Evaluating English Learner Reclassification Policy Effects across Districts

Joseph R. Cimpian, *New York University*

Karen D. Thompson, *Oregon State University*

Martha B. Makowski, *University of Illinois at Urbana-Champaign*

Published as: Cimpian, J. R., Thompson, K. D., & Makowski, M. (2017). Evaluating English learner reclassification policy effects across districts. *American Educational Research Journal, 54*(S1), 255S-278S.

Corresponding Author:

Joseph Robinson Cimpian

joseph.cimpian@nyu.edu

Note: This research was funded by a National Academy of Education/Spencer Foundation Postdoctoral Fellowship and a University of Illinois Hardie Fellowship awarded to Cimpian, as well as the U.S. Department of Education Institute of Education Sciences grant R305H140072, for which Thompson is the principal investigator. The research presented here does not necessarily reflect the opinions of the funding agencies. The authors thank their state partners, who provided access to data, collaboratively discussed research questions, and provided valuable insights to help interpret research findings. In addition, the authors thank Lupe Díaz for data management and research assistance, and Kathryn Ciechanowski, Andrei Cimpian, Kenji Hakuta, Stacey Lee, Sean Reardon, and Ilana Umansky for helpful feedback.

Abstract

Effectively educating the large English learner population requires policymakers to ensure developmentally appropriate settings and services throughout the time students are learning English, as well as during their transition to fluent English proficient status—a process termed reclassification. Using longitudinal student-level data from two US states (*N*=107,549), we implement recent advances in multi-site regression discontinuity designs to assess the effects of reclassification policies across districts. We find that reclassification decisions are heavily influenced by state criteria; however, there is considerable variability across districts in the extent of state-level influence. We also find robust evidence of between-district heterogeneity in the effects of reclassification on subsequent achievement and graduation. We discuss the implications of these findings for reclassification policies and future research on the topic. Looking toward the next century of education research, we discuss ways that multi-site regression discontinuity designs can be combined with qualitative research to enable policymakers and practitioners to better understand variation in effects of policies across contexts as well as the mechanisms underlying those effects.

Evaluating English learner reclassification policy effects across districts

By 2050, over one-third of school-aged children in the U.S are projected to be immigrants or the children of immigrants, up from 23% in 2005 (Pew Research Center, 2008). Many of these children will be labeled English learners (ELs). Effectively educating this population requires policymakers to ensure developmentally appropriate settings and services throughout the time students are in the process of learning English. However, recent research has suggested that the transition from EL to "Fluent English Proficient" (FEP)—a process known as reclassification, which often shifts settings and curricula for the student (Estrada, 2014; Linquanti, 2001; Parrish et al., 2006)—can result in academic disruptions to achievement trajectories and graduation (Estrada & Wang, 2015; Robinson, 2011; Robinson-Cimpian & Thompson, in press; Umansky, 2015).

While much research on ELs has focused on the speed with which they attain English proficiency and are reclassified (e.g., Conger, 2009; Thompson, 2015; Umansky & Reardon, 2014), a growing literature has turned its focus to the effects that reclassification can have on students' academic success (e.g., Callahan et al., 2009; Robinson, 2011; Robinson-Cimpian & Thompson, in press). These studies have focused particular attention on the influential role that policymakers have on the reclassification process through establishing test-based thresholds that students must attain in order to be eligible for reclassification.

Although reclassification criteria vary across states, across districts within a state, and even within districts, they consistently include a determination of whether an EL student is achieving at a pre-specified level on an assessment (or set of assessments) determined by policymakers (Linquanti, 2001; Linquanti & Cook, 2015). Because reclassification often entails

a change in instructional services and settings, prior work has argued that it is fundamentally important that the switch to the reclassified-FEP (R-FEP) setting does not occur when the student still receives added benefit from the EL setting. At the same time, the change to the R-FEP setting should not be delayed when a student is no longer benefiting from the EL setting. Thus, policymakers should set reclassification criteria at the point in the student's English language proficiency development when the student is able to "successfully achieve" in mainstream classroom settings (ESEA, s.9101(25)) and when the transition between settings is smooth and does not result in academic disruptions.

To test the smoothness of transitions induced by attaining policy-specified thresholds, a recent wave of research has applied a quasi-experimental technique known as regression discontinuity designs (RDD; e.g., Estrada & Wang, 2015; Robinson, 2011; Robinson-Cimpian & Thompson, in press; Umansky, 2015). These studies utilize large amounts of data to compare the outcomes of students who just barely attained the reclassification criteria with those who just barely failed to attain it—that is, compare nearly identical populations under different educational conditions—and thus provide estimates of the causal effects of reclassification at the policy threshold. To date, each of these studies has examined effects within a single school district.

The recent movement toward uniform criteria within a state (Every Student Succeeds Act, 2015) and toward common English language proficiency assessments across states (ELPA21, 2015; WIDA, 2015) presents new opportunities to use RDD methods to provide rigorous policy recommendations at the state level and beyond. Fortunately, we are at an advantageous point in EL education policy research where increased access to large-scale administrative data and recent methodological advances in the analysis of multi-site RDDs

(Bloom & Weiland, 2015; Raudenbush, Reardon, & Nomi, 2012) allow us to address questions around implementation and policy effects that arise when adopting common reclassification criteria.

Here we extend the literature using RDDs to evaluate district EL reclassification policies by examining statewide effects. Specifically, we examine: (1) the average effects of reclassification on later achievement and graduation across districts, and (2) between-district variability in these effects. Using longitudinal student-level data from two US states—one in the Southeast and one in the Northwest—with different types of quasi-uniform criteria, we examine the degree of policy adherence across districts within each state, finding a tremendous amount of variability. We also find remarkable between-district variability in the effects of reclassification on subsequent achievement and graduation. In addition to discussing the implications of the empirical results for the two states under study, we conclude with a discussion of the strengths and limitations of this quantitative analysis. We discuss results of supplementary qualitative analysis to better understand variability in reclassification effects across districts and describe a more in-depth qualitative study now underway, highlighting how mixed-methods approaches that incorporate causal analysis of effects and mechanisms with in-depth qualitative data about services and placement practices will be necessary in the next century of education research, in the area of EL policy and beyond.

Background

EL classification: A process marked by variation

State and district policies vary enormously with respect to identifying (Linquanti & Cook, 2013), serving (Estrada, 2014; Hopkins, Lowenhaupt, & Sweet, 2015), and reclassifying students in the process of acquiring English (Linquanti & Cook, 2015). We focus here on

variation in reclassification processes. The No Child Left Behind Act (2001) required that states assess ELs' English proficiency every year and establish English-proficient performance standards on these assessments. There is variation in ELP assessments across states, though as of 2015, 44 states are members of two major ELP assessment consortia, with the member states of each consortium administering the same ELP assessment (ELPA21, 2015; WIDA, 2015). When students attain the English-proficient performance standard on their state ELP assessment, students may become eligible for reclassification. However, most states and districts established additional criteria (e.g., content-area achievement test, student grades) that students must meet in order to actually be reclassified (Linquanti & Cook, 2015), complicating comparisons across states and districts. Numerous studies have documented that reclassification criteria vary across districts within states as well (e.g., Hill, Weston, & Hayes, 2014).

The two states we use for our analyses—hereafter referred to as State A and State B—illustrate the variability in reclassification criteria *across* states, as well as the potential created by state policies for variability *within* states. For instance, the policy in State A permits multiple pathways to reclassification by allowing decisions to be based on at least two of five different criteria, which include attaining the state threshold on an ELA content-area assessment (specifically, Level 3 out of 5) or a score of "Proficient" on an ELP assessment (Level 3 of 5).[1] State A's policy also allows parents, teachers, and administrators to reclassify a student before

[1] The ELP assessment data we received from State A is sparsely populated and contains values outside the possible range. Thus, after discussions with the state agency, we deemed these ELP data unreliable (perhaps because the data used were from the first year of a new statewide ELP assessment). Therefore, we are only able to use the state ELA assessment as the test-based exit criteria for our analyses. This is a potential limitation, particularly since researchers urge that ELP assessments should serve as the primary reclassification criteria (Linquanti & Cook, 2015). Nonetheless, because students were able to exit by attaining the ELA assessment criterion (regardless of their ELP scores) in State A during the time period under study, it is important to evaluate reclassification effects for students reclassified on the basis of this criterion.

attaining the exit criteria or to retain a student as EL if the student attains the exit criteria. Thus, there is a great amount of flexibility in State A, both in terms of the choice of test-based exit criteria and in terms of the allowances provided to school personnel and parents for overriding the test-based placement for students.

By contrast, in State B attaining a score of "Advanced" (Level 5 of 5) on the state ELP assessment is the lone required criteria, although the policy permits districts to consider additional criteria. Analysis of EL plans submitted by districts to the state department of education shows variation in the way the state reclassification policy has been implemented. Additional criteria considered in some districts include other language proficiency measures, teacher recommendations, classroom work-samples, and parent input. In some districts, all students who attain the threshold are automatically exited. In other districts, most students who attain the threshold are exited but teachers have the option of requiring that additional district-specific evidence be considered as a factor for some students. And in a small number of districts, all students who meet the ELP threshold must also meet other locally-defined criteria.

