within 1 year of first attempt: 2009: -0.002 (0.002) 2010: +0.005† (0.002) 2011: + 0.011** (0.002)	• On-time high school graduation by reading cohort: 2008: -0.011 (0.010) 2009: -0.002 (0.005) 2010: -0.011 (0.014) • Ever withdraw or dropout from high school within 2 years of first attempt: 2008: +0.010*** (0.001) 2009: + 0.012 (0.009) 2010: + 0.006 (0.006)
New Jersey (Ou, 2010)	• Pooled and by-cohort analyses across 4 cohorts • Analyses conducted for all students in state (n = 299,948)	• First attempt in 11th grade in spring 2002 to 2005 • Passing rate (math): 76% • Passing rate (LAL): 87%	High school dropout: +0.011*** (0.001)	High school dropout: +0.005*** (0.002)
Massachusetts (Papay, Murnane, & Willett, 2010)	• 1 cohort • Analyses conducted for all students in state (n = 66,347)	• First attempt in 10th grade in 2004 • Passing Rate (math): 87%	On-time high school graduation[a]: -0.017 (0.010)	On-time high school graduation[a]: -0.005 (0.017)
California (Reardon, Arshan, Atteberry, & Kurlaender, 2010	• Pooled analyses across 5 cohorts • Analyses conducted for students in 4 of 10 largest districts in CA (n = 106,454)	• First attempt in 10th grade, for cohorts scheduled to graduate in 2006 to 2010 • Passing rate (math): 78% • Passing rate (ELA): 79%	Effect of failing at least one section (math or ELA) on on-time graduation: +0.027* (0.011); The authors note that this result is sensitive to bandwidth, and mostly statistically non-significant at other bandwidth checks	
Texas (Martorell, 2004)	• Pooled analyses across 3 cohorts • Analyses conducted for all students in state (n = 505,291)	• First attempt in 10th grade in 1993 to 1995 • Passing rate (all 3 sections): 85%	Effect of failing exam at first attempt[b]: -0.001 (0.004)	

Notes:
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***} < .001$.

All impacts (point estimates and standard errors in brackets) expressed in this table are in terms of the impact of barely failing versus barely passing the exit exam, except in my study where I look at the impact of scoring barely below versus barely above the passing score.

[a] The original study compared the effects of barely passing versus barely failing the exit exam. In this table I flipped the direction of comparison.

[b] In Texas, the exit exam was structured differently in that students were required to take an exit exam consisting of math, reading, and writing, and were required to pass all three sections.

References

Ahn, T. (2014). A regression discontinuity analysis of graduation standards and their impact on students' academic trajectories. *Economics of Education Review*, *38*, 64-75.

Amrein, A. L., & Berliner, D. C. (2003). The Effects of High-Stakes Testing on Student Motivation and Learning. *Educational Leadership, 60*(5), 32–38.

Autor, D. H. (2014). Skills, education, and the rise of earnings inequality among the "other 99 percent". *Science*, *344*(6186), 843-851.

Bailey, M. J., & Dynarski, S.M. (2011). Inequality in postsecondary education. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity?* (pp. 117-131). New York, NY: Russell Sage Foundation.

Center on Education Policy (2010). *MN State Profile.* Retrieved from: https://eric.ed.gov/?id=ED530202

Center on Education Policy (2011). *State high school tests: Changes in state policies and the impact of the college and career readiness movement.* Washington, DC: Center on Education Policy. Retrieved from http://files.eric.ed.gov/fulltext/ED530163.pdf

Clark, D., & See, E. (2011). The impact of tougher education standards: Evidence from Florida. *Economics of Education Review*, *30*(6), 1123-1135.

Dougherty, S. (2012, November). *Bridging the discontinuity in adolescent literacy: Evidence of an effective middle grades intervention*. Paper presented at the APPAM Fall Research Conference.

Dee, T. S., & Jacob, B. A. (2008). Do high school exit exams influence educational attainment or labor market performance? In A. Gamoran (Ed.), *Standards-based reform and the poverty gap: Lessons for No Child Left Behind* (pp. 154-197). Washington, DC: Brookings Institution Press.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579-621). Westport, CT: American Council on Education/Praeger Publishers.

Frandsen, B. R. (2017). Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. In R. C. Hill, T. B. Fomby, J. C. Escanciano, E. Hillebrand, & I. Jeliazkov (Eds.), *Regression discontinuity designs: Theory and applications* (pp. 281-315). Bingley, UK: Emerald Publishing Limited.

Greene, J. P., & Forster, G. (2003). P*ublic high school graduation and college readiness rates in the United States*. Education Working Paper No. 3. New York, NY: Manhattan Institute.

Greene, J. P., & Winters, M. A. (2004). Pushed out or pulled up? Exit exams and dropout rates in public high schools. New York, NY: Manhattan Institute.

Grodsky, E., Warren, J. R., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971-2004. *Educational Policy*, *23*(4), 589-614.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201-209.

Holme, J. J., Richards, M. P., Jimerson, J. B., & Cohen, R. W. (2010). Assessing the Effects of High School Exit Examinations. *Review of Educational Research, 80*(4), 476–526.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615-635.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, *23*(2), 99-121.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, *86*(1), 226-244.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger Publishers.

Revisor of Statutes, State of Minnesota (2005). 2005 Minnesota Session Laws Chapter 120B.02 Versions. Retrieved from https://www.revisor.mn.gov/statutes/2005/cite/120B.02

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655-674.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, *122*(1), 159-208.

Martorell, F. (2004). *Do high school graduation exams matter? A regression discontinuity approach.* Job market paper, University of California, Berkeley. Retrieved from https://www.utdallas.edu/research/tsp-erc/pdf/wp_martorell_2004_high_school_graduation_exams.pdf

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698-714.

Minnesota Department of Education (2010a). *Minnesota Graduation-Required Assessments for Diploma (GRAD) test specifications for mathematics.* Retrieved from https://education.mn.gov/mdeprod/idcplg?IdcService=GET_FILE&dDocName=003182&RevisionSelectionMethod=latestReleased&Rendition=primary

Minnesota Department of Education (2010b). *Minnesota Comprehensive Assessments Series II (MCA-II) test specifications for grade 11 mathematics.* Retrieved from https://education.mn.gov/mdeprod/idcplg?IdcService=GET_FILE&dDocName=003244&RevisionSelectionMethod=latestReleased&Rendition=primary

Minnesota Department of Education (2011). *Standard setting technical report for Minnesota assessments*. Prepared by Pearson. Retrieved from https://education.mn.gov/mdeprod/idcplg?IdcService=GET_FILE&dDocName=042704&RevisionSelectionMethod=latestReleased&Rendition=primary

Minnesota Session Laws – Regular Session, 2007. Chapter 196, H.F. No. 2245 Retrieved from: https://www.revisor.mn.gov/laws/2007/0/146/%5E(%3FPlaws.2.9.0%5B0-9%5C.a-zA-Z%5Cs/%5C/%5D+)$#laws.2.9.0

Minnesota Session Laws – Regular Session, 2009. Chapter 96, H.F. No. 2. Retrieved from: https://www.revisor.mn.gov/laws/2009/0/96/

Minnesota Statutes, 2005. Chapter 120B.30 (Statewide Testing and Reporting System). Retrieved from: https://www.revisor.mn.gov/statutes/2005/part/EDUCATION%2520CODE%253A%2520PREKINDERGARTEN%2520-%2520GRADE%252012

Minnesota Statutes, 2009. Chapter 120B.30 (Statewide Testing and Reporting System). Retrieved from: https://www.revisor.mn.gov/statutes/2009/cite/120B.30

National Research Council. (1998). *High stakes: Testing for tracking, promotion, and graduation*. National Academies Press.

Ou, D. (2010). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review*, *29*(2), 171-186.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis*, *32*(1), 5-23.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2014). High-school exit examinations and the schooling decisions of teenagers: Evidence from regression-discontinuity approaches. *Journal of Research on Educational Effectiveness*, *7*(1), 1-27.

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis*, *32*(4), 498-520.

Reardon, S. F., & Kurlaender, M. (2009). Effects of the California high school exit exam on student persistence, achievement, and graduation. Policy Brief 09-3. *Policy Analysis for California Education, PACE (NJ1)*.

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, *5*(1), 83-104.

Shuster, K. (2012). Re-Examining exit exams: New findings from the Education Longitudinal Study of 2002. *Education Policy Analysis Archives, 20*(3). Retrieved from http://epaa.asu.edu/ojs/article/view/797

Taylor, E. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, *117*, 162-181.

Warren, J. R., Grodsky, E., & Lee, J. C. (2008). State high school exit examinations and postsecondary labor market outcomes. *Sociology of Education, 81*(1), 77–107.

Essay 3

District-SES Test Score Gaps Before and After an Assessment Change in Texas

District-SES Test Score Gaps Before and After an Assessment Change in Texas

Introduction

Socioeconomic gaps in achievement are one of the most persistent features in education. To promote greater educational equity, policymakers have sought to raise achievement for all students. One of the most widespread means is through high stakes accountability policies, such as the No Child Left Behind Act of 2001 (NCLB), which requires districts to report disaggregated scores by racial minority and economically disadvantaged groups (NCLB, 2002). The spotlight on these students have put intense pressure on districts serving predominantly disadvantaged students. Districts that do not meet the improvement targets may face sanctions. This pressure may lead districts to make meaningful changes that impact student learning (Rouse, Hannaway, Goldhaber, & Figlio, 2013; Winters, Trivitt, & Greene, 2010), which is the desired behavior targeted by accountability policies. It may also influence districts to use non-learning related strategies to increase student test scores (Haney, 2000; Jacob, 2005; Jennings & Sohn, 2014; Ladd & Lauen, 2010; Neal & Schanzenbach, 2010; Nichols & Berliner, 2007), which distorts behavior in the very districts that policymakers hope to improve.

Accountability policies typically require student performance to be measured based on a state-developed academic assessment. NCLB requires all states to test students in each grade from grades 3 to 8 in math and reading or language arts to measure progress towards targets. Even as the spotlight is on student performance on state tests, and rewards and sanctions for districts are instituted based on the results on these high-stakes tests, it is important to remember that the high-stakes state tests are but a

measurement tool to measure student learning in broader domains that educators are interested in, such as math, or reading.

For example, consider this statement that the Texas Education Agency (2008), from the state where NCLB accountability policies are largely modeled after, made regarding the state test, Texas Assessment of Knowledge and Skills (TAKS): "TAKS has been developed to better reflect good instructional practice and more accurately measure student learning. We hope that every teacher will see the connection between *what we test on this state assessment* and *what our students should know and be able to do to be academically successful* [emphasis added]" (Texas Education Agency, 2008, p.1). Although accountability policies call for many high-stakes decisions to be made based on performance on a single test, it would be safe to say that most policymakers and educators would want those decisions to be robust to the measurement tool used.

However, it is well documented that gains on high-stakes state tests for low-income children and racial minority children are not matched on audit tests at the state level such as the NAEP (Ho, 2007; Ho, 2009; Ho & Haertel, 2006), or other state specific low-stakes tests at the district level (Jacob, 2005), or researcher-administered audit tests (Klein, Hamilton, McCaffrey, & Stecher, 2000). This gives rise to the concern about whether the test score gains are due to improved learning, or due to non-learning related reasons.

What we cannot tell from the above studies is whether the lack of generalizability of test score gains across different tests happens for students across all districts, or is concentrated within specific districts. Much work has been done to study test score gaps at the student level, but it is also useful to monitor gaps at the district level. Districts are

the educational unit that have the most direct control over educational, human, financial and infrastructure resources that matter for students' education. Districts are also the unit that is held accountable under state accountability systems. It is thus useful to understand how assessments affect test scores and score gaps at the district level.

