ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-828

How Much Teacher Is in Teacher Rating Scales?

D. Betsy McCoach
University of Connecticut

Anthony J. Gambino
University of Connecticut

Scott J. Peters
NWEA

Daniel Long
University of Connecticut

Del Siegle
University of Connecticut

Teacher rating scales (TRS) are often used to make service eligibility decisions for exceptional learners. Although TRS are regularly used to identify student exceptionalism either as part of an informal nomination process or through behavioral rating scales, there is little research documenting the between-teacher variance in teacher ratings or the consequences of such rater dependence. To evaluate the possible benefits or disadvantages of using TRS as part of a gifted identification process, we examined the student-, teacher-, and school-level variance in TRS controlling for student ability and achievement to determine the unique information, consistency, and potential bias in TRS. Between 10% and 25% of a students' TRS score can be attributed to the teacher doing the rating, and between-teacher standard deviations represent an effect size of one-third to one-half standard deviation unit. Our results suggest that TRS are not easily comparable across teachers, making it impossible to set a cut score for admission into a program (or for further screening) that functions equitably across teachers.

VERSION: August 2023

Suggested citation: McCoach, D. Betsy, Anthony J. Gambino, Scott J. Peters, Daniel Long, and Del Siegle. (2023). How Much Teacher Is in Teacher Rating Scales?. (EdWorkingPaper: 23-828). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/6vpz-en33

How Much Teacher Is in Teacher Rating Scales?

D. Betsy McCoach[1], Anthony J. Gambino[1], Scott J. Peters[2], Daniel Long[1], and Del Siegle[1]

[1]University of Connecticut

[2] NWEA

Author Note

D. Betsy McCoach: ORCID 0000-0001-9063-6835

Anthony J. Gambino: ORCID 0000-0001-5820-1997

Scott J. Peters: ORCID 0000-0003-2459-3384

Daniel Long: ORCID: 0000-0003-1887-2993

Del Siegle: ORCID 0000-0001-5579-9217

We have no known conflicts of interest.

Correspondence concerning this article should be addresses to Betsy McCoach at betsy@uconn.edu

Abstract

Teacher rating scales (TRS) are often used to make service eligibility decisions for exceptional learners. Although TRS are regularly used to identify student exceptionalism either as part of an informal nomination process or through behavioral rating scales, there is little research documenting the between-teacher variance in teacher ratings or the consequences of such rater dependence. To evaluate the possible benefits or disadvantages of using TRS as part of a gifted identification process, we examined the student-, teacher-, and school-level variance in TRS controlling for student ability and achievement to determine the unique information, consistency, and potential bias in TRS. Between 10% and 25% of a students' TRS score can be attributed to the teacher doing the rating, and between-teacher standard deviations represent an effect size of one-third to one-half standard deviation unit. Our results suggest that TRS are not easily comparable across teachers, making it impossible to set a cut score for admission into a program (or for further screening) that functions equitably across teachers.

How Much Teacher Is in Teacher Rating Scales?

Teachers are often asked to provide recommendations, either informally or formally, as to whether a student should receive gifted and talented or special education services. Often the teacher completes a structured rating scale that contains items asking how often a student exhibits certain behaviors associated with a disability or the need for gifted and talented services. Within the last 10 years, the identification processes for both fields have received significant attention, both due to concerns about students being overlooked (Gentry et al., 2022; McBee et al., 2016; Morgan et al., 2015) and due to long-documented inequities within the populations of students served (Hosp & Reschly, 2003; Peters et al., 2019).

Educators and professional societies have long advocated for the use of multiple data points in making service eligibility decisions (American Education Research Association et al., 2014; Council for Exceptional Children, 2020). Multiple criteria identification systems commonly include teacher nominations and/or teacher rating scales (TRS). In fact, Callahan et al. (2014) reported that over 86% of school districts used teacher nominations. Similarly, a three-state survey of school districts conducted by the National Center for Research on Gifted Education (NCRGE) found that over 90% of school districts used teacher nominations and/or TRS to identify students for gifted services (Siegle et al., 2018).

TRS feature prominently in many multi-criteria identification systems and are often advocated as one way to diversify the pool of students that are identified as gifted (Harradine et al., 2014; Peters & Gentry, 2010). Peters and Gentry (2010) argued that combining TRS with information from achievement and ability tests creates a more comprehensive picture of a child and helps to determine the most appropriate services. The impact of including teacher

information depends on the method by which teacher ratings are incorporated into the identification decision as well as the reliability, validity, and equity of the teacher ratings.