Effects of reclassification and the surprising desirability of precise null effects

Because of the change in services often associated with reclassification, it is important to study its effects on other outcomes, such as achievement and graduation. Studying these effects, though, can prove difficult due to the inherent complexity and variability in reclassification procedures. Thus, when examining the literature on the relationship between reclassification and academic outcomes, it is important to consider how the methods used isolate the effects of reclassification from other factors.

One strand of research on the relationship between reclassification and later outcomes has simply compared students who have been reclassified to students who remain ELs using standard

regression frameworks and has concluded that reclassified students have higher levels of academic performance than students who remain ELs (Flores et al., 2009; Hill et al., 2014). However, students who are reclassified may differ in many ways from students who are not, and therefore this research cannot isolate the causal impact of reclassification itself.

Another strand of research has used quasi-experimental methods to analyze the causal impact of reclassification in ways that may eliminate the selection bias inherent in standard regression approaches. When exploring the effects of reclassification on later outcomes using quasi-experimental methods, researchers have found positive, negative, and null effects of reclassification, depending on the outcome and context, including the grade level and the reclassification criteria in place during a particular time period (Callahan et al., 2009, 2010; Estrada & Wang, 2015; Robinson, 2011; Robinson-Cimpian & Thompson, in press; Umansky, 2015). Because numerous factors other than the state-established test thresholds influence reclassification decisions (e.g., teacher judgment, work samples, parent input), studies using regression discontinuity designs (RDDs) have been used recently to disentangle the effects of threshold-induced reclassification from these other factors, thus isolating the policy effects.

It is important to note that reclassification effects, unlike treatment effects in other fields such as medicine, are not necessarily expected or desired; in fact, evidence of effects—whether positive or negative—does not connote a desired outcome, but rather may reflect a misalignment between the bundles of services and settings provided to ELs and R-FEPs at the policy threshold (Robinson, 2011). Perhaps counter-intuitively, RDD studies that yield precisely measured *null* effects of reclassification are desirable because they suggest that students in both the EL setting and the R-FEP setting are performing equally well at the policy threshold, and thus the available settings and services for students and the reclassification threshold are appropriately aligned.

Negative effects of reclassification suggest that students are exiting EL services before they are able to succeed in mainstream classroom settings. One way to address this would be to shift the threshold higher to reclassify fewer students so that students would remain in the more beneficial EL setting for longer. On the other hand, positive effects suggest students remained ELs past the point at which they were able to succeed in mainstream classroom settings without additional support, which policymakers could remedy by shifting the threshold lower to reclassify students sooner. As with any RDD, the analysis is testing the relationship between the threshold and the instruction/services provided. Importantly, no threshold is inherently universally too high or too low. Rather the threshold must also be considered in the context of the available instructional opportunities for students on either side of the threshold, wherever it may be. Thus, alternative policy solutions for addressing reclassification effects may involve realigning services or providing additional resources to some students, as we discuss in greater detail later.

Data

Data for this study come from two states, referred to as State A and State B. The samples for the main analyses are restricted to only those students within one standard deviation of the reclassification testing criteria in order to lessen reliance on observations far from the threshold, (though we vary this inclusion criterion for robustness checks), and to districts with at least two schools per gradeband, at least five ELs per school, and at least 10 students on either side of the threshold; these restrictions were included to help ensure reliable RDD estimates for each district (discussed in the next section). The sample for State A consists of all students who were at some point considered EL between the 2007-08 and 2009-10 school years, were enrolled in grades 4 through 9 at the time a reclassification decision was made, had complete demographic data, and had three years of complete test score data for either ELA or math (N=31,088). For analytic

purposes, State A's sample was partitioned into two gradebands: grades 4-5 and 6-9. The sample for State B includes students who were at some point classified as EL between the 2007-08 and 2011-12 academic years, were enrolled in grades 3-11 at the time a reclassification decision was made, had complete demographic data, and had two years of complete test score data for either ELA or math (N=65,243). For analytic purposes, students from State B were partitioned into three gradebands: grades 3-5, 6-8, and 9-11. Descriptive statistics for the main analytic sample for each state and gradeband are presented in Table 1. Notably, in both states and all gradebands, the proportion of Latino students is 75% or greater. This is similar to national data, in which over 70% of English learners are Spanish speakers. Data for State B also include 4-year graduation outcomes for a subset of the sample (see the final column of Table 1 for descriptive statistics on this sample; N=11,218). In supplementary analyses to better understand the heterogeneity of reclassification effects in State B, we draw on additional data, including: EL plans outlining reclassification practices in each district; statewide course-taking data about enrollment in English Language Development (ELD) courses; and conversations with practitioners.

Methods

To isolate the effects of the state policy from confounding factors, we employ several statistical techniques. We first describe the model for predicting the likelihood of reclassification and then discuss the model for examining the effects of reclassification on academic outcomes. Our model builds on previous work that examined effects within a single district (Robinson, 2011; Robinson-Cimpian & Thompson, in press; Umansky, 2015), but extends those models to examine statewide average effects and heterogeneity of effects across districts (building on a recent suggestion by Bloom & Weiland, 2015, in the context of Head Start evaluations, to conduct multi-site RDDs using a meta-analytic approach).

Estimating the statewide average policy-induced increase in reclassification likelihood and between-district variance

Recall that the decision to reclassify a student is based on whether the student attained the policy-based test-score criteria, teacher judgment, and possibly other criteria that can vary across and even within districts. Thus, the only known, directly observable, exogenously determined factor influencing reclassification decisions is whether the student attained the policy-specified test-score criteria. To separate the influence of the policy from other factors, we use a regression discontinuity design (see Imbens & Lemieux, 2008), where we predict reclassification status (R) for student i in district j as a function of the test score (T) from the prior year used for determining eligibility (linear and quadratic terms), an indicator for whether she met the policy-based threshold on that test (C), interactions between C and all T terms, grade-by-year fixed effects (ψ_{gy}), and (in some models) a vector **X** of demographic covariates (e.g., gender, race, age, length of time spent as EL, special education status, free and reduced-price lunch status). Because the extent to which attaining the threshold affects reclassification decisions likely varies across districts, our model estimates the relationships of the prior test score and threshold attainment to reclassification likelihood for *each district* uniquely but simultaneously. To do this, we create a series of variables that indicate whether a student attends school in a specific district d. If district j for a given student matches district d, then student i's value of $I_{idj} = 1$; otherwise (i.e., if $d \neq j$), $I_{idj} = 0$ (see Bloom & Weiland, 2015, for a similar approach). This series of D indicators (where D is the number of districts in the analysis) effectively switches on and off district-specific RDDs depending on whether the student is in that district:

$$R_{ij} = \sum_{d=1}^{D} \Big(I_{idj} \times \Big(\psi_{0d} + \psi_{1d} C_{ij} + \psi_{2d} T_{ij} + \psi_{3d} T_{ij}^2 + \psi_{4d} \big(T_{ij} \times C_{ij} \big) + \psi_{5d} \big(T_{ij}^2 \times C_{ij} \big) \Big) \Big) + \psi_{gy} \big[+ \mathbf{\psi X_{ij}} \big] + \varepsilon_{ij} \quad (1)$$

In the above equation, ψ_{1d} is the average increase in reclassification likelihood associated with just barely attaining the policy-based test-score criteria for students in a specific district. Equation 1 yields a total of D estimates of ψ_{1d} (i.e., one estimate per district). To obtain the *statewide average* increase in reclassification likelihood due to attaining the policy threshold (ψ_1), we estimate the following random-effects meta-analysis (Hedges & Olkin, 1985), where ν_{1d} is the district-specific deviation from the state mean:

$$\psi_1 = \psi_{1d} + \nu_{1d} \quad (2)$$

This meta-analytic approach also yields an estimate of the *between-district variance* in the compliance[3] of districts with the statewide policy. That is, we not only estimate the average policy effects on reclassification likelihood, but we also estimate how much the policy effects vary across the districts in each state. This random-effect variation is presented in standard deviation units and referred to a τ_1 (the variance is τ_1^2).

Importantly, this process plausibly isolates the district-specific and statewide average effects of attaining the policy-based threshold on reclassification likelihood from all other factors that might affect reclassification decisions (e.g., student motivation, teacher perceptions, familial support). This is because our estimates are driven by the outcome differences between students who *barely attained* and who *barely failed to attain* the state-specified threshold; and we assume that both observed (e.g., special education status, gender) and unobserved (e.g., motivation)

[3] Note that "compliance" is a technical term used when discussing regression discontinuity designs where attaining the thresholds predicts but does not determine treatment status (here, reclassification).

factors are equally distributed between these two sets of students. Appendix Table A demonstrates that this assumption is satisfied for the statewide average for most observed variables. However, nearly all of these observed variables exhibited significant heterogeneity in the differences across the districts. To ensure that the effect heterogeneity (e.g., in subsequent achievement and graduation; discussed later) was not simply the byproduct of covariate heterogeneity, we assessed the extent to which covariate heterogeneity was significantly correlated with effect heterogeneity. Our supplemental analysis, which used a Benjamini-Hochberg (1995) correction for multiple testing, found no statistically significant correlations. Thus, there is little to no evidence for concern that the primary results reported in this paper are due to either mean differences or systematic variation across districts in the covariates above and below the state policy threshold; rather, the observed effects are likely due to between-district differences in the instruction and services students received before and after reclassification.