In this study, I ask whether observed district-SES gaps change when the state assessment changes. The gold standard for examining whether measured district-SES gaps, or for that matter any other performance or gap that educators care about, differ due to a different measurement tool being used, i.e. the state test, is to conduct an experiment. One such experiment might randomly assign high-SES districts and low-SES districts to either the existing high-stakes state test or another test with high-stakes placed upon it, both based on the same curriculum, both designed and developed in a way to be representative of the tested domain, but are otherwise not predictably similar to each other. We would measure gaps in performance between the high-SES and low-SES districts using each test, and compare whether the measured gaps are statistically similar. If the measured gaps differ between the two tests, then there might be cause for concern that inferences about district-SES gaps based on the high-stakes state test might not be generalizable across the measurement tool used.

In the absence of such an experiment, this study turns to a natural experiment to answer the question. I make use of a change in Texas' high-stakes state assessment to serve as an alternative measure of district-SES gaps in test scores. I ask whether district-SES gaps, as measured by the old high-stakes test, differ when a new high-stakes test is introduced.

As previously mentioned, it is well known that scores on large-scale achievement tests drop when a new testing program is implemented (Koretz & Hamilton, 2006). Those studies are based on student-level test scores, nation-wide (e.g. Ho & Haertel, 2006) or state-wide using NAEP (e.g. Ho & Haertel, 2006; Koretz & Barron, 1998), within a district (e.g. Jacob, 2005; Koretz, Linn, Dunbar, & Shepard, 1991), or using purposive samples (e.g. Koretz & Barron, 1998 which selects students who took ACT in addition to the high-stakes test under study). This study contributes to the existing literature by looking at aggregated scores at the district-level, for district-SES performance gaps, and across an entire state.

Background

Educators and policymakers are often interested in measuring student performance in a particular domain of interest, such as in math or reading. However, the knowledge and skills encompassed in these domains of interest are often too broad to be entirely covered in a single test. Hence, the design and development of tests rely heavily on the "sampling principle of testing: test scores reflect a small sample of behavior and are valuable only insofar as they support conclusions about the larger domains of interest" (Koretz, 2008, p.21). Tests need to sample content, skills, and knowledge in a way that is representative of the domain of interest and reflects the broader goals of the domain.

At the same time, high-stakes testing has created a greater need for standardized tests – tests that are uniform in the sense that examinees "face the same tasks, administered in the same manner and scored in the same way" (Koretz, 2008, p.23) so that scores can be comparable from test to test (Koretz & Hamilton, 2006). Comparability

of scores is essential for fairness since high-stakes decisions on students, educators, and educational units are made based on these results.

Hence, one consequence of high-stakes standardized testing for test development is the creation of test forms that are similar to one another, to the extent that they are "predictable" (Koretz and Hamilton, 2006, p.568). The predictability of standardized tests has created opportunities for students and educators to engage in test preparation activities, some of which may improve learning, others which may increase scores without corresponding meaningful gains in learning (see Koretz and Hamilton, 2006 for a discussion of various test preparation activities). This raises concerns about whether scores obtained on high-stakes standardized tests reflect construct-relevant performance that can be generalized across different tests[28] used, or whether they reflect non-construct-relevant performance which hinge on the specificities of the test, and are not relevant for the target of inference[29].

The high-stakes nature of the tests has placed numerous technical demands on tests, and exposed them to greater legal scrutiny (e.g. see Schmeiser and Welch, 2006 pp. 312-313; Phillips and Camara, 2006) so much so that test design and development has become a complex and costly enterprise. This may be why states often invest in the development of a single assessment[30]. This single state assessment is often the sole measurement tool on which high-stakes accountability decisions regarding schools and districts are based on (Koretz and Hamilton, 2006). One particular means of evaluating

[28] that are representative of the domain of interest.
[29] Tests often perform different functions. Lazear (2006) proposes that from the perspective of providing incentives to learn, a good test may not have to be unpredictable. For high ability learners, an unpredictable test may provide incentives to learn a broader range of materials. For learners with high costs of learning, a predictable test may be a better motivator for learning.
[30] which may have multiple standardized test forms, across various grades.

the generalizability of student test performance across measurement tools is to develop another test that is representative of the domain of interest. However, for reasons discussed earlier, this is likely a costly and formidable enterprise in itself. Moreover, it may not be feasible to administer two high-stakes tests to students at the same time, nor be legally defensible to administer two high-stakes tests that are not built on the same test specification.

This study makes use of a change in assessment within Texas to study district-level SES gaps that is increasingly receiving attention in the literature (e.g. see Reardon, 2016). The new assessment, State of Texas Assessments of Academic Readiness (STAAR), is essentially based on the same curriculum as the old assessment, TAKS, and can serve as a "pseudo-audit" test for TAKS. The use of this new high-stakes assessment as a "pseudo-audit" tool for the old assessment has an advantage over most other studies. Earlier studies typically administer audit tests in a low-stakes context, while we examine gaps measured by the STAAR and TAKS in a high-stakes context[31]. Thus, students and educators would be motivated to do well on both tests.

One limitation in the use of STAAR as a "pseudo-audit" tool is that there is a runway period from the time the test specifications and eligible curriculum for STAAR is announced, which gives districts time to prepare for it, and which may give rise to differences and effectiveness in the ways districts prepare for the "pseudo-audit" assessment. However, the differential effectiveness with which low-SES and high-SES

[31] One caveat is that the state accountability system is temporarily suspended in the first year of assessment change (2011-2012) while the new accountability system based on STAAR is being developed. I assume that students and teachers will still be motivated to do well in the first year as part of the learning curve for the new assessment that will be used in the new accountability system. However, I also cannot rule out that districts such as lower-SES districts which are likely to face greater accountability pressure might try to depress their performance in a year where they are not held accountable, so that they have greater room to show improvement in the years when the accountability pressure is on.

districts prepare for the new high-stakes test may arguably be the very differences that we hope to capture in the performance gap. How well STAAR can function as a "pseudo-audit" measurement tool will depend on the extent that novel ways to test construct-relevant performance are introduced in the assessment, such that students cannot depend on non-construct relevant approaches to score points.

I make use of a change in Texas' assessment to estimate district-SES gaps in student test scores. In this paper, I discuss high-SES and low-SES district score gaps terms of the 75^{th}-25^{th} score gap (75^{th} district-SES percentile score versus 25^{th} district-SES score) and the 90^{th}-10^{th} score gap (90^{th} district-SES percentile versus 10^{th} district-SES percentile score). In spring 2012, Texas switched assessments, from TAKS to STAAR. STAAR is considered a more rigorous assessment that emphasizes postsecondary readiness.

Studies on Texas in the 2000s have found that the gap in average test scores between white students and racial minority students were narrower based on analyses using the existing high-stakes Texas Assessment of Academic Skills (TAAS) compared to NAEP (Klein, Hamilton, McCaffrey, & Stecher, 2000). Klein et al. (2000) also found that the relationship between SES and TAAS scores disappeared when schools were the unit of analysis, even though the relationship between SES and scores on non-TAAS tests administered by the researchers persisted, which suggests that results from TAAS scores may not generalize to findings when non-TAAS scores are used. I propose to revisit this issue with some modifications, by comparing district-level gaps, between high-SES and low-SES districts[32].

[32] The dataset that I use for this study contains data for all states. Technically, I can gain external validity by using all states which experienced an assessment change as the treatment. However, to limit the scope of

The TAKS and STAAR Assessment in Texas

Texas has a long history of standardized high-stakes testing. Each wave of testing change serves to raise the academic bar over time. From the Texas Assessment of Basic Skills (TABS) and Texas Educational Assessment of Minimum Skills (TEAMS) in the 1980s that tested "basic" or "minimum" academic skills, to the Texas Assessment of Academic Skills (TAAS) in the 1990s that includes assessment of problem-solving and complex thinking skills, to the Texas Assessment of Knowledge and Skills (TAKS) in the 2000s that measures students' mastery of the state-mandated curriculum, the State of Texas Assessments of Academic Readiness (STAAR) is the fifth test to be introduced in 2012 (Clark, 2011; Haney, 2000; Zyskowski, 2016).

The purpose of STAAR is to "increase the rigor of the assessments so that students have the academic knowledge and skills they need to meet the challenges of the 21st century" (Texas Education Agency, 2010c). The STAAR program consists of a series of assessments in grades 3-8 that are vertically linked to end-of-course assessments. These end-of-course assessments which replace the 11th grade TAKS are linked to "readiness for postsecondary endeavors" (Texas Education Agency, 2013, p.5).

STAAR differs from TAKS in many ways, particularly in rigor and test design (Texas Education Agency, 2010a). See Table 1 for a comparison. In most grades and subjects, assessments are lengthened. Items assessing skills at greater depth and level of cognitive complexity are added and the number of open-ended items ("griddable items")

---

this paper, I use Texas as a case study to delve deeper into the details of the assessment change. I also chose Texas as a case study because of the availability of (i) earlier studies that compare test score gains on Texas' state test to audit tests, and (ii) publicly available documentation of both the old and new assessments in question for this study.

are also increased. Performance standards are also set higher. Table 2 and Table 3 present the test blueprints for TAKS and STAAR for mathematics and reading respectively.

The test design for STAAR also differs from that for TAKS by focusing on fewer skills and testing those skills in a deeper way. Although the curriculum, the Texas Essential Knowledge and Skills (TEKS), which both the TAKS and STAAR are based on, did not change in 2011-2012 for mathematics nor reading[33], one difference is that the Texas Education Agency explicitly identifies a subset of TEKS that are "eligible[34]" to be assessed on STAAR. Within the "eligible" set of knowledge and skills from the TEKS, the Texas Education Agency further classifies them into "readiness standards" or "supporting standards" (Texas Education Agency, 2011; Texas Education Agency, 2012). The "readiness standards" which the Texas Education Agency defines to be "necessary both for success in the current grade or course and for preparedness in the next grade or course" (Texas Education Agency, 2010a) are emphasized on the STAAR. Even though 30% of the eligible content standards from TEKS are "readiness standards", they are covered by about 65% of the items on the STAAR (Texas Education Agency, 2013). The "supporting" standards may not be tested annually, but are still included in instruction and eligible for assessment.

Finally, STAAR also has greater speed demands than TAKS. There are more items on STAAR than on TAKS, but STAAR has a four-hour time limit, whereas there is no time limit for TAKS.

---

[33] The Mathematics TEKS in place for school year 2011-2012 was implemented from school year 2006-2007 to 2013-2014. The Reading TEKS in place for school year 2011-2012 was implemented from 2009-2010 to 2018-2019.

[34] "Eligible" standards must be amenable to being assessed on a paper and pencil test (Texas Education Agency, 2010a).

In short, STAAR differs from TAKS in that STAAR tests a smaller subset of the TEKS curriculum, but at a level of greater depth and cognitive complexity. Students also have a time limit to work on a greater number of items that are presumably more difficult.

In this study, I focus on how district-SES gaps change within Texas before and after the switch in assessment from TAKS to STAAR in spring 2012.

One working hypothesis is that district-SES gaps may widen after the state switches to a new assessment. As the hypothesis goes, lower-SES districts that face greater accountability pressure from the scrutiny on disadvantaged student groups may be more likely to teach to the old test. For example, they may teach a curriculum that focuses on knowledge and skills emphasized on the test while neglecting other important parts of the curriculum that is seldom tested (McNeil, 2000; McNeil & Valenzuela, 2001). Another working hypothesis is that district-SES may be highly correlated with family income and family investments in student enrichment and learning (see Kaushal, Magnuson, & Waldfogel, 2011). This may translate to advantages for higher-SES students within those districts as the state shifts to an assessment that focuses on higher cognitive skills.