There is a general belief that teacher ratings add unique information to the identification process, improve multi-criteria systems, and provide a more nuanced assessment of giftedness that should increase diversity.

However, TRS and nominations could also be counterproductive for at least two reasons. First, unlike traditional standardized tests, TRS are especially prone to rater dependence: the score a student receives depends, at least in part, on the teacher who provides the rating. Wide variability across teachers in their use of the TRS could introduce inconsistency or unreliability into the overall identification system. Second, although teachers have the potential to notice talents or unmet learning needs that might not manifest on standardized tests, recent research has shown teachers are just as likely to hold pro-White implicit and explicit biases as the overall population (Starck et al., 2020). Systems that rely on TRS must consider the ways in which between-teacher inconsistency can decrease precision and the degree to which teacher bias can compromise the equity of identification decisions. Inconsistencies between teachers in their use of the scale and teacher bias may compromise the validity of the inferences that we seek to make when using TRS.

Because few studies have evaluated whether TRS introduce construct irrelevant variance or bias, this study examined both consistency and bias in formal TRS. We were particularly interested in estimating the degree to which between-teacher differences in the use of TRS could influence identification decisions.

Literature Review

Pros and Cons of Teacher Ratings

There are two contrasting and somewhat contradictory perspectives on teacher judgments (Martínez et al., 2009). Some view teachers as the best source of information about student potential because they have a much deeper understanding of students' strengths and skills (VanTassel-Baska, 2008). However, there is good reason to be skeptical of teachers' ability to judge students' performance fairly and accurately. Teacher bias is often mentioned as a possible contributing factor to disproportionality in identification (Ford & King, 2014; Grissom & Redding, 2016). Teachers' beliefs about giftedness (Miller, 2009) and (implicit or explicit) biases can affect the way in which teachers rate students (Martínez et al., 2009). Multiple studies examining the degree of between-teacher variance on teachers' ratings of students' behavior suggest that 15%-39% of the variance on behavioral rating scales lies between teachers (Eklund et al., 2017; Martínez et al., 2009; Mashburn et al., 2006; Peters et al., 2012). Therefore, students' scores may partially depend upon which teacher completes the rating scale. Such construct-irrelevant variance can compromise the validity of the information TRS provide. Given these contradictory perspectives (Martínez et al., 2009), there has been surprisingly little research on the role TRS play in the identification of students as gifted.

Teacher Ratings in Gifted Identification

TRS are a diverse class of instruments. In some cases, the teacher rating is little more than a dichotomous recommendation that the student would benefit from specialized services or should undergo further testing. In gifted education, TRS usually ask teachers to rate each student on a list of characteristics or behaviors that are associated with advanced ability, domain specific ability (e.g., mathematics), or metacognitive characteristics (e.g., critical or creative thinking). The teacher rates how often a student demonstrates these behaviors or how well the characteristics describe the student. These same instruments may be collected from all teachers

on all students as an initial screener, or they may be collected on a subset of students to make a final eligibility decision. Some TRS are published and well-researched instruments. Examples of such TRS include the Gifted Rating Scales (GRS; Pfeiffer & Jarosewich, 2003), the Scales for Identifying Gifted Students (SIGS; Ryser & McConnell, 2004) and the HOPE Scale (Gentry et al., 2015). Other TRS have little or no psychometric information or research support. Different types of TRS may exhibit varying degrees of rater dependence, but this has yet to be examined.

Reliability of Teacher Judgements

Despite potential reliability challenges, assessments that rely on subjective teacher judgement are common in education. Harlen's (2005) systematic review of the literature on teacher summative judgements indicated that the overall reliability of scores from student artifacts, such as portfolios, was low. Reliability increased when there was greater structure to the scoring criteria or when the raters had taken part in the development of the scoring criteria. Several construct-irrelevant factors may influence teacher ratings. For example, Harlen (2005) found that student behavior, gender, prior achievement, and special education needs tended to bias teacher ratings. Harlen also emphasized the importance of task specificity, clarity of the criteria, and teacher training to improve the reliability of TRS data. In short, the broader and less defined the teacher rating task is, the less reliable the scores are likely to be, both within and across teachers. Many gifted TRS are general, and nearly all TRS were designed without the involvement of the general classroom teachers who conduct the ratings. Therefore, both inconsistency and bias are potential concerns.