Estimating the statewide average effect and between-district variance of policy-induced reclassification on student achievement scores and graduation

Thus far, we have only discussed analyses related to the extent to which the statewide policy is adhered to across a state. We now turn to estimating the effects of that adherence—that is, of reclassifying students because they attained the test-based criteria—on achievement and graduation. To estimate these effects, we use an instrumental variables approach (Angrist, Imbens & Rubin, 1996), similar to the approach implemented by Robinson (2011) in studying policy-induced reclassification effects. This requires two stages. First, we estimate the predicted likelihood of reclassification for each student using the estimated coefficients from Equation 1.[4]

[4] Raudenbush et al. (2012) note that the analytic approach we employed here (i.e., estimating independent RDDs for each district, albeit done simultaneously) requires the assumption that attaining the threshold influences treatment (here, reclassification likelihood) in *each* district. As

In the second stage, the predicted value of reclassification ($\hat{R}$) obtained from Equation 1 is then used as a predictor of an academic outcome (Y; e.g., subsequent ELA achievement, graduation) in Equation 3:

$$Y_{ij} = \sum_{d=1}^{D} \Big(I_{idj} \times \Big(\phi_{0d} + \phi_{1d}\hat{R}_{ij} + \phi_{2d}T_{ij} + \phi_{3d}T_{ij}^2 + \phi_{4d}\big(T_{ij} \times C_{ij}\big) + \phi_{5d}\big(T_{ij}^2 \times C_{ij}\big) \Big) \Big) + \phi_{gy}\big[+\mathbf{\phi X_{ij}}\big] + \omega_{ij} \quad (3)$$

By conditioning on all other covariates in Equation 3, the coefficients on $\hat{R}$ (i.e., the ϕ_{1d}s, one for each of the D districts) yield the district-specific effect of reclassification due to attaining the state-policy-based threshold. This is the district-specific average treatment effect for the group of students who would be reclassified if they attained the threshold but would not be reclassified if they failed to attain the threshold. These individuals are referred to as "compliers" in the instrumental-variables literature (Angrist et al., 1996) because their treatment status (i.e., reclassification) complies with the assignment associated with their threshold-passing status. This is the group of students for whom the policy has a direct effect, and is therefore a group of tremendous interest to policymakers who might be considering altering the policy to facilitate improved student outcomes. However, our analyses cannot speak to effects for students whose

---

we found in answering our first set of research questions, on how the policy affects reclassification likelihood, there was tremendous variability across districts with some districts showing no significant increase in likelihood associated with attaining the threshold. Thus, we did not include these districts in the analyses for our second set of research questions, on the effects of state-policy-induced reclassification. To satisfy the assumption noted by Raudenbush and colleagues, we included only districts with a highly significant increase in reclassification (i.e., F>10; Stock & Yogo, 2005) in State B; only one district met this high bar for State A, and so we do not estimate effects for State A. Regarding State B, the sample size and districts included decreased with this new restriction; in supplemental analyses, we re-estimated the model using all observations and districts and obtained the same patterns of findings. Thus, the restriction of the first-stage F>10 does not alter the pattern of findings, but it does theoretically reduce bias in the estimates.

reclassification status does not hinge on attaining the policy threshold, such as a student whose teacher would choose to not reclassify her regardless of whether she attains the state threshold, or a district or school site which had a practice of not reclassifying students in the early primary grades. Relatedly, our analyses cannot speak to effects in entire *districts* that do not adhere to the state policy and thus have very few complier students, because the precision is too low and potential for bias too high in such districts. Again, we use random-effects meta-analysis to estimate the average policy-induced reclassification effect in each state, as well as the between-district variability in the effects of reclassification.

For all RDD analyses in the main text, we restrict the sample to only students within 1SD of attaining their state's reclassification criteria to lessen reliance on observations far from the threshold; all results are generally robust to alternative choices of sample inclusion criteria (see Appendix Tables B-D). Further, to ensure sufficient numbers of observations for each district-level RDD and for estimation of the standard errors, only districts with at least two schools (each with at least five EL students) and a district total of 10 EL students on each side of the threshold per analysis were retained. All analyses are clustered at the school level, to account for the nesting of students within schools, and estimate percentile-*t* (i.e., asymmetric) 95% confidence intervals via 999 clustered bootstrapped replications (Cameron & Miller, in press).[5]

Results

[5] Percentile-*t* clustered bootstrapped confidence intervals with 999 replications were necessary due to the small number of clusters (i.e., schools) found within each district. Relying on asymptotic standard errors (and thus, confidence intervals) produces estimates that are artificially precise. To obtain accurate precision, we employed the computationally intensive cluster bootstrap approach and inspected our bootstrap distributions, following the suggestion of Cameron and Miller (in press). When possible (i.e., with OLS but not IV models), we estimated wild cluster bootstrapped standard errors (using the Stata command cgmwildboot, created by Judson Caskey) as well for a random subset of districts and obtained confidence intervals that were remarkably similar to the percentile-*t* ones in all cases, regardless of the district size.

In this section, we present the results of our analyses examining the effects of policies both on the likelihood of reclassification and on the effects of reclassification on subsequent achievement and graduation. For each of these outcomes, we first discuss our estimates of the statewide average effects and then discuss variability across districts within the state.

Policy effects on the likelihood of reclassification

In both states, attaining the policy-based criteria for reclassification increases the likelihood of reclassification, on average. However, attaining the criteria does not uniformly increase the likelihood across the states or grade levels within the states. Elementary school students in State A experience the smallest increases in the likelihood of reclassification. That is, attaining the test-based criteria increases a student's likelihood of reclassification by only 6.9 percentage points (pp; p=.002) according to Model 1 in Table 2. (Later in this section, we discuss why this increase might be so low.) Students in later grades in this state experience a greater increase, 27.0pp, in the likelihood of reclassification when they attain the test-based criteria (p<.001). In State B, the state policy increases the likelihood substantially more, with percentage point increases of 56.8, 56.7, and 56.9, in grades 3-5, 6-8, and 9-11, respectively (all ps<.001). These estimates are relatively robust to the inclusion of additional covariates (e.g., gender, race, time as EL) in Model 2. See Figure 1 for a visual representation.

Although attaining the state policy threshold generally increases the likelihood of reclassification, there is considerable between-district heterogeneity in this dimension. In State A, there is a standard deviation (SD) of 6.7 pp in the policy-induced reclassification rates across the districts in grades 4-5 in Model 1 (p<.001). This implies that attaining the threshold in a district with reclassification rates one SD below the state average does not increase a student's chance of reclassification, whereas attaining the same threshold in a district 1SD above the state

average would increase a student's chance by about 13.6 pp. These differences are even more pronounced in State B, where the between-district SD in reclassification likelihood in grades 9-11 is 42.0 pp ($p<.001$). If we look 1SD above and below the state average of 56.9, this variability highlights that attaining the state-specified threshold does *not* affect reclassification much in some districts (i.e., rates in the low teens) whereas other districts reclassify students *solely* on the basis of the state-specified threshold (i.e., rates of nearly 100%).

Differences between State A and B in state-level reclassification policies may help to explain the above patterns. First, State A has multiple ways in which a student can be reclassified (e.g., one pathway is through attaining a score of "Proficient" on the state ELA exam; a different pathway is through scoring a pre-specified threshold on the state ELP assessment, for which we do not have reliable data). In contrast, State B has only the state ELP test as the state-required criteria. The existence of multiple pathways in State A may explain why the jump in reclassification likelihood for attaining the threshold (for one of several pathways) is relatively low in that state, if districts tend to rely on one or more of the other pathways; this also may explain why we observe measurable variance across districts in reclassification likelihoods in State A. In State B, the extensive variability across districts both in terms of average jumps in reclassification likelihood may be explained by another feature of its policy: Although it requires scoring above a threshold on the ELP assessment, it also permits districts to consider additional factors in making reclassification decisions (e.g., writing samples). These additional factors may present new hurdles to some students who are otherwise reclassification-eligible.

Effects of state-policy-induced reclassification on subsequent achievement and graduation

Given the evidence to suggest that state reclassification policies substantially influence reclassification decisions, albeit to varying degrees across districts, we now turn to the effects of

this policy-induced reclassification on subsequent achievement on both math and ELA tests (Table 3) and on graduation in State B (Figure 2 and Appendix Table D). We do not report effects for State A because too few students were reclassified on the basis of attaining the ELA threshold, which introduces imprecision and potential bias into the effect estimates for that state. Because year-after effects may not show the full picture of how reclassification impacts student outcomes in the long-term, we focus our presentation of results on the effects of reclassification on graduation, which is the longest-term outcome available for analysis in K-12 administrative datasets.

We find little evidence of a statewide average effect on any outcome, suggesting that on average, students in State B near the state threshold perform equally well whether in the EL or R-FEP setting. For instance, just-barely reclassified students are about 2.4 percentage points more likely to graduate high school than just-barely non-reclassified students, but this difference is not statistically significant (p=.55). However, all effects (except at the elementary school level) vary substantially across districts, suggesting that the lack of a substantial average effect in the state does not imply null effects for the individual districts. Continuing the graduation example, the between-district standard deviation is 16 percentage points (p<.001). That is, in some districts, there is a large and significant negative effect of reclassification on graduation (e.g., district 2 in Figure 2), where students just-barely not reclassified are 80 percentage points less likely to graduate (p<.001). By contrast, some districts have large positive effects of reclassification on graduation (e.g., district 29 in Figure 2), where students who are just-barely reclassified are 38 percentage points more likely to graduate than otherwise similar peers who were not reclassified (p=.01).