While changes in the assessments will also affect the accountability system in Texas, the new accountability system is implemented only for the 2012-13 academic year and onwards (Texas Education Agency, 2010a). To avoid confounding changes in the accountability system and changes in assessment, I focus only on spring of 2012 as the treatment year. I ask:

RQ1: Does the district-average test score gap between low-SES and high-SES districts widen in the year immediately after the switch from TAKS to STAAR in spring 2012?

Data

I use the Stanford Education Data Archive (SEDA) 2.1 dataset (Reardon, Ho, Shear, Fahle, Kalogrides, & DiSalvo, 2018) in this study. The SEDA dataset links district-level performance for grades 3 to 8 for all U.S. states to a common NAEP scale (Reardon, Kalogrides, & Ho, 2017). I use data from the 2008-09 academic year, the first year of data in the dataset, to 2011-12, the first year of assessment change in Texas[35]. I did not use data from academic year 2012-13 and onwards because that is when a new accountability system was introduced in Texas and may confound the effect of assessment change. Additionally, districts may start to develop non-learning related strategies as they gain experience with the assessment.

The SEDA 2.1 achievement data is constructed using publicly available data on each state's standardized testing program from the ED*Facts* data system at the U.S. Department of Education. Briefly, the ED*Facts* data consists of "coarsened" data where aggregated student data is reported in terms of the number of students in each proficiency category, for each grade from grades 3 through 8 for math and reading/language arts. Using ordered probit models, SEDA estimates the mean and standard deviation of achievement for districts (and other geographic units) from the proficiency count data (Fahle, Shear, Kalogrides, Reardon, DiSalvo, & Ho, 2018; Ho & Reardon, 2012; Reardon

[35] For reading, I further constrain data used to include scores for spring of years 2011 and 2012 because the TEKS curriculum for reading changed in school year 2009-2010. See Appendix A for details about the scores used for analyses by year-grade.

& Ho, 2015). The estimated means and standard deviations for each state-subject-year-grade are standardized and placed on the NAEP scale[36] (Reardon, Kalogrides, & Ho, 2017). In this study, we use estimates from SEDA that are standardized using the CS scale, obtained by dividing the estimates by the national grade-subject-specific standard deviation for the cohort of students who were 4th grade in spring 2009 (and 8th grade in spring 2013). The group means (districts in our case) recovered from coarsened data in this way can be used to estimate between-group achievement gaps (Reardon, Shear, Castellano, & Ho, 2017) and allows us to study district gap trends over states, years, and grades[37, 38].

Measures

In this study, I use the district-average math and reading scores (by subject-grade-year for all students within the district) standardized in the SEDA dataset using the CS scale as described above.

For the district-SES measure, I use the standardized SES composite variable provided within SEDA 2.1. The source variables of this composite are obtained using the Education Demographic and Geographic Estimates (EDGE), the school district-level tabulation of the American Community Survey (ACS). It is computed as the first principal component score of the following standardized measures for the academic year

[36] Estimates for non-tested grades in NAEP assessment years are interpolated for grades 5, 6, and 7 and extrapolated for grade 8. Estimates for grades 3 to 8 in non-NAEP assessment years are interpolated across years (Reardon, Kalogrides, & Ho, 2017).

[37] The CS scale may not permit absolute comparisons across grades, but it allows us to compare gaps in district-SES percentile performance across grades. The grade (within cohort) standardized scale (gcs) within the SEDA dataset may allow comparisons across grades and years, but it does so by making more assumptions about the NAEP scale (see Fahle et al., 2018). Since, we are primarily interested in district-SES gap trends, we use the cohort standardized scale.

[38] Technically, this study can be conducted by using continuous test score data from Texas, which would not require using the models and assumptions that SEDA used to recover the test scores. However, the advantage of using SEDA is that it is publicly available and allows for cross-state analysis.

2008-09: median income, percent of adults ages 2 and older with a bachelor's degree or higher, poverty rate for households with children ages 5-17, SNAP receipt rate, single mother headed household rate, and employment rate for adults ages 25-64 (Fahle et al., 2018). Throughout the analyses, I use district-SES percentiles that are calculated for Texas[39].

Analytical Strategy

Our primary effect of interest is the gap in district-average performance between low-SES and high-SES districts measured when the assessment in Texas changes (in school year 2011-12, henceforth referred to as year 2012). However, the one year lag in the administration of the old and new test may be confounded by other education and economic changes happening in Texas and nation-wide. To provide counterfactual district-SES score trends, I construct a comparison group consisting of states that did not change their assessment and assessment policies, namely no change in assessment standards, mode (paper-and-pencil or computerized) of assessment, or performance cut scores. See Appendix B for further details.

I use a difference-in-difference estimation strategy to estimate differences in district-SES gaps measured using TAKS and STAAR. Substantively, I am interested in differences in the measured gap in district performance[40].

Even though there are multiple years of pre-assessment change data available for math, I use a difference-in-difference design rather than a comparative interrupted time series (CITS) design. The latter allows for both baseline average and growth trends but

[39] This avoids confounding the differences in SES gradients across states.
[40] To be consistent with the difference-in-difference convention, I refer to the gap in performance between high-SES and low-SES districts as the first difference, the pre-/post-differences in this gap as the second difference, and the difference between Texas versus comparison state as the third difference.

requires at least 6 time-points of baseline data (see Somers, Zhu, Jacob, & Bloom, 2013). In this study, each cohort has at most three, or sometimes less years of baseline data. To address serial correlation among the outcome from year to year (Bertrand, Duflo, & Mullainathan, 2004), I average the data at the district-level before the assessment change.

I estimate the district-SES test score gaps within Texas before and after the assessment change, i.e., the difference-in-difference effects, using the following model:

(1): $Score_{cd} = \beta_0 + \beta_1 f(\boldsymbol{SES}_d) + \beta_2 Post + \beta_3 f(\boldsymbol{SES}_d) \times Post + \beta_4 \boldsymbol{X}_d + \varepsilon_{cd}$

Score is the district-average test score in math or reading for cohort *c* in district *d* within Texas. For each cohort *c*, I average the district-average test score over the pre-test change years to address serial correlation among the outcome from year to year (see Bertrand, Duflo, & Mullainathan, 2004). *f*(***SES***) is a cubic polynomial function[41] of the district overall-SES composite[42] included in the SEDA dataset for 2009. *Post* is an indicator with a value of 1 if the year is 2012, the first year when Texas switched from the TAKS to STAAR, and a value of 0 otherwise. The interaction terms between *Post* and the cubic polynomial function of *SES* provide our main parameters of interest. By substituting in the 25th district-SES percentile and 75th district-SES percentile (or the 90th and 10th district-SES percentile) for Texas, I calculate the 75th-25th (or 90th-10th) district-SES percentile gap before and after the assessment change. Taking the before and after difference in 75th-25th (or 90th-10th) district-SES gap provides us with an estimate of the difference in measured gap between high-SES and low-SES districts after the assessment

[41] In Reardon's (2017) paper that uses the SEDA dataset, the relationship between district-average test scores (pooled across years and subjects) and district-SES composite (which I also use in this paper) is modeled as a cubic function. Based on visual inspection of the data used in my analyses (see Appendix C) and AIC statistics, a cubic polynomial function appears to best model the relationship between district-average test scores (by subject-year-grade) and district-SES among the various polynomial functions.
[42] See section on *Measures* for a description of how the SES composite is constructed in the SEDA dataset.

change relative to the pre-change years within Texas[43]. I then test the null hypothesis that the linear combination of the *Post*×*f*(**SES**) interaction terms evaluated at the 75th and 25th (90th and 10th) district-SES percentiles is equal to zero[44], i.e. that there is no difference in 75th-25th (90th-10th) district-SES percentile gap as measured by STAAR compared to TAKS.

I use cohort fixed effects to compare estimates within each cohort (or approximately the same group of students) over time. I include a vector of district-grade covariates (district-level covariates that vary across grades), ***X***, to control for any changes in demographics and background differences within the cohort over time[45]. Since the treatment assignment is at the state level across all cohorts at a single point in time (year 2012), I did not cluster standard errors (see Abadie, Athey, Imbens, & Wooldridge, 2017).

I also calculate a triple difference-in-difference estimate where I compare the difference-in-difference estimates obtained for Texas to a comparison group which comprises of other states that did not change assessment from 2009 to 2012. This triple

[43] If β_{3a}, β_{3b}, and β_{3c} are the parameters for the *SES*, SES^2, and SES^3 terms respectively, then the difference in gap between high-SES and low-SES districts before and after the assessment change is: $\beta_{3a}(p75 - p25) + \beta_{3b}(p75^2 - p25^2) + \beta_{3c}(p75^3 - p25^3)$. I use the notation *p75* to refer to the 75th district-SES percentile. A similar reasoning follows for the 90th 10th district-SES gap change.

[44] The standard error for this linear combination is obtained using the *lincom* command in Stata.

[45] The district-grade covariates include: log of student enrollment by district-grade-year, percentage of white students in the district-grade, percentage of students on free or reduced price lunch in the district-grade, percentage of all students in the district that are in special education, log of percentage of all students in the district that are English Language Learners, percentage of students living in the same house as the previous year, Gini coefficient for the district, percentage of 15-19 year olds giving birth in the district, log of number of schools in the district, percentage of schools in the district that are charter schools, square root of percentage of students in charter schools in the district, district-average of pupil-teacher ratio in students' school, log of total per-pupil expenditure in the district, and an indicator if the district is in a rural locale. All untransformed covariates are provided in the SEDA dataset, which in turn obtains the variables from the Education Demographic and Geographic Estimates (EDGE) and the Common Core of Data (CCD) (Fahle et al., 2018). I take the average of these covariates for each cohort within district across the pre-test change years.

difference takes into account changes in national education policies and other events or trends that may affect district-average performance over the period. I estimate the triple difference using the relationship between district-average test scores and district-SES as follows:

$$(2): Score_{cds} = \beta_0 + \beta_1 f(\boldsymbol{SES}_{ds}) + \beta_2 Post + \beta_3 f(\boldsymbol{SES}_{ds}) \times Post + \beta_4 f(\boldsymbol{SES}_{ds}) \times TX_s + \beta_5 Post \times TX_s + \beta_6 f(\boldsymbol{SES}_{ds}) \times Post \times TX_s + \boldsymbol{X}_d + \boldsymbol{W}_s + \varepsilon_{cds}$$

where the terms are defined as before. The subscript *s* refers to states. *TX* is an indicator for the treatment state, Texas, and $\boldsymbol{W}$ is a vector of state-level covariates that vary across years[46]. The interaction terms between *Post*, *TX*, and the cubic polynomial function of *SES* provide our main parameters of interest. By substituting in the 25th district-SES percentile and 75th district-SES percentile (or the 90th and 10th district-SES percentile) for Texas and evaluating the difference, I obtain the 75th-25th (90th-10th) district-SES percentile score gap before and after the assessment change, in Texas compared to the comparison states. I use cohort-by-state fixed effects to restrict the gap comparisons within each cohort and state. I cluster standard errors by state since the treatment assignment is at the state level (see Abadie et al., 2017).

Results

Figure 1 illustrates the trend in 75th-25th (90th-10th) district-SES percentile gaps in test scores within Texas between 2009 and 2012 for math and between 2011 and 2012 for

[46] The state-year covariates include: log of median household income, log of total expenditures per pupil, instruction as percentage of current expenditures, and percentage of white students in public schools. The untransformed variables are obtained from the National Center for Education Statistics (retrieved from http://nces.ed.gov/ccd/elsi/) with the exception of median household income which is obtained from the Department of Numbers (https://www.deptofnumbers.com/income/). I take the average of these covariates for each cohort within state across the pre-test change years.

reading, obtained by fitting equation (1) separately for each subject[47]. Visual inspection of these graphs suggest that the gap in mean district scores between high-SES and low-SES districts (75^{th}-25^{th} and 90^{th}-10^{th} district-SES percentile gaps) widened slightly when Texas switched from TAKS to STAAR.