The published reliability coefficients of psychometrically validated TRS within special education and gifted education appear to be high. For example, all the subscales of the *SIGS* demonstrated internal consistency values of .93 or greater (Ryser & McConnell, 2004).

Similarly, the HOPE Scale reported alpha reliability of .96 (Gentry et al., 2015). The Hispanic Bilingual Gifted Screening Instrument (HBGSI; Fultz et al., 2013) showed split-half reliability coefficients of .79 and alpha reliability coefficients of .92 or greater. However, reliability information for district created TRS is generally unavailable. More importantly, most of these reliability estimates assess internal consistency reliability, which is the consistency with which a rater answers multiple items on a scale. Far less research exists on the inter-rater reliability of TRS in gifted education. Generally, TRS in gifted education ask teachers to assess student traits such as creative thinking or problem solving or general student performance. Such questions require the respondent to have worked with the student for a substantial period of time in a classroom environment. Therefore, in elementary school, only the classroom teacher knows the student well enough to complete the TRS. Consequently, gathering information about the inter-rater reliability of TRS measures in gifted education is virtually impossible.

However, a substantial body of research has examined the inter-rater reliability of TRS that are used to assess behavior and academic challenges. For example, Volpe et al. (2005) conducted a review of seven coding schemes used to observe and assess student classroom behavior. Interobserver agreement tended to be high: .93 to .98 for the Behavioral Observation of Students in Schools, .96 for the Academic Engaged Time Code of the Systematic Screening of Behavioral Disorders instrument, and .81 for the State-Event Classroom Observation System. Like Harlen (2005), Volpe et al. emphasized that structure, training to assure observers know what to look for when using these instruments, and even collaborating with a second observer can decrease bias and increase consistency of ratings, which is essential for obtaining useful information from these TRS. These published coefficients suggest relatively strong reliability for TRS within the behavioral domain. However, gifted TRS generally ask teachers to make

inferences about unobservable traits such as creativity, motivation, cognitive ability, etc. rather than documenting specific, observable behaviors. Therefore, it is unclear whether gifted TRS would exhibit similar levels of inter-rater reliability.

Validity

Concurrent validity coefficients between TRS and academic achievement tests, cognitive ability tests, individual intelligence tests, and even creativity and nonverbal tests are common. TRS tend to correlate moderately well with student-level assessments of behavior, achievement, or ability. For example, the technical manual for the SIGS reported correlations between the general intellectual subscale and the WISC-III and CogAT of .67 and .48 respectively (Ryser & McConnell, 2004). The authors of the Gifted Rating Scales (Pfeiffer & Jarosewich, 2003) reported correlations of 0.62 between the intellectual subscale and the WIAT-II composite and 0.54 between the Academic subscale and the WISC-IV full scale IQ score. Gentry et al. (2015) reported correlations in the mid .50s between HOPE Scale scores and academic achievement test scores. Some degree of non-overlap should be expected given these instruments are designed to measure a constellation of complex behaviors and characteristics associated with giftedness and are meant to enhance or augment the information provided by more traditional identification instruments (Matthews, 2018).

In special education research, Meissel et al. (2017) evaluated the degree to which teacher judgments of academic skills aligned with student achievement. The correlation was .72 for reading and .73 for writing, but there was wide variation in these correlations across schools. Some schools showed a correlation of -.50 between teacher judgment and reading scores whereas others showed a .94 correlation between teacher judgment and writing achievement. After controlling for student achievement, nearly 75% of the remaining variance was at the student

level. Importantly, certain students received lower average teacher ratings even *after* controlling for academic achievement. Specifically, English learners (ELs) and students with special needs scored -.14 and -0.57 standard deviation units lower in reading, suggesting potential teacher bias.

Bias in TRS

Reliability and validity evidence is far more commonly reported in gifted TRS than is information about assessment bias. However, several studies have examined differential item functioning or measurement invariance. In an analysis of data from the GRS, Pfeiffer and Jarosewich (2007) found no differences across racial/ethnic groups. Similarly, Ryser and McConnell (2004) eliminated items that showed significant differential effects via a logistic regression that included demographic group as a predictor. Finally, the HOPE Scale has undergone the most testing for measurement invariance (Peters & Gentry, 2010; Peters & Gentry, 2012; Pereira, 2021). All three analyses of HOPE Scale data used multi-group confirmatory factor analysis (MCFA) to evaluate measurement invariance across racial/ethnic, gender, language, and income groups. Pereira (2021) found a slightly worse fitting model for students who were still learning English when constraining all indicator intercepts. Similarly, Peters and Gentry (2012) found the model fit slightly worse for females than males when constraining indicator intercepts. However, when testing invariance across racial/ethnic groups, the scalar model did not result in any worse model fit. Similarly, constraining equal latent means across groups showed no decrease in fit. These TRS findings suggest some small degree of group-specific non-invariance.