Given the large amount of between-district heterogeneity both in terms of reclassification

likelihood and in terms of the effects of reclassification, a reasonable question to ask is: Are the districts where reclassification confers benefits more likely to reclassify students who attain the state criteria? To explore this question, we looked at the correlations between the likelihood of reclassification in a district and its reclassification effects. We found no compelling evidence for such a relationship. For example, the correlation between the graduation effect and the likelihood of reclassification was $r(34)=0.07$, $p=.69$. One might suspect, based on this very low correlation, that districts are not considering—or are unaware of—the effects of reclassification in that particular district on students close to the threshold when they make reclassification decisions. Moreover, when looking at the correlations of reclassification likelihood first with the subsequent math score and then again with the subsequent ELA score at each gradeband, we found one correlation to be non-significantly negative and the other to be non-significantly positive. Thus, there was no consistency even in the directionality of the correlations within a gradeband, further suggesting these correlations represent random patterns rather than a careful consideration of effects when making reclassification decisions. Importantly, although these correlations may represent random patterns, the between-district heterogeneity in effects of reclassification on academic outcomes does not appear to be random, as we discuss later when exploring potential mechanisms for these effects.

Discussion

This research is the first to examine between-district variability in threshold-induced reclassification likelihoods and effects of reclassification, and as such, it provides a framework for future studies of this nature. As reclassification criteria become more standardized across districts within states, as stipulated by the new Every Student Succeeds Act (2015), and also across states, similar analyses can be used to identify and learn from educational systems that are

more effective in educating English learners. We found evidence that statewide thresholds influence a student's likelihood of reclassification, but that the magnitude of the impact varies considerably across and within states. Turning to the effects of reclassification on achievement and graduation in one state, we found no evidence of an average effect in the state, but consistent evidence for heterogeneity of effects across districts—with some districts having negative effects, and others having positive ones. We begin this section by discussing potential mechanisms for the effect heterogeneity. We then conclude by discussing the implications of our findings—and of the research approach employed—for EL policies and research practices in the years to come.

Exploring heterogeneity of effects

To explore possible mechanisms underlying the significant variation in reclassification effects across districts, we used three approaches, focusing our analysis on districts with significant positive or negative effects of reclassification on later outcomes. First, to better understand the services students receive when classified as ELs, we reviewed district EL plans submitted to the state department of education. Second, we analyzed a new statewide course-taking dataset to determine the proportion of secondary ELs enrolled in English Language Development classes in particular districts. Third, we engaged in conversations with district and state personnel to gain their insights into possible explanations for the mechanisms underlying the observed effects.

Because district EL plans are compliance documents, all plans outlined specific services ELs receive, such as ELD classes, but provided limited insight into tangible differences in services across districts. Analysis of course-taking data revealed more variation in services. Among the districts with the most consistently significant effects (positive or negative) of

reclassification on graduation, the proportion of secondary ELs enrolled in ELD classes ranged from 51% to 100%.[6] We generally found that the larger the magnitude of the reclassification effect, the higher the proportion of secondary ELs enrolled in ELD classes. For example, in district 33 (in Figure 2), with 94% of secondary ELs enrolled in ELD classes, reclassification is associated with a 67 percentage-point higher likelihood of graduating. If ELD classes in the districts are not rigorous, if enrollment in ELD limited students' ability to earn other credits they needed for graduation, and/or if enrollment in ELD carried with it stigma in the district, that could potentially explain why reclassification was associated with positive outcomes for students.

There are important limitations to this course-taking analysis. First, course-taking data were only available for 2013-14, the year after the other data used for analysis ended. Second, the only type of English-learner-specific course with its own code in the course-taking dataset was ELD. The course-taking data contain no codes to indicate whether particular content-area courses were "sheltered" versions of those courses designed specifically for English learners, so we could not determine the extent to which districts placed ELs in these "sheltered courses," which prior research suggests may be less rigorous than mainstream versions of the courses (Dabach, 2014). Third, courses with equivalent titles and codes can vary widely in rigor and effectiveness, in large part because individual teachers can have substantial impact on student outcomes. In one district with a consistent negative effect of reclassification on later outcomes, a

[6] Federal and state regulations require that ELs receive designated language development instruction, which is typically accomplished through a dedicated ELD class. However, a variety of other models might be used. For example, an ELD teacher might "push in" to a content-area class, providing targeted small group instruction to a cluster of ELs within the larger class setting. The fact that a student is not coded as being enrolled in an ELD class might indicate that the student was receiving language development instruction in some way other than a conventional, stand-alone ELD class period. At the secondary level in our data, 0.01% of non-ELs are enrolled in ELD classes, compared to 72% of ELs.

relatively high proportion (88%) of secondary ELs were enrolled in ELD courses. However, in this district, if the ELD teachers were particularly effective, if ELs had access to rigorous core content courses, and/or if mainstream teachers had limited training in how to meet the needs of recently reclassified students, this could explain why students close to the reclassification threshold seemed to benefit from EL services.

To gain additional insight into heterogeneity of effects, we also engaged in conversations with district and state personnel. For example, we found across all secondary-school outcomes that one district had large positive effects of reclassification (i.e., suggesting it was beneficial to exit EL status in that district). Through discussions with that district's former EL coordinator (and without mentioning anything about the quantitative patterns we observed), we learned that during the majority of our data panel, the ELD teacher at the high school in that district experienced health problems that caused periods of extended absences and ineffective instruction when present. In essence, the district coordinator had a suspicion that students might benefit from exiting EL status in that district given the circumstances—a suspicion that our quantitative analyses corroborate.

Implications for EL policies and practices

The implications of this research for policymaking are many, and yet must be interpreted with the specific context of this study in mind and not broadly generalized. Moreover, interpretation must be accompanied by recognition that the RDD-based estimates might best be viewed as the first step in an evaluation rather than as an end point. Studies such as ours often begin from a policy need to evaluate the threshold that a district or a state has established for reclassification eligibility. However, as mentioned before, assessment thresholds cannot be evaluated in isolation. Rather, RDD-based evaluations analyze how thresholds interact with the

instruction and services available to students on both sides of the threshold to produce effects.

Thus, the multi-site RDD approach used in the current study can provide policymakers with evidence of misalignment between thresholds and instruction/services in the state on average and in specific districts—misalignment that policymakers and educators may or may not otherwise suspect. The precise change(s) needed in the threshold and/or the instruction/services provided to students cannot be determined by the RDD approach alone, but rather must involve more in-depth data collection and professional judgment after the RDD identifies areas of misalignment. Possible strategies for remedying the misalignment might include: (1) either lowering or raising the reclassification threshold (depending on whether a positive or negative reclassification effect was found, respectively); and/or (2) modifying instruction/services for students near the threshold. For example, if a negative effect of reclassification were found, a district might consider providing additional language development support for students after reclassification, perhaps by providing additional professional development for teachers of reclassified students. Of course, modifying instruction/services is a complex endeavor and might involve teacher training, teacher recruitment, and/or changes in the way teachers are assigned to courses (cf. Dabach, 2015). Recent evidence suggests that changing the threshold can significantly alter the reclassification effects in a single district (Robinson-Cimpian & Thompson, in press), but we are not aware of any research examining how changing instruction/services or resources for a threshold-specific subset of students affects changes in reclassification effects.

The present research also informs conversations about whether state policy should establish a single reclassification threshold for all districts, which is now required under the Every Student Succeeds Act (2015). Establishing a single threshold has the benefit of facilitating

comparisons across districts and providing a common metric by which to assess EL status for students who move across district boundaries. Yet, requiring a common threshold across the state restricts the ability of a district to adjust the threshold to meet the needs of its own students given the services the district provides. State B in effect employed a hybrid approach. Because districts in this state could consider additional criteria in reclassification decisions, districts had an indirect means for altering the common criteria, even though our findings show that the state threshold operated as the main gatekeeper criterion. As we discussed earlier, one might suspect that a district would reclassify fewer students who attained the state criterion—perhaps by adding additional criteria—if they suspected negative effects of reclassification in the district; the converse could be true as well. However, the small and inconsistently-signed correlations between reclassification likelihood and effects provided no evidence for a link between the two. Moreover, our examination of district EL plans revealed no clear pattern between various types of additional criteria and reclassification likelihood. Rather, the weak evidence of any links among reclassification likelihood, effects, and additional criteria suggests that the flexibility that districts currently have to raise or lower criteria does not relate to the observed effects, and thus may not be leading to more optimal outcomes for students as currently implemented. We suspect that districts are simply unaware of their reclassification effects, and thus the lack of a relationship between additional criteria at the district level and effects is unintentional.