Table 4 shows the results of fitting equation (1) for Texas to obtain the difference-in-difference estimates for the 75^{th}-25^{th} (90^{th}-10^{th}) district-SES percentile gaps before and after the assessment change. The 75^{th}-25^{th} district-SES percentile gap differs by about +0.037 standard deviation units for math and about +0.041 standard deviation units for reading after the switch to STAAR relative to the baseline years when TAKS was used. The 90^{th}-10^{th} district-SES percentile gap is also wider when measured using STAAR compared to TAKS by about +0.073 standard deviation units for math and +0.088 standard deviation units for reading. The point estimates are quite stable across models with no covariates and with district-by-grade covariates. The difference in gap after the assessment change is statistically significant for both math and reading.

While these results suggest that there may be differences in district-SES test score gaps when the measurement tool changed from TAKS to STAAR within Texas, this may be confounded with national trends happening between 2009 and 2012, due to the time lag in the administration of these assessments. For example, economic conditions may be deteriorating due to the great recession around that period, which may cause the district-SES performance gap to change over time. To account for such national trends, we turn our attention to cross-state comparisons.

[47] I do not draw trend lines that connect average outcomes for pre- and post- years because the CS scale used for the outcomes does not permit absolute comparison of performance across grades. However, this scale allows us to compare gaps across grade. See Data section for details.

Figure 2 compares the magnitude of the difference-in-difference estimates obtained for Texas relative to each of the comparison states. These estimates are obtained from fitting equation (1) and calculating the difference-in-difference estimates for the 75^{th}-25^{th} (90^{th}-10^{th}) district-SES percentile score gaps[48] separately for each state and subject. The point estimates for the district-SES performance gaps in Texas, though small, appear larger than that compared to most of the other comparison states. Unlike for most other comparison states, we can reject the null hypothesis that the difference in district-SES score gaps measured before and after the assessment change in Texas is zero for both math and reading.

Figure 3 graphically shows the results when we pool the estimates across the comparison states, obtained from fitting equation (2). This is essentially a triple difference-in-difference estimate that provides the difference in mean district scores between high-SES and low-SES districts, before and after the assessment change, in Texas relative to the comparison states. Figure 3 suggests that while the pre-/post-gradient difference in 75^{th}-25^{th} (90^{th}-10^{th}) district-SES percentile gaps for the comparison states is relatively flat for both math and reading, the corresponding pre-/post- gradient difference in Texas is steeper, i.e., the difference in observed 75^{th}-25^{th} (90^{th}-10^{th}) district-SES percentile gap in Texas after the switch from TAKS to STAAR in 2012 is larger in Texas compared to the comparison states pooled together.

Table 5 shows the corresponding estimates displayed in Figure 3. After accounting for the secular trend among comparison states in observed gaps between high-SES and low-SES districts over the same period, the observed 75^{th}-25^{th} and 90^{th}-10^{th}

[48] evaluated at the relevant district-SES percentiles for Texas.

district-SES percentile gap in Texas is wider by about 0.029 and 0.057 standard deviation units respectively for math after the switch from TAKS to STAAR. For reading, the corresponding 75^{th}-25^{th} and 90^{th}-10^{th} district-SES percentile gap is wider by about 0.021 and 0.051 standard deviation units respectively after taking into account the secular trend in comparison states across the same time period. All of these differences in measured gap between the two tests are statistically significant.

Sensitivity Checks

I conduct checks on the sensitivity of findings to various specifications for the triple difference estimates (Table 5) that indicate the difference in measured gaps using STAAR relative to TAKS, in Texas relative to the comparison states.

First, I added district fixed effects. The district fixed effects would limit comparisons of changes in district-average scores within districts. The results in Table 5 Panel A suggests that the substantive findings do not change with the inclusion of state-cohort-district fixed effects.

I also check the sensitivity of findings to the functional form specification (Panels B and C) for the relationship between district-average scores and district-SES. For the cross-state analysis, the checks suggest that the findings for 75^{th}-25^{th} district-SES percentile score gaps may be sensitive to the functional form specified, but the findings for 90^{th}-10^{th} district-SES gaps are robust.

Threats to Validity

One assumption of the difference-in-difference strategy is that in the absence of the assessment change, the slope of change in district-SES score gaps from 2011 to 2012 in Texas would be the same as the slope change in district-SES score gaps in comparison

states over the same period. This is a strong assumption and cannot be directly tested. However, having parallel trends before the change would add confidence that this method is suitable. In Appendix D, I plot a graph that estimates the mean district-average math[49] score separately for each year within the study period, at the 25th and 75th district-SES percentiles. This graph suggests approximately parallel trends before the change.

Another threat to validity is that I am relying on a district-SES composite based on SES variables measured in 2009. One possibility for the observed widening district-average performance between high-SES and low-SES districts may be due to increased sorting of students by SES, such that higher-SES students are sorting into higher-SES districts (or students in those districts are growing richer), and lower-SES students are sorting into lower-SES districts (or students in those districts are growing poorer). I am unable to check for this directly, because some of the variables used to form the district-SES composite are collected in the SEDA dataset only for the year of 2009. In Table 7, I use other SES-related variables that vary by district-year for Texas available within the SEDA dataset to check for this possibility. I look at how the percentage of students with free and reduced-price lunch change within each decile. I find that in general, across all district-SES deciles within Texas, the percentage of students on free and reduced-price lunch has increased. Within the same period, the total per-pupil expenditure (total expenditure/enrollment) as well as the per-pupil instructional expenditure (instruction expenditure/enrollment) also dropped across all deciles. It does not appear that higher-SES students are sorting into higher-SES districts or higher-SES districts are spending more on education, and vice versa for the lower-SES districts.

[49] I did not plot a corresponding graph for reading since only one time-point is used for the pre-change period.

This study relies on the linking of coarsened test score distributions (by state, district, subject, year, and grade) to an interpolated NAEP score for even years when NAEP was not administered, and for grades 3, and 5 to 7, for the state-average test scores. If actual even year test scores deviate from the linear trend between odd years, then the interpolated scores for the even years would be affected. Reardon, Kalogrides, and Ho (2017) provide a series of empirical checks on the validity of this approach and find that the interpolation is generally sound. I also look at NAEP data for Texas as well as NAEP-TUDA (Trial Urban District Assessment) data (data not shown here) for districts within Texas, and find that score trends are quite stable across administrations of the assessments within the study period. I assume that there is no wild fluctuation in scores for year 2012, and that this stability of score trend allows for linear interpolation of scores across years.

Discussion and Conclusion

This study examines whether district-SES test score gaps differ when measured using the old and new state assessment in Texas. I find that for both math and reading, district-average test score gaps between high-SES and low-SES districts are slightly wider in 2012, the first year when the STAAR assessment was administered in Texas, compared to the baseline years (2009 to 2011) when TAKS was administered. This is the case whether we compare districts at the 75^{th} and 25^{th} district-SES percentile, or districts at the 90^{th} and 10^{th} district-SES percentile, and whether we solely look within Texas or compare Texas to other states that did not change assessment over the same period. The finding for the 90^{th}-10^{th} gap is robust across a number of specifications, but the finding

for 75th-25th gap is sensitive to polynomial specifications. The magnitude of these differences, however, are very small[50].

One caveat about interpreting these effects is that we are looking at district-average test scores. The findings are relevant for test score gaps between districts at high district-SES percentiles and districts at low district-SES percentiles. The ecological fallacy warns us that findings about district aggregate scores does not imply the same for individual student test scores. Hence, the findings on district-SES score gaps apply only at the district-level but not the student-level. In other words, the findings do not imply that test score gaps between high-SES *students* and low-SES *students* widened in 2012 within Texas.

The results from this study is in some ways consistent with the results from earlier studies in that score trends found using an existing high-stakes test may be different when a new test is introduced, or when an audit test is used, although the magnitude of differences that I find is much smaller. One advantage that this study has over other studies is that the sample includes nearly all districts from the state, and hence is generalizable to the state. Another advantage is that the tests used are both administered under high-stakes conditions.

There are a few possible reasons for the difference in magnitude found between this study and earlier studies. First the nature of the gap is different. In this study, we look at district-SES gaps while earlier studies look at student-level SES gaps. Another reason is because our sample consists of state-wide test scores, whereas earlier studies are based on representative samples, or constrained to specific districts. Finally, the difference in

[50] This may be because we are looking at not only district-average test scores, but also differences in gaps.

magnitude between this study and earlier studies could be because here, both tests are administered under high-stakes conditions, which may motivate students to do well on each of those tests.

The findings of this study also suggest that when there is an assessment change, lower-SES districts are more disadvantaged compared to higher-SES districts on the new test, or higher-SES districts are more advantaged compared to lower-SES districts. Unfortunately, the scale that we used for this study does not allow us to distinguish year-to-year performance on an absolute scale, hence we are unable to distinguish whether the difference in gaps is due to lower-SES districts performing less well on the new assessment, or performing better on the new assessment but not as well relative to higher-SES districts.

One peculiar aspect of STAAR is that while it attempts to be more "rigorous" by including more open-ended items and items that test greater cognitive complexity, it appears to intentionally focus on a narrower part of the curriculum[51]. The "readiness standards" which comprises 30% of the eligible content standards from the curriculum are given a weightage of about 65% on STAAR (Texas Education Agency, 2010c). In addition, STAAR only focuses on the curriculum taught within the school year whereas TAKS assesses the cumulative curriculum up till the present year. Thus, while STAAR may increase the cognitive difficulty of material tested, it also presents opportunities to narrow the curriculum. Furthermore, where there is no time limit for TAKS, there is a time limit of four hours for STAAR. Our results suggests that lower-SES districts may not be performing as well as higher-SES districts on such a test.

[51] albeit narrowing the focus to parts of the curriculum with greater relevance for "college- and career-readiness".

There may be a number of reasons why this is the case. One reason may be that the time limit may make a difference. Students from lower-SES districts may need the extra time to complete the test, or the psychological burden of having a time limit may affect their test performance. Another reason may be because lower-SES districts are not preparing their students as well on skills requiring greater cognitive complexity despite an explicit narrower focus on the state curriculum, relative to the higher-SES districts. Lower-SES districts may also have been focusing on different parts of the curriculum that are more emphasized on TAKS than on STAAR. We also cannot rule out lower-SES districts having a greater reliance on test preparation practices on TAKS due to the enormous accountability pressures that they face, but which may not carry over well in the preparation for STAAR.

Another major reason may be due to the high correlation between SES and the availability of educational resources. Higher-SES districts may have more resources to invest in building up capacity to better prepare their students for STAAR. Families of students from higher-SES districts may also have the resources to invest in enrichment programs that deepen student learning on the state curriculum, or broaden learning beyond the state curriculum in a way that translates into an advantage for them on the new assessment.

Although I cannot determine the reasons causing the differences in measured district-SES gaps using the old and new assessment, I propose that these are the relative advantages or disadvantages that give rise to differential learning opportunities and performance between high-SES and low-SES districts to be captured through measuring district-SES gaps. The advantage of using an alternative assessment is to minimize

construct-irrelevant factors that may affect one type of district more than another, such as use of test preparation activities that do not impact learning.