Although an instrument may exhibit no evidence of measurement non-invariance at the item level, this does not guarantee that the instrument is unbiased, nor does it guarantee that the inclusion of such an instrument in the identification process would eliminate biases in the

identification process. Most analyses of bias examine whether there are group differences at the item level, holding constant group performance at the construct level. Scale score difference across subgroups could be indicative of either pervasive bias on the instrument or true differences between the two groups on the underlying construct.

Bias in Teacher Judgements

The disproportional underrepresentation of students of color and those from low-income families is a persistent problem in gifted education (Grissom et. al., 2019; Hamilton et. al, 2018; Long et al., 2023; Peters et al., 2019). American K-12 students are far more racially/ethnically diverse than the teacher workforce. In the 2017–2018 school year, approximately 79% of public-school teachers were White (NCES, 2020a), whereas only 48% of public-school students were White (NCES, 2020b). This has led to concerns over bias in measures that require subjective teacher judgement.

Two studies used nationally representative datasets (National Education Longitudinal Study (NELS) of 1998 and the Early Childhood Longitudinal Study (ECLS-K) to evaluate whether students who had same-race teachers received different subjective evaluations. Ehrenberg et al. (1995) found that same-race teachers viewed their students more favorably when recommending the student for academic honors or college attendance. Using the ECLS-K to conduct a similar study, Downey et al. (2004) found that Black students were typically rated as having more externalizing behaviors than White students but this was partially moderated by the race of the teacher. Black teachers provided more favorable evaluations of Black students' effort, but there was no difference between Black and White teachers on ratings of Black students' disruptiveness in class. Overall, Downey et al. found that the size of the effect due to student-teacher racial matching ranged from .05 to .10 standard deviation units. Although the effect is

quite modest, even after controlling for other factors, the race of the teacher and the race of the student interacted to influence subjective ratings, even in cases where the criteria were relatively clear and explicit (e.g., behavior).

Grissom et al. (2019) and Grissom and Redding (2016) both noted the influence of teachers in the gifted identification process and how this influence allows for teacher bias, but also indirect parental bias. Using the ECLS-K, Grissom and Redding estimated the effect of student-teacher racial matching on a student's probability of receiving gifted services. After controlling for SES, achievement in math and reading, and other student-level characteristics, White-Hispanic and White-Asian gaps in probability of being identified were not statistically significant. However, the White-Black gap remained, and it was partially explained by whether a Black student also had a Black teacher. The proportion of Black students in the school was also a large and statistically significant predictor of Black students being identified as gifted. This suggests teacher subjectivity may play a role within the identification process; substantial discretion on the part of the teacher may help some students while holding others back.

Grissom et al. (2019) further explored this line of research by evaluating whether SES moderated the role of student race in predicting receipt of gifted services. Black and White students in the lowest two SES deciles showed similar rates of gifted identification; however, these gaps grew as SES increased. Within the highest income decile, there was a 10-point Black-White gap in identification rates. Having at least one parent in a "very high" prestige occupation or a parent whose income is at least $200,000+ per year were predictive of gifted identification. One theory is that such cultural capital likely manifests itself through advocacy at the school or teacher level, given how commonly referrals are a catalyst for identification.

Teacher-Level Variance

By their very nature, students' TRS scores are not independent. Instead, they are nested within teachers, and the degree of this intraclass correlation (ICC) has implications for both the reliability of the scores and validity of the inferences that educators can make from the data.