Despite the complexities involved in centralized and quasi-centralized threshold settings, this and other recent research offers a path forward for EL policymaking and evaluation. Considering the construct-relevance and validity of reclassification criteria is an essential step (Linquati & Cook, 2015), and a variety of techniques can inform establishment of test-based thresholds (Cook, Linquanti, Jung, & Chinnen, 2012; Linquanti & Cook, 2015). Once criteria are

established and implemented, policymakers may wish to conduct multi-site RDDs to examine reclassification effects and likelihoods. By informing districts of their effects, and facilitating exploration of the mechanisms of these effects, leaders can then make more informed decisions. A district with identified threshold/service misalignment might consider making changes to instructional services. A state with consistent patterns of positive effects across districts might consider lowering the threshold. Moving forward, we recommend that policymakers carefully evaluate the effects at both the state level and the district level for any EL reclassification policy they implement, and that they make adjustments as needed in the criteria, services provided, and resources available to struggling students and districts. As EL policies continue to set thresholds at the state level—or as groups of states that belong to assessment consortia consider setting common thresholds, which are even further removed from district control—it will be increasingly important to implement rigorous evaluations to determine where misalignment exists and what resources could provide remedies.

Finally, even though many findings of this study cannot be easily generalized to other contexts, there is one important finding with clear policy implications for other states and districts—namely, the tremendous amount of heterogeneity in the effects of reclassification on later outcomes. This research is the first to examine reclassification effects in multiple districts simultaneously, and the findings revealed a wide array of district-level effects, ranging from large negative effects to large positive effects, even when considering students subject to the same state-level policy threshold. Thus, policymakers and researchers should not default to a belief that reclassification is universally beneficial or detrimental. Further research is needed to understand the circumstances that lead to these varied reclassification effects, as it is clear from the present study that it cannot be traced to threshold placement alone.

Implications for research in the next century

Within education, rigid classification systems are common: A student either does or does not qualify for English learner services. The same is true for special education and for gifted programs, among many types of services. Because such services are often costly and because, in some cases, legal rights to the services are protected under federal law, there is a need for clear classification systems, enabling education agencies to tabulate the number of students requiring particular services and enabling enforcement of laws stipulating the provision of those services. Yet a student who barely attains the criteria for exiting EL services may have needs that are quite similar to the needs of current ELs. Sorting students into discrete categories obscures this situation. In the next century of education research, the tension between the need for rigid classification systems and the need for a continuum of services responsive to individuals' complex and shifting needs is likely to remain an enduring theme.

We see multi-site RDDs, combined with follow-up qualitative analysis about mechanisms for effects, as one important tool for ameliorating the potential pitfalls of rigid classification systems and enabling smooth transitions for students across settings and services. Specifically, the quantitative analysis presented here can serve as a first step to evaluate reclassification policy and practice, identifying if there is something amiss with either reclassification criteria or instructional services in particular districts, as indicated by significant positive or negative effects of reclassification on later outcomes in these districts. If significant effects are found, then additional qualitative data can be gathered about both the criteria and services in the particular districts with significant effects, employing methods such as those used by Estrada (2014) and Kanno & Kangas (2014) to understand course placement and instructional services for current and former ELs within individual schools.

In fact, in State B, we are building on the findings of our present analysis by launching a companion study in which we will partner with five different districts that vary in their reclassification criteria, their likelihood of reclassifying students who attain the ELP threshold, and their effects of reclassification on later outcomes. Through interviews, observations, and follow-up quantitative analysis, we will learn much more about the mechanisms for reclassification effects. This overarching strategy—using sophisticated, multi-site RDD approaches to identify districts with significant positive and negative effects and then following up with qualitative research within those districts to explore the mechanisms for those effects—can be applied in any context in which test-based criteria determine eligibility for services, such as intervention programs or gifted programs.

In addition to mixed methods research that draws on the causal inference aspects of experimental and quasi-experimental designs, we expect that research-practitioner partnerships will be central to addressing critical education issues in the next century of education research and to fully understanding the mechanisms at play. As Gutierrez and Penuel (2014) wrote, "[C]onsequential research on meaningful and equitable educational change requires a focus on persistent problems of practice, examined in their context of development, with attention to ecological resources and constraints, including why, how, and under what conditions programs and policies work" (p. 19). While researchers have long partnered with practitioners in a variety of ways, dedicated grant programs at both the Institute of Education Sciences and the Spencer Foundation now specifically provide support to such partnerships, fostering their stability and facilitating sustained focus. Because our analysis in State B occurred within the context of an ongoing researcher-practitioner partnership, we have been able to build upon our initial quantitative findings and craft the next mixed methods phase of our research in collaboration

with practitioners, with findings from one study immediately shaping the design of the next study. In the next century of education research, we see a vital need to expand researcher-practitioner partnerships to facilitate sustained focus on persistent problems of practice such as reclassification in order to ensure the success of English learners and all students.

References

Angrist, J.D., Imbens, G.W., & Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444-455.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 289-300.

Bloom, H.S., & Weiland, C. (2015). *Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study.* New York, NY: MDRC.

Callahan, R., Wilkinson, L., Muller, C., & Frisco, M. (2009). ESL placement and schools effects on immigrant achievement. *Educational Policy*, *23*(2), 355-384.

Conger, D. (2009). Testing, time limits, and English learners: Does age of school entry affect how quickly students can learn English? *Social Science Research*, *38*(2), 383-396.

Cook, G., Linquanti, R., Chinen, M., & Jung, H. (2012). *National evaluation of Title III implementation supplemental report: Exploring approaches to setting English Language Proficiency performance criteria and monitoring English learner progress*. Washington, DC: U.S. Department of Education.

Dabach, D.B. (2014). "I am not a shelter!": Stigma and social boundaries in teachers' accounts of students' experience in separate "sheltered" English learner classrooms. *Journal of Education for Students Placed at Risk,19*(2), 98-124.

Dabach, D.B. (2015). Teacher placement into immigrant English learner classrooms: Limiting access in comprehensive high schools. *American Educational Research Journal, 52*(2),

243-274.

ELPA21. (2015) *ELLs in ELPA21*. Retrieved from http://www.elpa21.org/standards-initiatives/ells-elpa21

Estrada, P. (2014). English learner curricular streams in four middle schools: Triage in the trenches. *The Urban Review*, *46*(4), 535-573.

Estrada, P. & Wang, H. (2015*). The consequences for access to core curriculum of continuing English learner status in secondary school.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Flores, E., Painter, G., & Pachon, H. (2009). *¿Qué pasa? Are ELL students remaining in English learning classes too long?* Los Angeles, CA: Tomás Rivera Policy Institute.

Gutiérrez, K.D., & Penuel, W.R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher*, *43*(1), 19-23.

Hedges, L.V., & Olkin, I. (1985). *Statistical Method for Meta-analysis*. Orlando, FL: Academic Press.

Hill, L.E., Weston, M. & Hayes, J. (2014). *Reclassification of English learner students in California.* San Francisco, CA: Public Policy Institute of California.

Hopkins, M., Lowenhaupt, R., & Sweet, T.M. (2015). Organizing English learner instruction in new immigrant destinations: District infrastructure and subject-specific school practice. *American Educational Research Journal*, *52*(3), 408-439.

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

Kanno, Y., & Kangas, S.E.N. (2014). "I'm not going to be, like, for the AP": English language learners' limited access to advanced college-preparatory courses in high school.

American Educational Research Journal, 51(5), 848-878.

Linquanti, R. (2001). *The reclassification dilemma: Challenges and choices in fostering meaningful accountability for English learners* (UC Language Minority Research Institute Policy Report 2001-01). Davis, CA: University of California, Davis.

Linquanti, R. & Cook, H.G. (2015). Re-examining reclassification: Guidance from a national working session on policies and practices for exiting students from English learner status. Washington, DC: Council of Chief State School Officers.

Pew Research Center (2008). *U.S. Population Projections: 2005-2050.* Washington, DC: Pew.

Raudenbush, S.W., Reardon, S.F., & Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness*, *5*(3), 303-332.

Robinson, J.P. (2011). Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis, 33*(3), 267-292.

Robinson-Cimpian, J.P., & Thompson, K.D. (in press). The effects of changing test-based policies for reclassifying English learners. *Journal of Policy Analysis and Management.* DOI:10.1002/pam.21882

Stock, J.H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*.

Thompson, K.D. (2015). English learners' time to reclassification: An analysis. *Educational Policy*. Advance online publication. DOI:10.1177/0895904815598394

Umansky, I.M. (2015). *The impact of English learner status on academic course-taking.* Paper

presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Umansky, I.M. & Reardon, S.F. (2014). Reclassification patterns among Latino English learner students in bilingual, dual immersion, and English immersion classrooms. *American Educational Research Journal 51*(5), 879-912.

WIDA (2015). *Consortium members*. Retrieved from https://www.wida.us/membership/states/

Table 1. Descriptive statistics, by state and analytic gradeband

	State A		State B			
	Gr 4-5	Gr 6-9	Gr 3-5	Gr 6-8	Gr 9-11	Graduate
Male	51.45	50.07	49.44	52.83	51.89	51.75
Race/ethnicity						
Hispanic	77.40	75.86	77.82	80.88	75.23	85.37
White	4.12	4.84	8.74	7.93	8.98	5.13
Black	13.86	14.27	1.48	1.55	2.89	1.38
Asian/Pac.Isl.	3.67	4.14	10.18	8.98	12.28	7.44
Other	0.95	0.90	1.78	0.66	0.62	0.67
Special education	11.58	8.19	8.37	15.57	13.49	12.17
Free/red.-price lunch	83.51	78.13	90.77	92.34	89.34	91.51
Attained policy	54.90	34.85	29.52	33.98	20.73	24.26
Reclassified (R-FEP)	19.61	25.75	29.68	35.80	28.61	30.79
Age (in years)	9.85 (0.79)	12.57 (1.36)	9.97 (0.85)	12.60 (0.84)	15.47 (0.88)	14.17 (1.51)
Time as EL (in years)	4.02 (1.45)	4.48 (2.44)	4.70 (1.22)	6.65 (1.97)	7.50 (3.27)	7.33 (2.72)
Number of districts	28	25	58	41	22	35
Sample size	16,831	14,257	38,479	20,753	6,011	11,218

Notes. Means for dichotomous variables as presented as percentages and appear in the upper portion of the table. Means for continuous variables appear in the lower portion of the table with standard deviations in parentheses.