Our results suggest that overall to the changes that have been made to the STAAR assessment, the gap in district-average test scores between higher- and lower-SES districts is wider when measured on the new assessment, STAAR, compared to that measured on the old assessment, TAKS. This is despite the opportunities to focus on a narrower segment of the curriculum. This may suggest that the wider gap is due to students in lower-SES districts faring poorer on average, relative to students in higher-SES districts (or students in high-SES districts faring better than those in low-SES districts), on the more challenging items tested.

One shortcoming of this study is that it only looks at results on the aggregate, and does not examine the processes that may lead to these results. I suggest future work in two directions.

The first is to understand district responses when there is an assessment change. What strategies do districts adopt to prepare their students for a change in assessment? Which of these impact teaching and learning in the classroom, and which only impact test scores? Are all students similarly affected by these strategies? Are there differences in how higher-SES and lower-SES districts respond to a change in assessment and in what ways are they different?

A second area for future work is to continue to replicate findings made regarding schools, districts, or groups about their test scores and performance gaps using assessments other than that obtained via long-administered high-stakes tests. As many earlier studies have shown, the findings from high-stakes tests that have been

administered for a while and have gained familiarity by districts may sometimes not be generalizable to other audit tests or low-stakes tests. Such work has focused mainly on student-level performance. With the availability of a dataset such as SEDA which makes district-level data available for nearly all districts in the state and nation-wide, there are opportunities for more studies to replicate the findings made at the district-level.

Finally, high-stakes decisions regarding schools and districts are often made based on results from the state's only assessment. What exactly do high-stakes tests help us infer about the progress that schools and districts are making towards preparing their students on the academic knowledge and skills that have relevance for the workplace, for college, and for the future, as opposed to the knowledge and skills emphasized by state assessments? Would the identification and decisions made regarding schools and districts be consistent if the predictability is taken out of high-stakes standardized tests without sacrificing representativeness of the tested domain? Would periodic use of such audit tests help assessments become better reform tools to inform and motivate educators towards improving education for all students?

Tables and Figures

Table 1. Comparison of TAKS and STAAR

	TAKS	**STAAR**
Assessed curriculum	Texas Essential Knowledge and Skills (TEKS). Educator committees identified student expectations that should be assessed on a statewide assessment, which the TEA later developed into TAKS objectives[1] with further inputs from Texas educators and the public.	Texas Essential Knowledge and Skills (TEKS). Educator committees identified a subset of Texas Essential Knowledge and Skills that are "eligible" to be tested on STAAR, and further classify this eligible subset into "readiness standards" and "supporting standards". Assessment of "readiness standards" are emphasized in STAAR[2].
Rigor of assessment	Focuses on mastery of the TEKS curriculum.	Overall test difficulty increased by including more "rigorous" items that assess skills at a "greater depth" and level of "cognitive complexity".
Number of items	See Table 2 and Table 3 for the number of items by grade and subject.	The number of items for STAAR are increased for most grades and subjects. See Table 2 and Table 3 for the number of items by grade and subject.
Item format	In math, most items on TAKS are in multiple-choice format with a limited number of open-ended griddable items.	In math, the number of open-ended griddable items on most tests are increased.
Test duration	Untimed	4-hour limit
Mode of administration	Paper administration for grades 3-8	Paper administration for grades 3-8
Test contractor	Pearson	Pearson

[1] See Table 2 and Table 3 for examples of TAKS objectives in mathematics and reading respectively.

[2] See Table 2 and Table 3 for the distribution of "readiness standards" and "supporting standards", and the distribution of items across the standards in mathematics and reading respectively.

References:

Texas Education Agency (2010a). *House Bill 3 transition plan. A report to the 82nd Texas Legislature from the Texas Education Agency.* Retrieved from https://tea.texas.gov/student.assessment/hb3plan/

Texas Education Agency (2011). 2011 District and Campus Coordinator Manual. Retrieved from https://web.archive.org/web/20110220045422/http://www.tea.state.tx.us/student.assessment/manuals/dccm/

Texas Education Agency (2012). 2012 District and Campus Coordinator Manual. Retrieved from https://web.archive.org/web/20120829142834/http://www.tea.state.tx.us/student.assessment/manuals/dccm/

Table 2. Test blueprints for TAKS (2011) and STAAR (2012) mathematics

TAKS						
	No. of Items by Grade					
TAKS Objectives[1]	**3**	**4**	**5**	**6**	**7**	**8**
Numbers, Operations, and Quantitative Reasoning	10	11	11	10	10	10
Patterns, Relationships, and Algebraic Reasoning	6	7	7	9	10	10
Geometry and Spatial Reasoning	6	6	7	7	7	7
Measurement	6	6	7	5	5	5
Probability and Statistics	4	4	4	6	7	8
Mathematical Processes and Tools	8	8	8	9	9	10
Total no. of items[2]	**40**	**42**	**44**	**46**	**48**	**50**
STAAR						
	No. of Items by Grade					
STAAR Reporting Categories[1]	**3**	**4**	**5**	**6**	**7**	**8**
Numbers, Operations, and Quantitative Reasoning	15	17	18	16	13	11
Patterns, Relationships, and Algebraic Reasoning	8	6	6	12	13	14
Geometry and Spatial Reasoning	9	12	7	8	10	8
Measurement	8	8	8	8	8	13
Probability and Statistics	6	5	11	8	10	10
Mathematical Processes and Tools[3]	See note 3.					
Total no. of items on test	**46**	**48**	**50**	**52**	**54**	**56**
Breakdown of items by format						
Multiple Choice	43	45	47	48	50	52
Griddable	3	3	3	4	4	4
Breakdown of items by type of standard						
No. of items testing readiness standards	28-30	29-31	30-33	31-34	32-35	34-36
No. of items testing supporting standards	16-18	17-19	17-20	18-21	19-22	20-22
	No. of Eligible Standards by Grade					
Total no. of readiness standards	9	10	10	10	12	11
Total no. of supporting standards	19	23	20	21	23	22

[1] The TAKS objectives are “umbrella statements” that serve as headings under which student expectations from the TEKS can be meaningfully grouped. They are called "reporting categories" under STAAR.

[2] Most items on math TAKS are in multiple-choice format with a limited number of open-ended griddable items.

[3] The STAAR blueprints state: "Underlying Processes and Mathematical Tools is not a separate reporting category. These skills will be incorporated into at least 75% of the test questions from reporting categories 1–5 and will be identified along with the content standards."

Note: Tests blueprints for TAKS and STAAR are retrieved from the following websites:
2011 TAKS Blueprints
https://web.archive.org/web/20110220050407/http://www.tea.state.tx.us/student.assessment/taks/blueprints/
2012 STAAR Test Blueprints
https://web.archive.org/web/20120626183326/http://www.tea.state.tx.us/student.assessment/staar/blueprints/

Table 3. Test blueprints for TAKS (2011) and STAAR (2012) reading

TAKS

TAKS Objectives[1]	No. of Items by Grade 3	4	5	6	7	8
The student will demonstrate a basic understanding of culturally diverse written texts.	15	15	13	13	12	12
The student will apply knowledge of literary elements to understand culturally diverse written texts.	7	8	8	8	10	10
The student will use a variety of strategies to analyze culturally diverse written texts.	6	7	8	8	10	10
The student will apply critical-thinking skills to analyze culturally diverse written texts.	8	10	13	13	16	16
Total no. of items	**36**	**40**	**42**	**42**	**48**	**48**

STAAR

STAAR Reporting Categories[1]	No. of Items by Grade 3	4	5	6	7	8
The student will demonstrate an ability to understand a variety of written texts across reading genres.	6	10	10	10	10	10
The student will demonstrate an ability to understand and analyze literary texts.	18	18	19	20	21	22
The student will demonstrate an ability to understand and analyze informational texts.	16	16	17	18	19	20
Total no. of items on test	**40**	**44**	**46**	**48**	**50**	**52**
Breakdown of items by type of standard						
No. of items testing readiness standards	24-28	26-31	28-32	29-34	30-35	31-36
No. of items testing supporting standards	12-16	13-18	14-18	14-19	15-20	16-21
	No. of Eligible Standards by Grade					
Total no. of readiness standards	12	13	15	13	14	13
Total no. of supporting standards	11	14	19	21	20	21

[1] The TAKS objectives are “umbrella statements” that serve as headings under which student expectations from the TEKS can be meaningfully grouped. They are called "reporting categories" under STAAR.

Note: Tests blueprints for TAKS and STAAR retrieved from the following websites:
2011 TAKS Blueprints
https://web.archive.org/web/20110220050407/http://www.tea.state.tx.us/student.assessment/taks/blueprints/
2012 STAAR Test Blueprints
https://web.archive.org/web/20120626183326/http://www.tea.state.tx.us/student.assessment/staar/blueprints/

Table 4. Differences in 75th-25th (DiD7525) and 90th-10th (DiD9010) district-SES percentile gaps within Texas measured using STAAR relative to TAKS, and parameter estimates from models fitted using equation (1), by subject

	Math		Reading	
	(1)	(2)	(3)	(4)
Calculated Difference				
DiD7525	**0.047****	**0.037***	**0.042***	**0.041****
	(0.018)	(0.017)	(0.016)	(0.015)
DiD9010	**0.086****	**0.073****	**0.085*****	**0.088*****
	(0.027)	(0.025)	(0.024)	(0.022)
Parameter Estimates				
SES	0.144***	0.053***	0.176***	0.068***
	(0.012)	(0.014)	(0.012)	(0.013)
SES2	0.071***	0.048***	0.061***	0.049***
	(0.007)	(0.007)	(0.007)	(0.007)
SES3	0.021***	0.019***	0.021***	0.021***
	(0.004)	(0.004)	(0.005)	(0.004)
Post	0.129***	0.122***	0.052***	0.045***
	(0.011)	(0.011)	(0.010)	(0.010)
PostxSES	0.041*	0.032*	0.036*	0.034*
	(0.017)	(0.016)	(0.015)	(0.014)
PostxSES2	0.005	0.008	0.009	0.012
	(0.010)	(0.009)	(0.009)	(0.008)
PostxSES3	-0.001	0.002	0.004	0.007
	(0.006)	(0.006)	(0.005)	(0.005)
Intercept	-0.080***	-0.245	-0.161***	-0.054
	(0.008)	(0.221)	(0.008)	(0.185)
Number of districts	4667	4598	5762	5714
District-grade covariates	no	yes	no	yes
State-year covariates	no	no	no	no

Note: Standard errors and estimates come from regression of equation (1) with fixed effects by cohort. Parameter estimates for covariates for models fitted in columns (2) and (4) are not shown.
†$p < .10$. *$p < .05$. **$p < .01$. *** $< .001$.