Multiple studies examining the degree of between-teacher variance on teachers' ratings of students' behavior indicate that 15%-39% of the variance on behavioral rating scales lies between teachers (Eklund et al., 2017; Martinez et al., 2009; Mashburn et al., 2006; Peters & Gentry, 2012). For example, in an evaluation of HOPE Scale scores from 1,671 students who were rated by 71 teachers, Peters and Gentry (2012) reported ICCs of .13 and .15 for the Academic and Social subscales. In a more recent analysis of HOPE Scale data that included 1,467 students, 572 of whom were English language learners, Pereira (2021) reported ICCs of .19 and .20 for the Academic and Social subscales. Splett et al. (2018) conducted similar analyses on teacher ratings of student behavior and emotional risk as measured by the Behavioral and Emotional Screening System (BESS). A total of 68 teachers completed ratings on 1,241 students. Between 8% and 20% of the variance in student-level scores was between teachers. Including other student- and teacher-level covariates (e.g., achievement, race, teacher years of experience) in the model reduced the rater/teacher effect to 17.8%.

Rambo-Hernandez et al. (2023) conducted the most-relevant study on between-teacher variance in TRS. The authors analyzed Scales for Rating the Behavioral Characteristics of Superior Students (SRBCSS) scores on creativity, motivation, mathematics, and science from 282 first- and 144 second-grade students who were rated by their classroom teachers (n = 16). Results showed ICCs on the creativity, motivation, mathematics, and science scales of .28, .21, .17, and .36, respectively, thus indicating that between 17% and 36% of the variability in the scores students received was due to between-teacher variability. Assuming students were not

assigned to classrooms based on their level of creativity, motivation, mathematics, or science, TRS scores in any of these areas should not show such large between-teacher variability. The large proportion of between-teacher variance suggests that teachers vary in their understanding and use of TRS.

Because a substantial proportion of the score variance in TRS is attributable to the teacher, students' scores depend, at least partially, on which teacher completes the rating scale. This construct-irrelevant variance may compromise the validity of the information provided by the TRS. A review of the literature on the most-common TRS in the field of gifted education did not reveal any additional research on between-teacher variability or rater effects in TRS. This is especially concerning given the frequent use of high, inflexible cut scores in gifted student identification (Dai & Chen, 2013; Hertzog, 2009).

Rater Effects

The term *rater effects* refers to variance that is attributable to the rater rather than to the performance or ability of the ratee (Anthony et al., 2022; Myford & Wolfe, 2003). "Ratings are rooted in observation, interpretation, and perhaps most importantly, the exercise of personal and professional judgement" (Myford & Wolfe, p. 389). There are several different types of rater effects: leniency/severity, halo effects, central tendency, and restriction of range (Myford & Wolfe, 2003). Leniency refers to a rater's tendency to consistently assign higher than average ratings (across ratees) whereas severity refers to a rater's tendency to consistently assign lower than average ratings (across ratees). The term halo effect refers to a rater's tendency to allow their perception of a ratee in general, or in another domain, influence their ratings. Restriction of range refers to the tendency of a rater to cluster responses around any point of the rating continuum. Central tendency is one type of range restriction in which the rater overuses the

middle points of the scale and underuses the end points of the scale. If some raters engage in more range restriction than others, the scores of raters who exhibit greater range restriction exhibit lower within-rater variance than the scores of raters who more fully use the rating scale continuum.

The identification of these three general types of rating effects (leniency/severity, halo effects, and range restriction/central tendency) requires three different analyses. Imagine that a set of raters rate an identical set of examinees. Leniency/severity differences across raters and rater effects would lead to differences in the raters' average scores. In contrast, halo effects would result in different scores across ratees (but not necessarily different average scores across raters), which would increase the within-rater correlations. Differences in range restriction/central tendency would lead to differences in the standard deviations (variances) of ratings across raters. Rater effects may also be due to between-rater differences in the rater's interpretation of the rating tasks. Rater biases also represent rater effects (Anthony et al., 2022). For example, a rater may be biased for or against a particular group of people based on their race/ethnicity, gender identity, appearance, or other characteristic. Like halo effects, such rater biases would result in different scores across students, but not necessarily different average scores across teachers. The current study examines between-rater effects that manifest in between-rater differences, such as leniency/severity and restriction of range as well as rater effects that result between-student differences, such as rater bias.

Summary

In the United States, at the elementary level alone there are more than 1.8 million teachers (NCES, 2020a). This represents the potential for 1.8 million different raters, each of whom possess idiosyncrasies and see the world through different lenses that might contribute to

inconsistency and/or bias in their ratings of students. Understanding the degree to which teachers can provide consistent, unbiased ratings of student behaviors and attributing that variability in teacher ratings to student-, teacher-, or school-level factors is essential to their ability to provide valid and equitable information for use in gifted and talented identification decisions. In this study, we examined the degree to which TRS are likely to influence the gifted identification process.