Table 2. Effects of attaining state policy threshold on the likelihood of reclassification, by state, gradeband, and model

	State A		State B		
	Gr 4-5	Gr 6-9	Gr 3-5	Gr 6-8	Gr 9-11
Model 1: No student covariates					
Fixed effects					
Attained threshold	0.069**	0.270***	0.568***	0.567***	0.569***
Standard error	(0.022)	(0.029)	(0.036)	(0.030)	(0.096)
Random effects					
Attained threshold	0.067***	0.071	0.233***	0.160***	0.420***
Chi-squared (*df*)	56.65 (27)	33.97 (24)	1583.16 (57)	364.08 (40)	846.77 (21)
I-squared	52.3	29.4	96.4	89.0	97.5
Model 2: Student covariates added					
Fixed effects					
Attained threshold	0.078***	0.271***	0.569***	0.568***	0.565***
Standard error	(0.022)	(0.030)	(0.036)	(0.029)	(0.096)
Random effects					
Attained threshold	0.063**	0.073	0.235***	0.151***	0.413***
Chi-squared (*df*)	52.84 (27)	34.40 (24)	1619.43 (57)	364.40 (40)	665.19 (21)
I-squared	48.9	30.2	96.5	89.0	96.8
Number of districts	28	25	58	41	22
Sample size	16,831	14,257	38,479	20,753	6,011

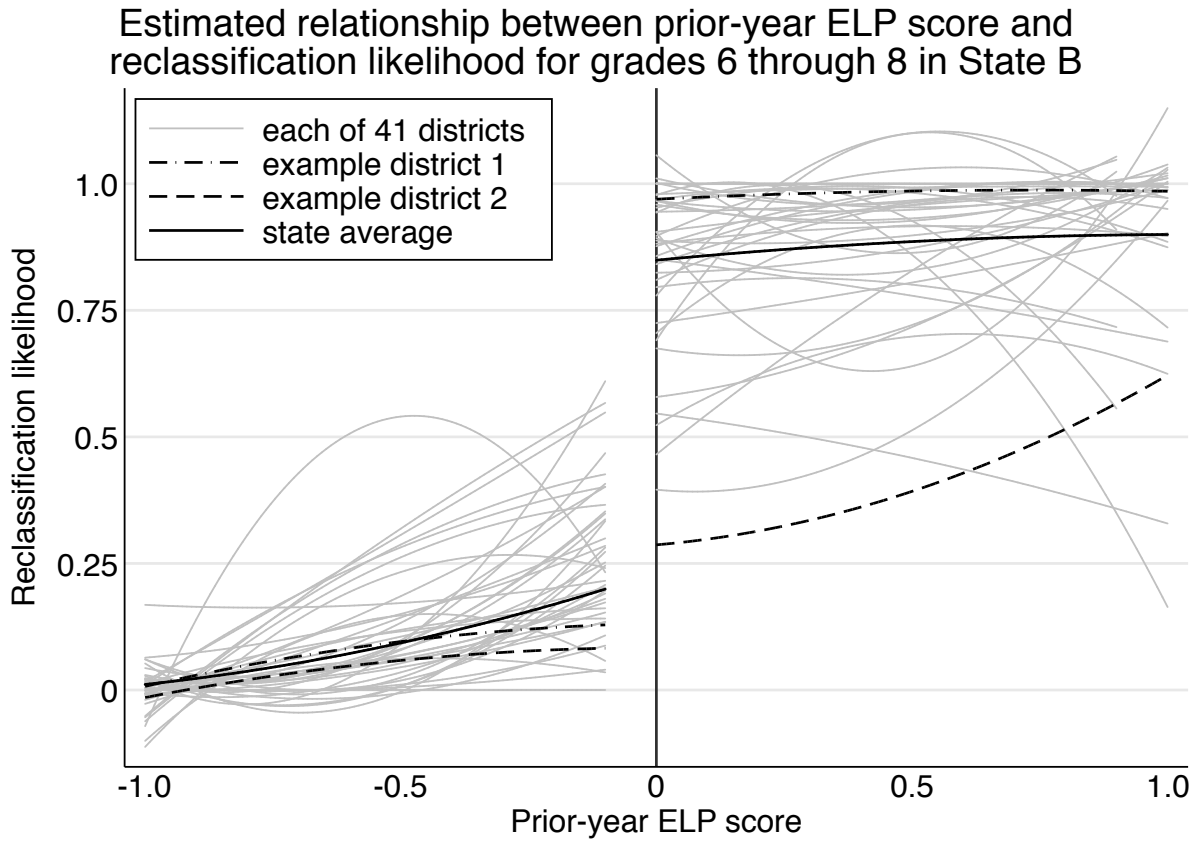
Notes. *p<.05; **p<.01; ***p<.001. All models include an indicator for whether the student attained the test threshold, linear and quadratic terms for the test containing the threshold, interactions between the test-score terms and the threshold indicator, and grade-by-year fixed effects. Model 2 adds covariates for gender, race/ethnicity, age, time as EL, lunch status, and special education status. The fixed effect for attained threshold is the statewide average effect for that attaining the threshold has on the likelihood of reclassification, with its standard error below it in parentheses. The value for the random effect for attained threshold is in standard deviation units, for ease of interpretation. The chi-squared test tests for significant between-district variance.

Table 3. Effects of threshold-induced reclassification on next-year achievement tests, by subject, gradeband, and model, for State B.

	Gr 3-5	Gr 6-8	Gr 9-11
Next year ELA achievement score			
Model 1: No student covariates			
Fixed effects			
Reclassified	-0.006	0.029	0.037
Standard error	(0.012)	(0.060)	(0.056)
Random effects			
Reclassified	<0.01	0.289***	0.142**
Chi-squared (*df*)	39.34 (40)	637.45 (31)	33.19 (15)
I-squared	0.0	95.1	54.8
Model 2: Student covariates added			
Fixed effects			
Reclassified	-0.006	0.036	0.056
Standard error	(0.013)	(0.040)	(0.060)
Random effects			
Reclassified	<0.01	0.172***	0.160**
Chi-squared (*df*)	33.19 (40)	180.31 (31)	36.70 (15)
I-squared	0.0	82.8	59.1
Number of districts	41	32	16
Sample size	34,391	18,928	4,646
Next year math achievement score			
Model 1: No student covariates			
Fixed effects			
Reclassified	-0.018	0.041	0.073
Standard error	(0.010)	(0.043)	(0.099)
Random effects			
Reclassified	<0.01	0.184***	0.242**
Chi-squared (*df*)	34.15 (40)	195.22 (31)	31.89 (15)
I-squared	0.0	84.1	53.0
Model 2: Student covariates added			
Fixed effects			
Reclassified	0.007	0.047	0.008
Standard error	(0.021)	(0.042)	(0.133)
Random effects			
Reclassified	0.054*	0.180***	0.345***
Chi-squared (*df*)	61.86 (40)	204.87 (31)	47.49 (15)
I-squared	35.3	84.9	68.4
Number of districts	41	32	16
Sample size	34,391	18,928	4,646