Table 5. Impact of switch from TAKS to STAAR on measured 75th-25th (TDiD7525) and 90th-10th (TDiD9010) district-SES percentile gaps in Texas relative to comparison states, and parameter estimates estimated from models fitted using equation (2), by subject

	Math			Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
Calculated Difference						
TDiD7525	**0.035*****	**0.030***	**0.029***	**0.034*****	**0.028****	**0.021***
	(0.008)	(0.012)	(0.011)	(0.004)	(0.009)	(0.009)
TDiD9010	**0.064*****	**0.058****	**0.057****	**0.062*****	**0.057*****	**0.051****
	(0.014)	(0.019)	(0.017)	(0.008)	(0.015)	(0.014)
Parameter Estimates						
SES	0.221***	0.092***	0.092***	0.243***	0.107***	0.110***
	(0.013)	(0.017)	(0.016)	(0.014)	(0.013)	(0.015)
SES2	0.040***	0.035***	0.035***	0.040**	0.038***	0.035***
	(0.009)	(0.005)	(0.005)	(0.012)	(0.006)	(0.005)
SES3	0.008**	0.009***	0.009***	0.007*	0.009***	0.008**
	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
Post	-0.010	-0.004	0.000	-0.001	0.004	0.002
	(0.006)	(0.008)	(0.012)	(0.006)	(0.008)	(0.013)
PostxSES	0.011	0.005	0.006	0.013**	0.016†	0.014
	(0.007)	(0.012)	(0.011)	(0.004)	(0.009)	(0.009)
PostxSES2	0.001	0.003	0.003	0.002	0.004	0.007**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
PostxSES3	0.000	0.001	0.001	0.000	0.001	0.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
TXxSES	-0.078***	-0.060***	-0.060***	-0.077***	-0.057***	-0.051***
	(0.013)	(0.012)	(0.012)	(0.014)	(0.008)	(0.009)
TXxSES2	0.031**	0.002	0.002	0.013	-0.008	-0.009
	(0.009)	(0.009)	(0.009)	(0.012)	(0.009)	(0.009)
TXxSES3	0.012***	0.008***	0.008***	0.016***	0.011***	0.008*
	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
PostxTX	0.139***	0.130***	0.133***	0.046***	0.039***	0.059**
	(0.006)	(0.009)	(0.024)	(0.006)	(0.009)	(0.016)
PostxTXxSES	0.030***	0.026*	0.025*	0.031***	0.024*	0.018*
	(0.007)	(0.011)	(0.010)	(0.004)	(0.009)	(0.008)
PostxTXxSES2	0.004	0.004	0.004	0.015***	0.008**	0.009**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
PostxTXxSES3	0.000	0.001	0.001	0.001	0.003	0.006**
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Intercept	0.004	-0.598*	7.212	-0.027*	-0.676*	4.157
	(0.008)	(0.249)	(5.158)	(0.012)	(0.258)	(4.695)
Number of districts	32417	31373	31373	40388	39111	39294
District-grade covariates	no	yes	yes	no	yes	yes
State-year covariates	no	no	yes	no	no	yes

Note: Standard errors are clustered by state. All estimates come from regression of equation (2) with fixed effects by cohort-state. Parameter estimates for covariates for models fitted in columns (2), (3), (5), and (6) are not shown.
†$p < .10$. *$p < .05$. **$p < .01$. *** $< .001$.

Table 6. Sensitivity checks for differences in measured 75th-25th (TDiD7525) and 90th-10th (TDiD9010) district-SES percentile gaps in Texas relative to comparison states, estimated from equation (2), by subject

	Math	Reading
Panel A. Add district fixed effects		
TDiD7525	0.035*	0.045***
	(0.014)	(0.013)
TDiD9010	0.066**	0.092***
	(0.020)	(0.019)
No. of districts	31373	39215
Panel B. Include 4th polynomial degree		
TDiD7525	0.012	0.006
	(0.008)	(0.008)
TDiD9010	0.049**	0.029*
	(0.014)	(0.011)
No. of districts	31373	39172
Panel C. Include 5th polynomial degree		
TDiD7525	0.009	0.010
	(0.008)	(0.008)
TDiD9010	0.048**	0.028*
	(0.013)	(0.012)
No. of districts	31373	39172

Note: Standard errors are clustered by cohort-state. All estimates come from regression of equation (2) with fixed effects by cohort-state unless otherwise noted.
†$p < .10$. *$p < .05$. **$p < .01$. *** $< .001$.

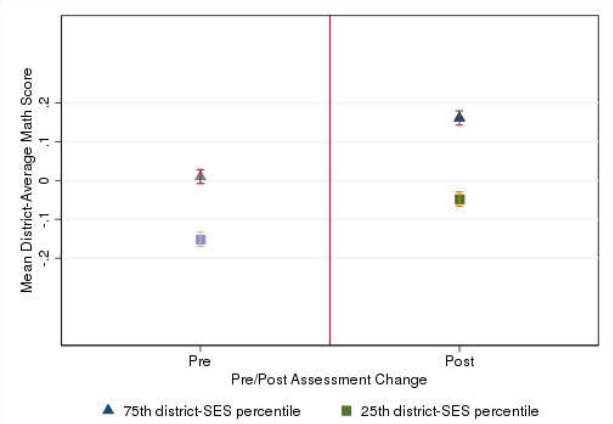
Table 7. Within grade-year mean of SES variables that vary by year and grade, by deciles of district-SES (1=lowest decile, 10=highest decile) for cohort 2006 within Texas

cohort	year	grade	district-SES decile	mean perfrl	mean ppe_tot	mean ppe_inst
2006	2009	3	1	0.514	12343	5726
2006	2010	4	1	0.572	12759	5939
2006	2011	5	1	0.593	11764	5787
2006	2012	6	1	0.557	11449	5430
2006	2009	3	2	0.581	11796	5698
2006	2010	4	2	0.610	12640	5989
2006	2011	5	2	0.627	11783	5778
2006	2012	6	2	0.599	10971	5454
2006	2009	3	3	0.578	12199	5602
2006	2010	4	3	0.602	12191	5929
2006	2011	5	3	0.607	11661	5795
2006	2012	6	3	0.557	10842	5428
2006	2009	3	4	0.535	11549	5594
2006	2010	4	4	0.560	11410	5796
2006	2011	5	4	0.565	11456	5671
2006	2012	6	4	0.534	10832	5347
2006	2009	3	5	0.494	11713	5420
2006	2010	4	5	0.527	11565	5690
2006	2011	5	5	0.542	10762	5554
2006	2012	6	5	0.506	10164	5155
2006	2009	3	6	0.459	12258	5592
2006	2010	4	6	0.503	12043	5879
2006	2011	5	6	0.511	11844	5784
2006	2012	6	6	0.465	10891	5388
2006	2009	3	7	0.399	11785	5513
2006	2010	4	7	0.436	11810	5698
2006	2011	5	7	0.449	11305	5601
2006	2012	6	7	0.415	10105	5269
2006	2009	3	8	0.380	12606	5425
2006	2010	4	8	0.412	12186	5702
2006	2011	5	8	0.428	11409	5513
2006	2012	6	8	0.399	11181	5262
2006	2009	3	9	0.309	12427	5394
2006	2010	4	9	0.350	12042	5726
2006	2011	5	9	0.349	11313	5620
2006	2012	6	9	0.315	10629	5283
2006	2009	3	10	0.198	13764	5363
2006	2010	4	10	0.220	13000	5446
2006	2011	5	10	0.228	11715	5459
2006	2012	6	10	0.206	11666	5206

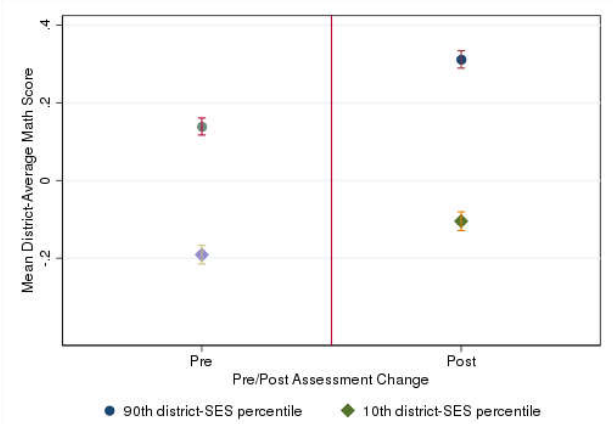
perfrl: percentage of students on free or reduced-price lunch in district
ppe_tot: Total per-pupil expenditure - Total expenditure/enrolment
ppe_inst: Current per-pupil expenditure - Instructional expenditure/enrolment

Figure 1. Mean district scores at 25th and 75th (Panels A and C) and 10th and 90th (Panels B and D) district-SES percentile within Texas before and after switch from TAKS to STAAR, by subject

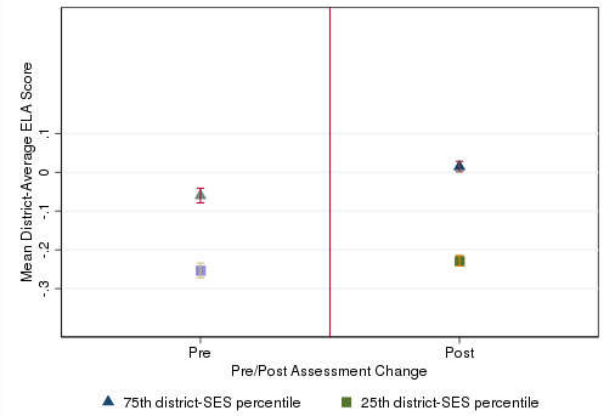
Panel A. Mean math score at 75th and 25th district-SES percentile



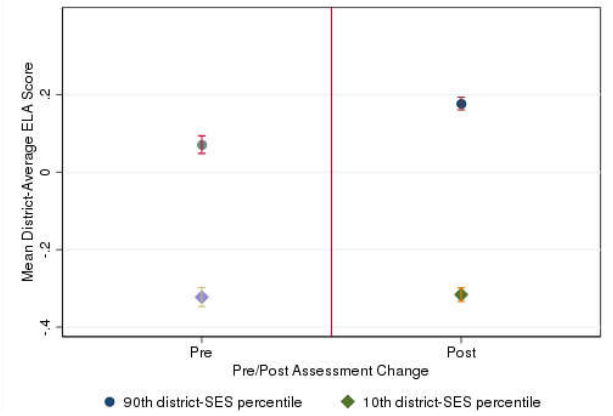
Panel B. Mean math score at 90th and 10th district-SES percentile



Panel C. Mean reading score at 75th and 25th district-SES percentile



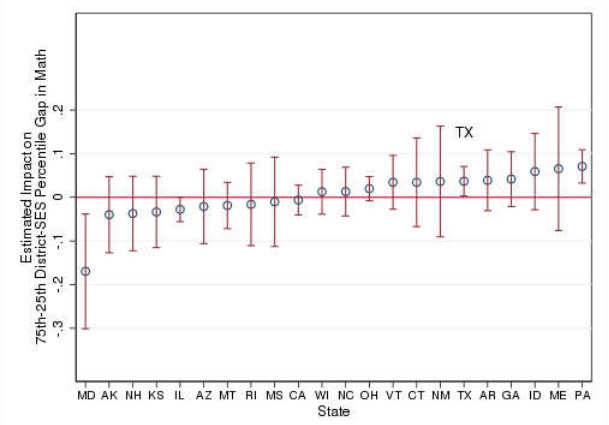
Panel D. Mean reading score at 90th and 10th district-SES percentile



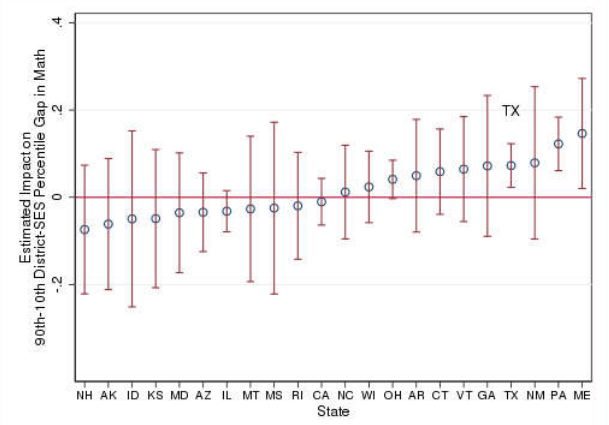
Note: Estimates and 95% confidence intervals obtained from fitting equation (1) using cohort fixed effects and evaluating the average pre/post scores for the relevant district-SES percentiles within Texas. Standard errors are obtained via a test of linear combination of the estimators. Models fitted without any covariates. As explained in text, pre-treatment math outcome is taken as average of outcomes for years 2009-2011. Pre-treatment reading outcome is taken only for year 2011. Post-treatment outcome is taken only for year 2012 for math and reading respectively.