Methods

We examined the degree of between-teacher variance in gifted TRS, both before and after controlling for cognitive ability and academic achievement. In addition, we examined the degree to which teachers' ratings were predicted by student demographics such as race/ethnicity, EL and/or free-and-reduced lunch (FRL) status[1], after controlling for ability and achievement. Finally, we used the variance estimates from the results to estimate the degree to which teacher ratings and subsequent identification decisions are likely to be influenced by which teacher completes the TRS.

The Overall Logic of the Study

Our data contain a three-level hierarchy: Students are nested within teachers who are nested within schools. The unconditional three-level model contains three orthogonal variance components: the variance that lies between students within teachers (level-1 variance, σ^2), the variance that lies between teachers within schools (level-2 variance, τ_{00_π}), and the variance that

[1] District H did not provide an FRL status indicator. Instead, they provided a low-income status indicator that identified students as being low-income if they were eligible for SNAP, Medicaid, or other government benefits. For simplicity, we refer to this low-income status variable as the FRL status variable throughout the rest of this paper due to the other districts' datasets containing true FRL status variables (except for District C, which did not provide any indicator for FRL or low-income status).

lies between schools (level-3 variance, τ_{00_β}). The total variance in the TRS is the sum of these three variances in the unconditional model: $\sigma^2_{Total} = \sigma^2 + \tau_{00_\pi} + \tau_{00_\beta}$.

The Most Probable Sources of TRS Variance at Each of the Three Levels

Between-Student (Within-Teacher) Variance

Unconditional Between-Student (Within-Teacher) Variance. Generally, most of the variance in variables that measure individual differences lies between students (within teachers). Students within classrooms vary greatly in terms of their abilities, prior experiences, prior achievement, home experiences, and backgrounds. In fact, Pedersen et al. (2023) found that 69% of Grade 4 math classrooms included students across all four achievement benchmarks and 68% of the variance lied within classrooms at each grade level. Because TRS are designed to measure individual differences among students, most of the variability in TRS should lie between students (within teachers).

Residual Between-Student (Within-Teacher) Variance. Residual between-student variance in TRS is the between-student variance in scores that is *unexplained by ability or achievement*. This residual variance results from three potential sources:

1.) *Confounders at level 1.* Some between-student differences in traits/factors that influence teacher ratings are not completely explained/predicted by ability and/or achievement. For example, creativity and motivation are traits that are not perfectly correlated with ability and/or achievement. The unique variance in creativity or motivation that is not predicted by ability and/or achievement may help predict teacher ratings. However, these variables are unobserved. Therefore, we cannot account for within-student differences in the unique variance in traits that have not been measured. If teachers detect important components of giftedness that both ability

and achievement tests miss and those qualities vary across students, then we would expect to see substantial ***student-level*** residual variance in models that condition on ability and achievement.

2.) *Teacher Bias:* If teachers rate students of equal ability and achievement levels differently based on student demographics such as race/ethnicity or gender or based on student behavior or other student characteristics, these differences would be included in the residual between-student (within-teacher) variance. Our final model conditions on student demographics, and we can determine what percentage of the between-student variance in TRS is uniquely explained by demographic variables such as race/ethnicity.

3.) *Measurement error.* Another potential source of between-student residual variance is measurement error in the TRS.

Given these three potential sources, interpreting residual between-student variance is not clear-cut. Even after controlling for ability, achievement, and demographics, we expected a substantial amount of residual between-student variance in TRS.

Between-Teacher (Within-School) Variance

Unconditional Between-Teacher (Within-School Variance). Conceptually, between-teacher (within-school) variance represents variance in *true* class average scores within the same school. Some differences in class means are due to sampling error, but the level-2 variance component is designed to capture the *true* between-teacher variance (Raudenbush & Bryk, 2002). If class means differ from each other only by the magnitude that would be expected due to sampling error, then the level-2 variance component should be 0 (or very near zero). Therefore,

the level-2 variance component is a measure of *true* variability between classrooms in terms of the outcome variable of interest.

Why might classrooms in the same school differ from each other in terms of achievement, ability, and/or TRS? The most obvious explanation is that students were not randomly assigned to classes. Grouping students in such a way that high ability/high achieving students tend to be clustered within certain classes and low ability/low achieving students tend to be clustered within other classes would result in substantial between-classroom variance in ability, achievement, and TRS. Differences in teacher quality/effectiveness might also influence students' achievement, ability, and/or TRS. In other words, more effective teachers might increase the achievement and cognitive skills of their students more than less effective teachers, which would raise the mean achievement and ability level of their class relative to other classes in the school. If this were the case, teachers might rate their students higher because their students actually performed better. Finally, differences in the way that teachers conceptualize giftedness and/or use the TRS could result in between-teacher differences in the TRS.