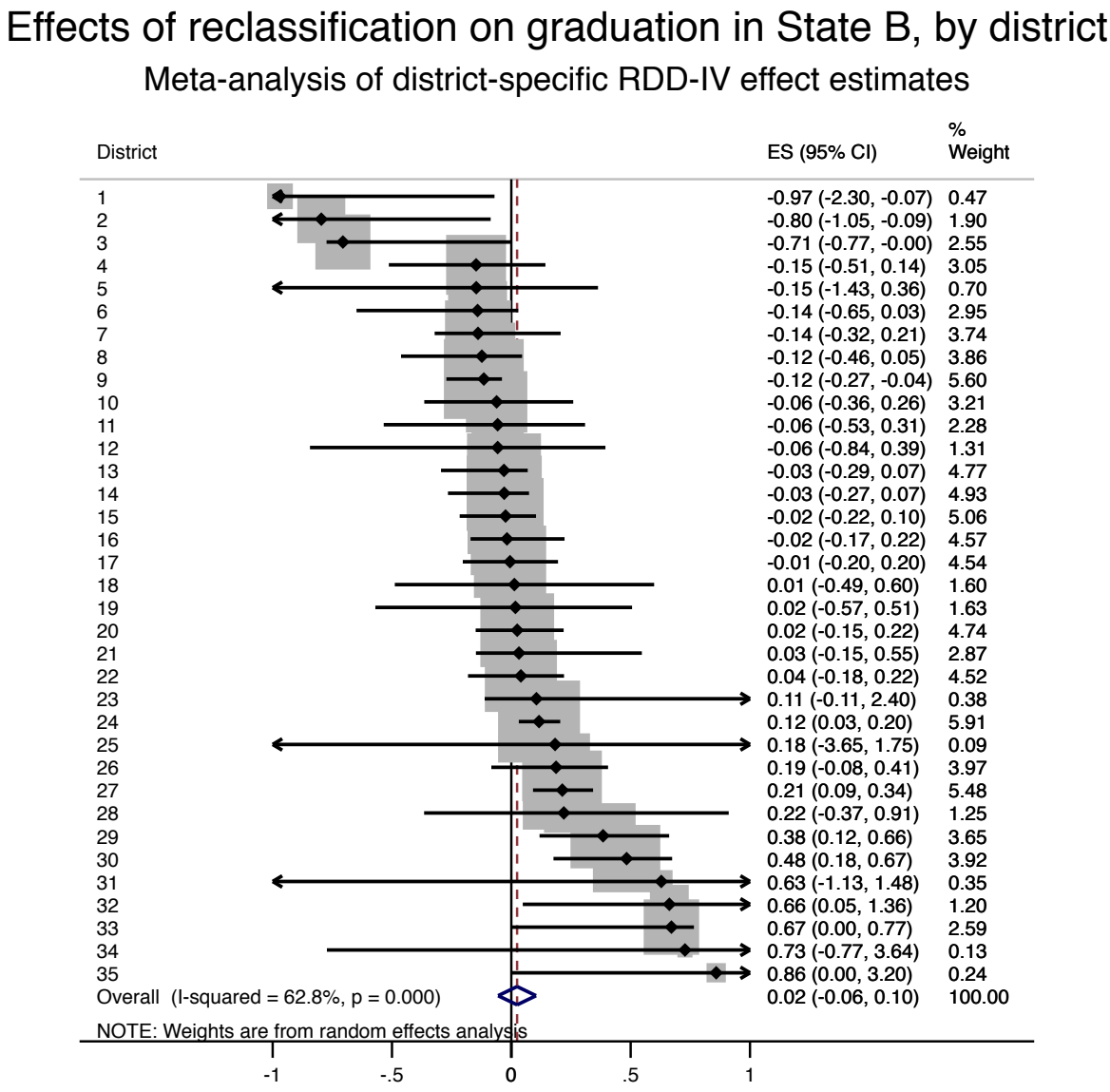
Notes. *p<.05; **p<.01; ***p<.001. All models include an indicator for whether the student attained the test threshold, linear and quadratic terms for the test containing the threshold, interactions between the test-score terms and the threshold indicator, and grade-by-year fixed effects. Model 2 adds covariates for gender, race/ethnicity, age, time as EL, lunch status, and special education status. The fixed effect for reclassified is the statewide average effect for threshold-induced reclassification, with its standard error below it in parentheses. The value for the random effect for reclassification is in standard deviation units, for ease of interpretation. The chi-squared test tests for significant between-district variance.

Figure 1.



Notes. The model represents the one described as Model 1 in Table 2. Each light-gray curve represented the quadratic relationship between reclassification likelihood and prior-year ELP score in a district, estimated separately above and below the state threshold. The dashed-and-dotted black curves are the estimates for a single district, to illustrate the jump in reclassification likelihood in one example district with a large jump (about 80 pp). The set of dashed black curves illustrate the jump in another example district, one with a smaller jump (about 20 pp). The solid black curves show the relationship for the state, averaged over the 41 districts in the analysis. The difference between where the state curves approach the threshold from above and below is the average effect of attaining the threshold on increasing the likelihood of reclassification (56.7 pp, $p < .001$). The standard deviation of the district-specific jumps in likelihood is 16 pp ($p < .001$).

Figure 2.



NOTE: District-specific 95% confidence intervals are constructed from 999 bootstrapped samples

Notes. Each district estimate reflects the district-specific policy-induced reclassification effect on graduation for the compliers. For each of the 35 district estimates, horizontal lines represent the percentile-*t* school-cluster bootstrapped 95% confidence interval for the estimate, and the size of the gray box represents the weight the district has in the meta-analysis. The weights are obtained from random-effects meta-analysis, and districts with more precise estimates are given greater weight when obtaining the state average effect. An open diamond in the last row represents the state average effect estimate of 2.4 percentage points (rounded to 2 in the graphic; $p = .55$).

Supplemental Online Materials

[All tables beyond this point are intended for publishing online only]

Appendix Table A. Tests for discontinuities in observed covariates, by state, grade level, and covariate-as-outcome

	State A		State B
	Gr 4-5	Gr 6-9	Graduation
Time as ELL			
Fixed effect	0.153 (0.081), $p = 0.060$	-0.028 (0.104), $p = 0.784$	0.215 (0.149), $p = 0.149$
Random effect	**0.214, $p = 0.024$**	0.076, $p = 0.432$	**0.462, $p = 0.010$**
Race: White			
Fixed effect	0.000 (0.000), $p = 0.754$	-0.006 (0.006), $p = 0.288$	0.001 (0.004), $p = 0.806$
Random effect	0.000, $p = 0.084$	0.010, $p = 0.072$	**0.017, $p < 0.001$**
Race: Black			
Fixed effect	0.000 (0.000), $p = 0.754$	0.000 (0.000), $p = 0.905$	-0.001 (0.001), $p = 0.252$
Random effect	0.000, $p = 0.706$	0.000, $p = 0.827$	0.000, $p = 0.281$
Race: Asian			
Fixed effect	0.001 (0.001), $p = 0.281$	-0.002 (0.008), $p = 0.811$	-0.003 (0.002), $p = 0.144$
Random effect	0.000, $p = 0.319$	0.017, $p = 0.064$	0.000, $p = 0.409$
Race: Other			
Fixed effect	0.000 (0.000), $p = 0.844$	0.000 (0.000), $p = 0.847$	0.000 (0.000), $p = 0.856$
Random effect	0.000, $p = 0.065$	0.000, $p = 0.438$	**0.000, $p = 0.005$**
Free/reduced price			
Fixed effect	0.000 (0.000), $p = 0.979$	0.010 (0.023), $p = 0.661$	-0.018 (0.019), $p = 0.340$
Random effect	0.037, $p = 0.073$	**0.052, $p = 0.020$**	**0.071, $p < 0.001$**
Male			
Fixed effect	**0.061 (0.023), $p = 0.009$**	-0.001 (0.033), $p = 0.976$	0.003 (0.038), $p = 0.934$
Random effect	0.039, $p = 0.299$	0.000, $p = 0.617$	**0.151, $p < 0.001$**
Age			
Fixed effect	0.037 (0.036), $p = 0.297$	-0.030 (0.027), $p = 0.263$	-0.006 (0.026), $p = 0.818$
Random effect	0.085, $p = 0.080$	0.000, $p = 0.884$	0.059, $p = 0.150$
Special education			
Fixed effect	0.017 (0.021), $p = 0.421$	-0.008 (0.013), $p = 0.545$	0.011 (0.014), $p = 0.430$
Random effect	**0.066, $p = 0.002$**	0.028, $p = 0.168$	**0.046, $p < 0.001$**

Notes. Boldface type indicates $p < .05$. Models include an indicator for whether the student attained the threshold, linear and quadratic terms for the test containing the threshold, interactions between the test-score terms and the threshold indicator, and grade-by-year fixed effects. For each outcome (e.g., Time as EL), the row labeled fixed effect contains the estimated above-below difference at the threshold across the state, its standard error in parentheses, and the associated p value. The random effect row presents the between-district SD and p value. The ideal here is to find little to no evidence of either mean discontinuities or between-district variation. Having found significant variation in many instances, we performed additional tests to assess whether the patterns of discontinuities in the above table correlated with the heterogeneity in effects found in the main analysis; as we discuss in the paper, we found no such correlations.

Appendix Table A (continued). Tests for discontinuities in observed covariates, by state, grade level, and covariate-as-outcome

	State B		
	Gr 3-5	Gr 6-8	Gr 9-11
Time as ELL			
Fixed effect	0.001(0.033), p =0.977	**0.148(0.075), p =0.049**	-0.270(0.200), p =0.176
Random effect	**0.203, p < 0.001**	**0.271, p < 0.001**	**0.543, p < 0.001**
Race: White			
Fixed effect	0.001(0.003), p =0.737	-0.005(0.004), p =0.181	0.007(0.007), p =0.347
Random effect	**0.010, p =0.002**	**0.010, p =0.001**	**0.020, p < 0.001**
Race: Black			
Fixed effect	**0.001(0.000), p =0.008**	0.001(0.001), p =0.181	-0.005(0.004), p =0.169
Random effect	0.000, p =0.500	**0.000, p =0.001**	**0.010, p < 0.001**
Race: Asian			
Fixed effect	0.002(0.002), p =0.388	0.003(0.003), p =0.251	0.018(0.012), p =0.127
Random effect	**0.010, p < 0.001**	**0.000, p =0.006**	**0.032, p < 0.001**
Race: Other			
Fixed effect	-0.001(0.002), p =0.606	0.000(0.000), p =0.763	0.001(0.002), p =0.545
Random effect	**0.000, p < 0.001**	**0.000, p =0.008**	**0.000, p =0.001**
Free/reduced price			
Fixed effect	-0.008(0.010), p =0.442	-0.006(0.007), p =0.400	-0.028(0.019), p =0.149
Random effect	**0.050, p < 0.001**	**0.022, p < 0.001**	**0.060, p < 0.001**
Male			
Fixed effect	-0.014(0.015), p =0.360	-0.034(0.029), p =0.242	0.012(0.050), p =0.810
Random effect	**0.060, p =0.002**	**0.127, p < 0.001**	**0.171, p < 0.001**
Age			
Fixed effect	-0.005(0.016), p =0.760	**-0.048(0.023), p =0.038**	-0.012(0.075), p =0.874
Random effect	**0.074, p < 0.001**	**0.107, p < 0.001**	**0.290, p < 0.001**
Special education			
Fixed effect	-0.004(0.012), p =0.738	0.035(0.018), p =0.055	0.037(0.039), p =0.338
Random effect	**0.062, p < 0.001**	**0.083, p < 0.001**	**0.147, p < 0.001**

Notes. Boldface type indicates $p < .05$. Models include an indicator for whether the student attained the threshold, linear and quadratic terms for the test containing the threshold, interactions between the test-score terms and the threshold indicator, and grade-by-year fixed effects. For each outcome (e.g., Time as EL), the row labeled fixed effect contains the estimated above-below difference at the threshold across the state, its standard error in parentheses, and the associated p value. The random effect row presents the between-district SD and p value. The ideal here is to find little to no evidence of either mean discontinuities or between-district variation. Having found significant variation in many instances, we performed additional tests to assess whether the patterns of discontinuities in the above table correlated with the heterogeneity in effects found in the main analysis; as we discuss in the paper, we found no such correlations.