Figure 2. Impact of switch from TAKS to STAAR on 75th-25th and 90th-10th district-SES percentile gap in Texas relative to comparison states for math and reading

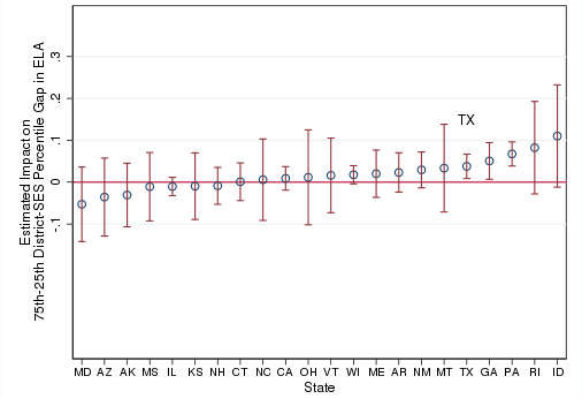
Panel A. Math 75th-25th district-SES percentile gap



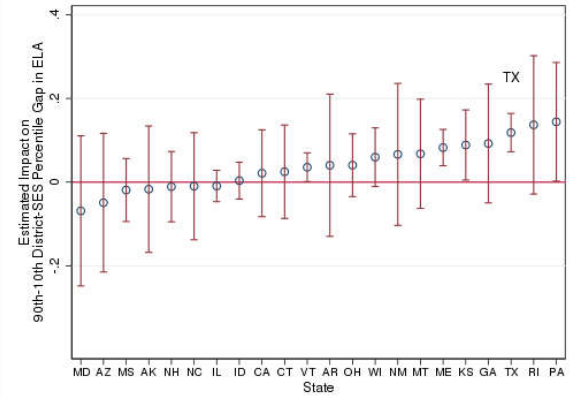
Panel B. Math 90th-10th district-SES percentile gap



Panel C. Reading 75th-25th district-SES percentile gap



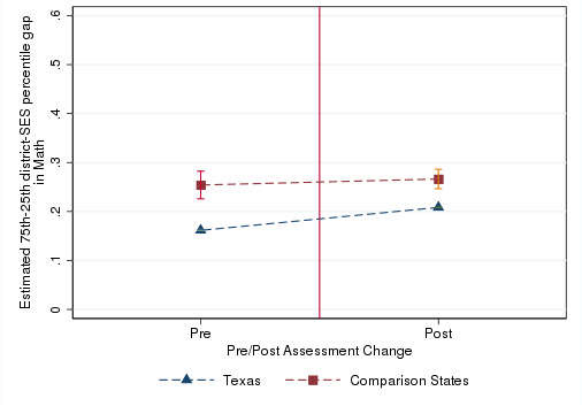
Panel D. Reading 90th-10th district-SES percentile gap



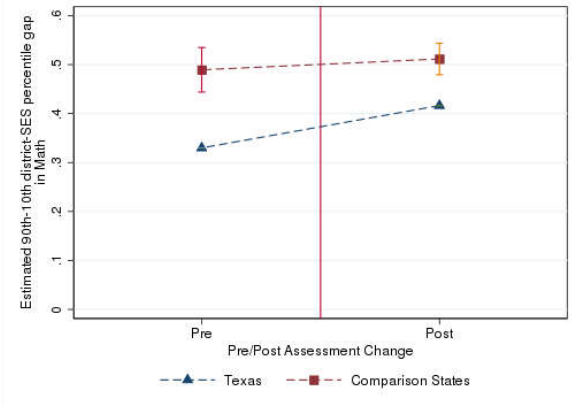
Note: Difference-in-difference estimates and 95% confidence intervals obtained from fitting equation (1) separately for each state using cohort fixed effects and evaluating the difference in average scores at the district-SES values corresponding to the 75th and 25th (or 90th and 10th) district-SES percentile in Texas. Standard errors for the 75th-25th (or 90th-10th) district-SES percentile difference are obtained via a test of linear combination of the estimators. Models fitted with district covariates. As explained in the main text, the pre-treatment outcome is taken as the average of district-level outcomes in years 2009-2011 for math; and in year 2011 only for reading. Post-treatment outcome is taken only for year 2012 for math and reading respectively.

Figure 3. Pre/Post 75th-25th and 90th-10th district-SES score gaps in Texas versus comparison states, by subject

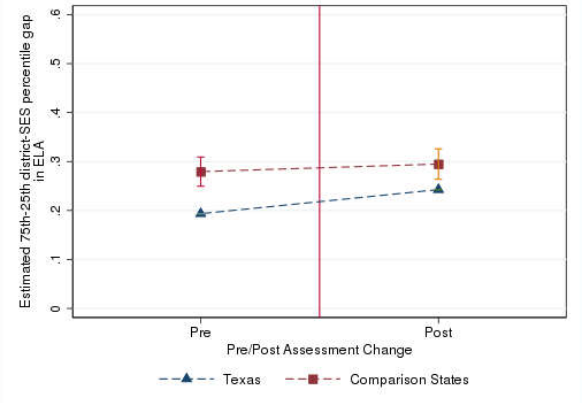
Panel A. Math 75th-25th district-SES percentile gap



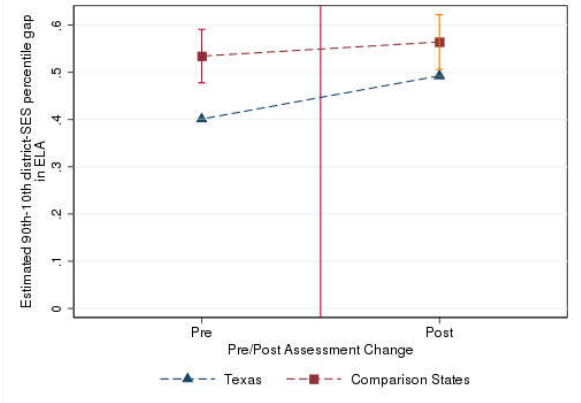
Panel B. Math 90th-10th district-SES percentile gap



Panel C. Reading 75th-25th district-SES percentile gap



Panel D. Reading 90th-10th district-SES percentile gap



Note: Estimates and 95% confidence intervals obtained from fitting equation (2) (see text) using cohort fixed effects and evaluating the average pre/post scores corresponding to the relevant district-SES percentiles within Texas. Standard errors are obtained via a test of linear combination of the estimators. Models fitted without any covariates. As explained in the main text, the pre-treatment outcome is taken as the average of district-level outcomes in years 2009-2011 for math; and in year 2011 only for reading. Outcome for post-treatment includes only year 2012 for math and reading respectively.

Appendix A
Cohorts Used in Analyses

Table A1. Cohorts and year-grades used in analyses, by subject

	School Year			
Cohort[1]	**2008-2009**	**2009-2010**	**2010-2011**	**2011-2012**
Math[2]				
2006	Grade 3	Grade 4	Grade 5	Grade 6
2007	X	Grade 3	Grade 4	Grade 5
2008	X	X	Grade 3	Grade 4
Reading[3]				
2004	X	X	Grade 7	Grade 8
2005	X	X	Grade 6	Grade 7
2006	X	X	Grade 5	Grade 6
2007	X	X	Grade 4	Grade 5
2008	X	X	Grade 3	Grade 4

Note: X denotes scores not included in analyses.
[1] Cohorts are named for the year in which students enter 1st grade in the Fall.
[2] The SEDA dataset excludes data for grade 7 and grade 8 math within Texas because some students take end-of-course exams in these grades, resulting in different assessments within subject-grade-year.
[3] Reading scores from 2009 and 2010 are excluded from analyses due to a change in the TEKS curriculum for reading in school year 2009-2010.

Appendix B
Summary of Assessment Changes in the U.S. States, 2009-2012

Table B1. States chosen as comparison states or if not, reasons for exclusion

State	Comparison State?	Change in Assessment?
Alabama	No	ARMT to ARMT+ (2012)
Alaska	Yes	
Arizona	Yes	
Arkansas	Yes	
California	Yes	
Colorado	No	CSAP to TCAP (2012)
Connecticut	Yes	
Delaware	No	DSTP (paper) to DCAS (online) (2011)
Florida	No	FCAT to FCAT 2.0 (2010)
Georgia	Yes	
Hawaii	No	Online adaptive testing (2011)
Idaho	Yes	
Illinois	Yes	
Indiana	No	Changed cut score (2009)
Iowa	No	ITBS to Iowa Assessments (2011)
Kansas	Yes	
Kentucky	No	KCCT to K-PREP (2012)
Louisiana	No	LEAP administered over two seatings (2010 and 2011)
Maine	Yes	
Maryland	Yes	
Massachusetts	No	Transition from 2000/2004 standards to 2011 standards (2012 and 2013 onwards)
Michigan	No	New cut scores (Fall 2011)
Minnesota	No	MCA-II to MCA-III Math (2010); MCA-II to MCA-III Reading (2012)
Mississippi	Yes	
Missouri	No	Re-administered previous form of grade-level MAP for budgetary reasons
Montana	Yes	
Nebraska	No	School-based student assessments to NeSA (Reading - 2010; Math - 2011)
Nevada	No	New cut scores (Math - 2010; Reading - 2011)
New Hampshire	Yes	
New Jersey	No	New tests for NJ ASK introduced over 2008 to 2009 for grades 3-8
New Mexico	Yes	New school accountability system (2012)[1]
New York	No	Changed cut score (2010)
North Carolina	Yes	
North Dakota	No	Contractor change in implementation of state assessments 2011
Ohio	Yes	
Oklahoma	No	Switched from paper- to computer-based administration (2011 to 2012)
Oregon	No	Changed cut score (Math - 2011; Reading - 2012)
Pennsylvania	Yes	Classroom-based diagnostic tests offered in 2010 as a resource for Pennsylvania System of State Assessment[2]
Rhode Island	Yes	

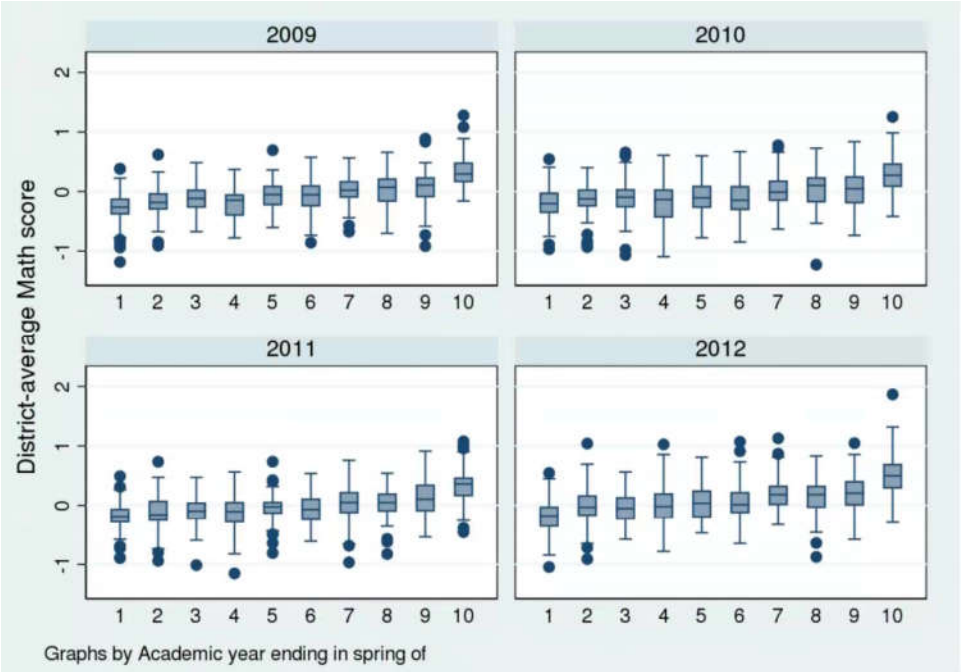
State	Comparison State?	Change in Assessment?
South Carolina	No	PACT to PASS (2009)
South Dakota	No	New content standards assessed (Reading - 2009); Curriculum realignment to new content standards (Math - academic year 2011-2012)
Tennessee	No	Changed cut score (2009)
Texas	**Treatment State**	**TAKS to STAAR (2012)**
Utah	No	New test designs and standards (over 2008 and 2009)
Vermont	Yes	
Virginia	No	New content standards assessed (Reading - 2012; Math - 2011)
Washington	No	WASL to MSP (2011)
West Virginia	No	Changed scale score (2010)
Wisconsin	Yes	
Wyoming	No	Technical problems in administration of computer-based tests (2010); Paper-based tests (2011)

Note: Unless otherwise specified, year refers to spring of stated year
[1] Since the assessment did not change, New Mexico was retained as a comparison state.
[2] Since the tests were offered as a diagnostic tool with no change to the state assessment, Pennsylvania was retained as a comparison state.
Data obtained through internet search of EdFacts reports, state assessment technical manuals, department of education announcements, presentation slides, and newspaper articles. References available upon request.