Residual Between-Teacher (Within-School) Variance. After controlling for ability and achievement, the residual between-teacher variance in TRS represents between-teacher variance in TRS that cannot be explained by between-teacher/class differences in ability or achievement. The true class TRS varies more than would be expected by sampling error in ways that cannot be explained by the ability and/or achievement level of the class. As mentioned above, if teachers' ratings are dependent on other student characteristics (such as creativity or motivation), then we would expect to see substantial ***student-level*** (within teacher) residual variance in models that condition on ability and achievement, but this should ***not*** result in substantial between-teacher residual variance. Between-teacher variance functions at the ***teacher-level***: Teachers' scores

differ from each other in ways that cannot be explained by either *student* or *school* characteristics. This is an important point: The between-teacher variance is of greatest interest in this study. Measurement error, in the traditional sense, is part of the residual level-1 variance. Residual between-teacher variance represents between-teacher differences in TRS. This means that whatever contributes to this residual variance is consistent within teacher but varies across teachers.

If the between-teacher variance were due to compositional differences in the classes (i.e., ability and achievement is not randomly distributed across classes, but rather, some classes are higher achieving/higher ability than others), then controlling for both ability and achievement should eliminate the between-teacher variability in TRS. Controlling for class average ability and achievement should eliminate between-teacher variance in TRS if the between-teacher variance is due to between-class differences in ability and/or achievement, regardless of the source (assuming those differences are detectable on the administered ability and/or achievement tests).

After controlling for ability and achievement, theoretically, the residual between-teacher variance in TRS scores should be near 0 if between-class/between-teacher differences in achievement and/or ability explain between-teacher differences in the TRS. Residual between-teacher variance in TRS that remains after controlling for class ability or achievement represents differences between classrooms/teachers (***not*** between students within classes) on the TRS that cannot be explained by the average ability or average achievement of the class.

If teachers use the TRS differently in ways that are unrelated to classroom-level differences in achievement and/or ability, between-teacher variance in the TRS would remain, even after controlling for ability and achievement. The remaining (residual) between-teacher variance might capture between-teacher differences in the way that teachers conceptualize

giftedness and/or it might capture between-teacher differences in their use of the TRS. Whatever the source of the residual between-teacher variance after controlling for ability and achievement, this residual variance indicates that teachers differ from each other in terms of how they rate students with similar achievement and ability profiles. In other words, systematic variance in TRS is attributable to the teacher who completes the rating scale: Some teachers' ratings are systematically higher and some teachers' ratings are systematically lower, even after controlling for students' ability and achievement. Such differences are akin to rater effects (Wind, 2020). For instance, teachers could differ in terms of the severity with which they judge students in their class. Some teachers could be more lenient raters (i.e., they tend to give higher scores in general, even after controlling for ability) and some teachers could be stricter raters (i.e., they tend to give lower scores in general, even after controlling for ability). The larger the proportion of residual between-teacher variance in the TRS, the more TRS scores are explained by between-teacher differences, suggesting that teachers interpret/use the TRS differently. Although such between-teacher differences may not represent systematic bias against particular individuals or subgroups, the effect of such differences in the use of the rating scale would advantage students from certain classrooms and disadvantage students from other classrooms.

We hypothesized that adding ability and achievement as covariates would reduce between-teacher (within-school) variability, but that *a non-negligible amount of residual between-teacher variance in TRS would remain*. In other words, some portion of the TRS is teacher-specific, and as such, represents construct irrelevant variance.

In summary, between-teacher (within-school) variance in TRS could indicate that

1. Students within the same school are non-randomly sorted into classes (teachers), based on underlying traits that are measured by (or related to traits measured by the gifted rating scale(s).
2. Teachers differ in terms of how they use the TRS, resulting in variability across teachers in terms of class means.
3. Some combination of 1 and 2.