Appendix Table B. Effects of attaining state policy threshold on the likelihood of reclassification, by state, gradeband, and model (for alternative samples: 0.8-SD and 1.2-SD samples)

	State A		State B		
	Gr 4-5	Gr 6-9	Gr 3-5	Gr 6-8	Gr 9-11
0.8-SD Population					
Model 1: No student covariates					
Fixed effects					
Attained threshold	0.072**	0.248***	0.589***	0.576***	0.575***
Standard error	-0.026	-0.034	-0.038	-0.041	-0.066
Random effects					
Attained threshold	0.077***	0.093*	0.240***	0.224***	0.251***
Chi-squared (*df*)	60.44 (27)	39.84 (23)	1573.31 (53)	594.17 (39)	213.55 (19)
I-squared	55.3	42.3	96.6	93.4	91.1
Model 2: Student covariates added					
Fixed effects					
Attained threshold	0.067**	0.251***	0.589***	0.574***	0.571***
Standard error	-0.026	-0.033	-0.037	-0.041	-0.066
Random effects					
Attained threshold	0.078***	0.087*	0.233***	0.228***	0.251***
Chi-squared (*df*)	59.43 (27)	36.47 (23)	1599.40 (53)	668.18 (39)	177.05 (19)
I-squared	54.6	36.9	96.7	94.2	89.3
Number of districts	28	24	54	40	20
Sample size	14,779	11,949	32,226	17,981	4,373
1.2-SD Population					
Model 1: No student covariates					
Fixed effects					
Attained threshold	0.075***	0.264***	0.576***	0.557***	0.538***
Standard error	-0.02	-0.041	-0.035	-0.032	-0.067
Random effects					
Attained threshold	0.061**	0.144***	0.227***	0.172***	0.293***
Chi-squared (*df*)	56.66 (29)	68.87 (25)	1886.96 (58)	277.26 (40)	369.60 (22)
I-squared	48.8	63.7	96.9	85.6	94
Model 2: Student covariates added					
Fixed effects					
Attained threshold	0.083***	0.260***	0.578***	0.558***	0.528***
Standard error	-0.02	-0.041	-0.034	-0.031	-0.065
Random effects					
Attained threshold	0.058**	0.149***	0.219***	0.165***	0.284***
Chi-squared (*df*)	54.82 (29)	72.78 (25)	1600.53 (58)	268.34 (40)	342.84 (22)
I-squared	47.1	65.6	96.4	85.1	93.6
Number of districts	30	26	59	41	23
Sample size	18,354	16,160	44,223	23,008	6,928

Notes. *p<.05; **p<.01; ***p<.001. See notes for Table 2 for model descriptions. The top panel uses a smaller sampling criterion, including only students within 0.8 SDs of attaining the threshold in valid districts. The bottom panel uses a larger sampling criterion, including only students within 1.2 SDs of attaining the threshold in valid districts. These alternative samples are used to test for model sensitivity and consistency of estimates. The results suggest that the models are robust to reasonably sized changes in sample inclusion criteria. See Appendix Table D for additional notes.

Appendix Table C. Effects of threshold-induced reclassification on next-year achievement tests, by subject, state, gradeband, and model (for alternative samples: 0.8-SD and 1.2-SD samples)

	0.8-SD Population			1.2-SD Population		
	Gr 3-5	Gr 6-8	Gr 9-11	Gr 3-5	Gr 6-8	Gr 9-11
	Next year ELA assessment					
Model 1: No student covariates						
Fixed effects						
Reclassified	-0.021	-0.020	0.006	-0.014	0.072	0.012
Standard error	(0.018)	(0.032)	(0.086)	(0.012)	(0.049)	(0.044)
Random effects						
Reclassified	0.056*	0.122***	0.246***	0.028	0.233***	0.091
Chi-squared (*df*)	62.90 (42)	173.14 (33)	39.24 (13)	48.20 (42)	358.63	25.64 (16)
I-squared	33.2	80.9	66.9	12.9	91.1	37.6
Model 2: Student covariates added						
Fixed effects						
Reclassified	-0.028	-0.015	0.011	-0.015	0.068	0.027
Standard error	(0.015)	(0.065)	(0.092)	(0.011)	(0.038)	(0.051)
Random effects						
Reclassified	0.032	0.326***	0.254***	<0.01	0.167***	0.126**
Chi-squared (*df*)	49.33 (42)	768.77 (33)	37.13 (13)	40.87 (42)	239.96	33.49 (16)
I-squared	14.9	95.7	65.0	0.0	86.7	52.2
Number of districts	43	34	14	43	33	17
Sample size	28,903	16,751	3,549	39,477	20,207	5,410
	Next year math assessment					
Model 1: No student covariates						
Fixed effects						
Reclassified	-0.001	-0.120	0.130	-0.020	0.035	-0.107
Standard error	(0.025)	(0.096)	(0.160)	(0.016)	(0.033)	(0.118)
Random effects						
Reclassified	<0.01	0.497***	0.437***	0.041	0.121***	0.296**
Chi-squared (*df*)	23.93 (42)	1100.20 (33)	52.01 (13)	51.82 (42)	99.59 (32)	36.60 (16)
I-squared	0.0	97.0	75.0	18.9	67.9	56.3
Model 2: Student covariates added						
Fixed effects						
Reclassified	-0.001	-0.086	0.126	-0.022	0.042	-0.200
Standard error	(0.020)	(0.061)	(0.177)	(0.020)	(0.032)	(0.125)
Random effects						
Reclassified	<0.01	0.291***	0.441**	0.069**	0.118***	0.327***
Chi-squared (*df*)	26.90 (42)	329.31 (33)	33.03 (13)	72.25 (42)	102.35	42.17 (16)
I-squared	0.0	90.0	60.6	41.9	68.7	62.1
Number of districts	43	34	14	43	33	17
Sample size	28,903	16,751	3,549	39,477	20,207	5,410

Notes. *p<.05; **p<.01; ***p<.001. See notes for Appendix Table D.

Appendix Table D. Effects of threshold-induced reclassification on 4-year graduation in State B, by model and sample inclusion criteria

	Sample inclusion criteria		
	0.8 SD	1.0 SD	1.2 SD
Model 1: No student covariates			
Fixed effects			
Reclassified	0.055	0.038	-0.005
Standard error	(0.039)	(0.045)	(0.038)
Random effects			
Reclassified	0.122**	0.185***	0.170***
Chi-squared (*df*)	56.20 (32)	114.62 (34)	99.39 (32)
I-squared	43.1	70.3	67.8
Model 2: Student covariates added			
Fixed effects			
Reclassified	0.049	0.024	-0.01
Standard error	(0.042)	(0.041)	(0.038)
Random effects			
Reclassified	0.138***	0.161***	0.145***
Chi-squared (*df*)	62.94 (32)	91.47 (34)	88.92 (32)
I-squared	49.2	62.8	64.0
Number of districts	33	35	33
Sample size	8,401	11,218	12,636

Notes. *p<.05; **p<.01; ***p<.001. All models include an indicator for whether the student attained the test threshold, linear and quadratic terms for the test containing the threshold, interactions between the test-score terms and the threshold indicator, and grade-by-year fixed effects. Model 2 adds covariates for gender, race/ethnicity, age, time as EL, lunch status, and special education status. The fixed effect for reclassified is the statewide average effect for threshold-induced reclassification, with its standard error below it in parentheses. The value for the random effect for reclassification is in standard deviation units, for ease of interpretation. The chi-squared test tests for significant between-district variance. Results are presented for samples that include students within 0.8, 1.0, and 1.2 SDs of the threshold, respectively, in districts that meet criteria for inclusion (i.e., at least 10 students above and below threshold, and first-stage F > 10). The districts included in the analysis change slightly from bandwidth to bandwidth based on whether the district meets the inclusion criteria just mentioned. The similarity of results across the bandwidths then speaks to the robustness of the results to not only changes in the bandwidth and sample size but also in subtle differences in which districts are included in the analysis.