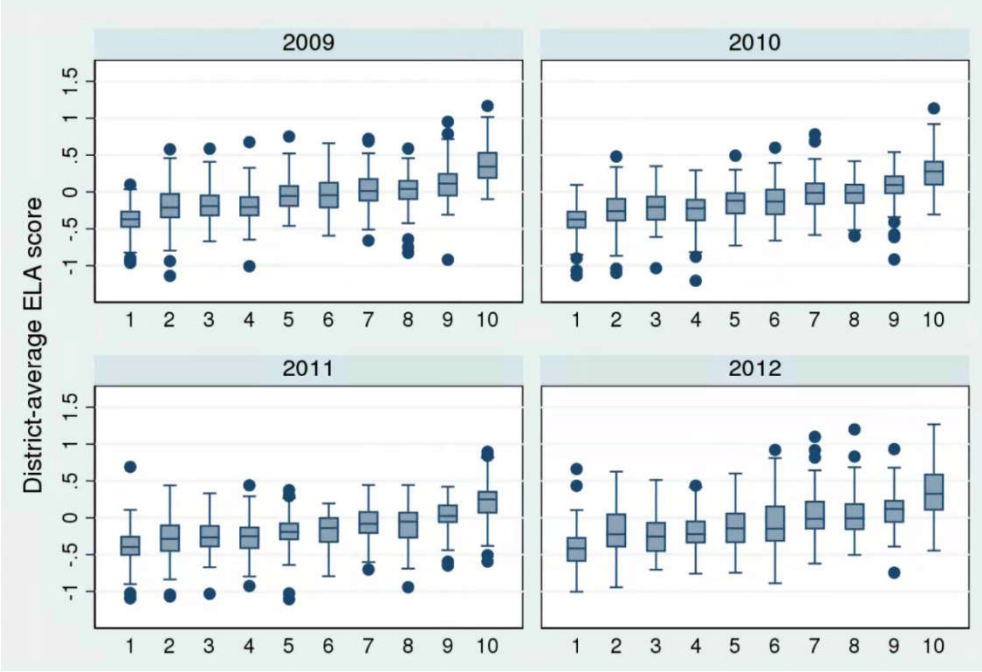
Appendix C
Relationship between District-Average Test Scores and District-SES Deciles

Figure C1. Boxplots of district-average test scores over deciles of district-SES by year for cohort 2006 within Texas.

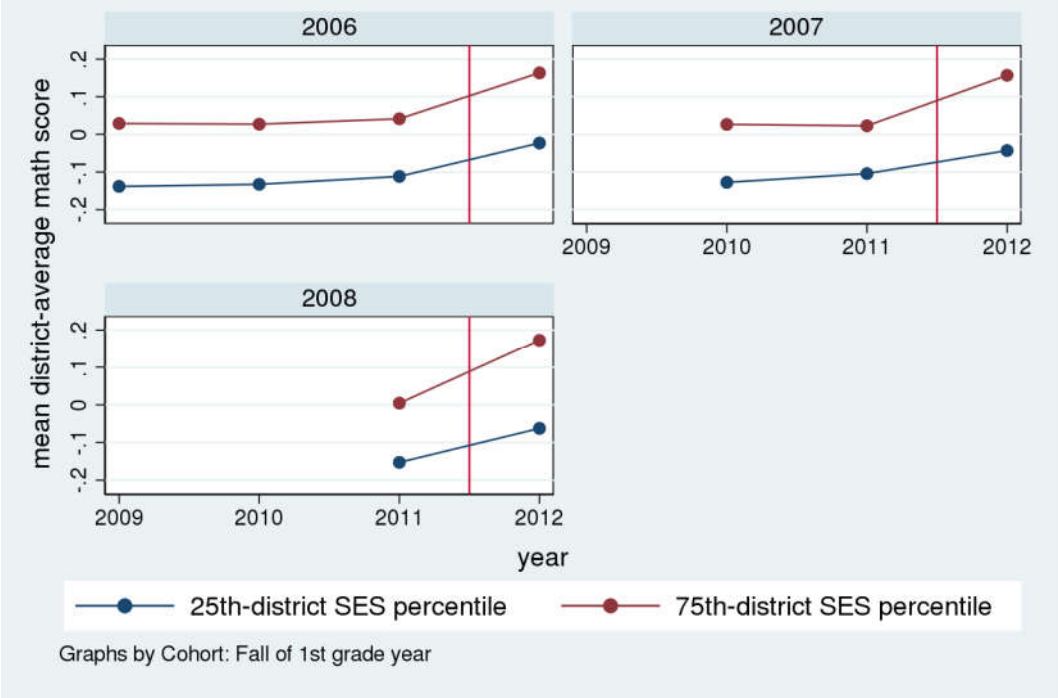
Panel A. Math



Panel B. Reading



Appendix D
75th-25th District-SES Performance Gaps Before and After Switch from TAKS to STAAR in Texas, by Cohort for Math



Note: To obtain the datapoints, I first estimate the relationship between district-average math scores (*Score*) and district-SES (*SES*)[52] for each year separately using the model:

$$Score_d = \beta_0 + \beta_1 SES_d + \beta_2 SES_d^2 + \beta_3 SES_d^3 + \varepsilon_d$$

I then calculated the mean district-average math score at the 25th and 75th district-SES percentiles within Texas separately for each year and plotted them on the graph.

[52] See main text for explanation of why a cubic polynomial function is used.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). When should you adjust standard errors for clustering? (No. w24003). Cambridge, MA: National Bureau of Economic Research.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249-275.

Clark, C. (2011). Testing, testing. Texas standardized exam moves from TAKS to STAAR. *Texas Lone Star*, 18-21. Retrieved from https://www.mytexaspublicschool.org/documents/april-may2012-testing.aspx

Fahle, E. M., Shear, B. R., Kalogrides, D., Reardon, S. F., DiSalvo, R., & Ho, A. D. (2018). Stanford Education Data Archive: Technical Documentation (Version 2.1). Retrieved from http://purl.stanford.edu/db586ns4974

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved from https://epaa.asu.edu/ojs/article/view/432/828

Ho, A. D. (2007). Discrepancies between Score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice*, *26*(4), 11-20.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*(2), 201-228.

Ho, A. D., & Haertel, E. H. (2006). Metric-free measures of test score trends and gaps with policy-relevant examples. CSE Report 665. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489-517.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5), 761-796.

Jennings, J., & Sohn, H. (2014). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. *Sociology of Education*, *87*(2), 125-141.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND Corporation. Retrieved from: https://www.rand.org/pubs/issue_papers/IP202/index2.html

Kaushal, N., Magnuson, K., & Waldfogel, J. (2011). How is family income related to investments in children's learning? In G. J. Duncan and R. J. Murnane (Eds.), *Whither Opportunity?* (pp. 187-206). New York, NY: Russell Sage Foundation.

Koretz, D. M. (2008). *Measuring up*. Cambridge, MA: Harvard University Press.

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System* (*KIRIS*) (MR-1014-EDU). Santa Monica. CA: RAND.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531-578). Westport, CT: American Council on Education/Praeger Publishers.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing: Preliminary evidence about generalization across tests. In R. L. Linn (Chair), *The effects of high stakes testing*. Symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, *29*(3), 426-450.

Lazear, E. P. (2006). Speeding, terrorism, and teaching to the test. *The Quarterly Journal of Economics*, *121*(3), 1029-1061.

McNeil, L. (2000). *Contradictions of school reform: The educational costs of standardized testing*. New York, NY: Routledge.

McNeil, L. & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers?* (pp. 127-150). New York, NY: The Century Foundation.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263-283.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage*. Cambridge, MA: Harvard Education Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733-755). Westport, CT: American Council on Education/Praeger Publishers.

Reardon, S.F. (2017). Educational opportunity in early and middle childhood: Variation by place and age (CEPA Working Paper No.17-12). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp17-12

Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, *40*(2), 158-189.

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., & DiSalvo, R. (2018). Stanford Education Data Archive (Version 2.1). Retrieved from http://purl.stanford.edu/db586ns4974

Reardon, S.F., Kalogrides, D., & Ho, A. (2017). Linking U.S. School District Test Score Distributions to a Common Scale (CEPA Working Paper No.16-09). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-09

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, *42*(1), 3-45.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy*, *5*(2), 251-81.

Schmeiser, D. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: American Council on Education/Praeger Publishers.

Somers, M. A., Zhu, P., Jacob, R., & Bloom, H. (2013). The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation. *MDRC*.

Texas Education Agency (2008). *TAKS information booklet mathematics grade 3*. Retrieved from https://web.archive.org/web/20080822064906/http://www.tea.state.tx.us/student.assessment/taks/booklets/index.html

Texas Education Agency (2010a). *House Bill 3 transition plan. A report to the 82nd Texas Legislature from the Texas Education Agency.* Retrieved from https://tea.texas.gov/student.assessment/hb3plan/

Texas Education Agency (2010b). *Mathematics blueprint*. Retrieved from https://tea.texas.gov/Student_Testing_and_Accountability/Testing/State_of_Texas_Assessments_of_Academic_Readiness_(STAAR)/STAAR_Mathematics_Resources_Archive/

Texas Education Agency (2010c). *Test design and setting student performance standards for STAAR Grades 3-8 and STAAR end-of-course*. Retrieved from https://tea.texas.gov/student.assessment/hb3plan/hb3-sec1ch2.pdf

Texas Education Agency (2011). *2011 District and Campus Coordinator Manual.* Retrieved from https://web.archive.org/web/20110220045422/http://www.tea.state.tx.us/student.assessment/manuals/dccm/

Texas Education Agency (2012). *2012 District and Campus Coordinator Manual.* Retrieved from https://web.archive.org/web/20120829142834/http://www.tea.state.tx.us/student.assessment/manuals/dccm/

Texas Education Agency (2013). *STAAR standard setting technical report*. Retrieved from https://tea.texas.gov/student.assessment/staar/performance-standards/

Winters, M. A., Trivitt, J. R., & Greene, J. P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review*, *29*(1), 138-146.

Zyskowski, G. (2016). State assessments: Past, present, and future. Presented at the meeting of Commission on Next Generation Assessments & Accountability on 20 January 2016 in Texas. Retrieved from https://tea.texas.gov/2804commission.aspx