If between-teacher variability in the TRS were solely due to differences in the composition of the classroom in terms of ability and/or achievement, then we would expect (a) the proportion of between-teacher variance in the TRS to be fairly similar to the proportion of between-teacher variance in ability and achievement and (b) after controlling for ability and achievement, the residual between-teacher variance in TRS would drop to near zero. However, a substantial amount of residual between-teacher (within school) variance in TRS after controlling for ability and achievement provides support for explanation #2—that teachers differ in terms of how they use the TRS.

Between-School Variance

Unconditional Between-School Variance. Between-school variability in achievement, ability, and TRS is most likely due to differences in the demographics/clientele of the schools. Schools within the same district often differ considerably in terms of the community that they serve and their average achievement scores. Locally norming an assessment (creating school specific norms) is akin to group mean centering, which eliminates between-school variance.

Residual Between-School Variance. Substantial between-school variance in ratings, after controlling for school ability and school achievement indicates that schools differ from each other in terms of average teacher ratings for students with similar ability and achievement

profiles. This suggests that schools may differ in the way in which they use the TRS. For example, perhaps some schools provided training or professional development on the rating scale and others did not. If schools create local (school-based) norms, then residual between-school variance is not a concern. However, if a district does not use school-based norms, students in schools with higher mean rating scale scores benefit from such between-school differences; they are more likely to be identified/nominated, holding ability/achievement constant. In contrast, students in schools with lower mean rating scale scores are less likely to be identified/nominated, holding ability/achievement constant. We hypothesized that after controlling for ability and achievement, there would be fairly little (to no) between-school variability in TRS data.

Research Questions

The purpose of this study was to determine the degree to which TRS were influenced by between-teacher differences in TRS and whether there were consistent demographic differences in TRS after controlling for ability and achievement.

1. **After controlling for ability and achievement, do student demographic variables (race/ethnicity, FRL status, EL status, and gender) predict students' ratings? How much is between-student variance due to demographics likely to influence the screening and identification process?**

We hypothesized that race/ethnicity might predict residual between-student variance on the TRS, suggesting that students from certain demographic groups are rated differently, even after controlling for ability and achievement. There are two potential reasons that teachers might rate students from a particular demographic group systematically higher (or lower) than other students, even when students are equated on both ability and achievement.

a.) Teachers might exhibit either an implicit or explicit bias either for or against a particular demographic group.

b.) Students from a particular demographic group who have equivalent ability and achievement scores are systematically higher or lower on crucial unobserved variables (such as creativity or motivation or teacher pleasing behaviors) that teachers weigh in the TRS.

Ideally, none (or very little) of the residual between-student variance in TRS scores should be explained by student demographics (after controlling for ability and achievement). Given the current focus on increasing the identification of traditionally underserved students as gifted, it is possible that traditionally underserved students (i.e., Black, Latinx, and Native American students) might actually receive *higher* teacher ratings than students of other ethnicities, after controlling for ability and achievement. On the other hand, given the long history of underrepresentation in programs for the gifted, traditionally underserved students might receive *lower* ratings, even after controlling for ability and achievement. To test this hypothesis, we statistically evaluated the level-1 slopes for the demographic variables, after controlling for ability, achievement, and other demographics, and we computed Cohen's *d* effect sizes to contextualize the magnitude of any statistically significant effects.

2. **How much between-teacher variance is there in TRS? After controlling for ability, achievement, and demographics, how much between-teacher variance remains? How much is between-teacher variance likely to influence the screening and identification process?**

We hypothesized that (a) a substantial proportion of the variance in TRS would lie between teachers (within schools), (b) a substantially greater proportion of the TRS variance

would be between teachers (within schools), when compared to both achievement and ability scores, and (c) after controlling for ability, achievement, and student demographics, there would still be considerable between-teacher variance in TRS scores. Substantial residual between-teacher variance would indicate that teachers differ from each other in terms of their rating scale usage for students that have similar profiles in terms of ability, achievement, and demographics. If students' ratings are even partially a function of the teacher to which they were assigned, the same student would be rated higher by one teacher and lower by another, calling into question the comparability of ratings across teachers. This, in turn, undermines the ability of a cut score for admission into a program (or for further screening) to function equitably across teachers. To evaluate this hypothesis, we computed the proportion of residual between-teacher variance that is unaccounted for by ability and achievement as well as the proportion of the total rating scale variance that lies between teachers (within schools) after controlling for ability and achievement (as explained earlier). After quantifying the degree of between-teacher variance in TRS, we then examined the degree to which between-teacher differences in TRS were likely to influence the nomination/identification process. In other words, how much more likely are students to be nominated or identified by a high rating teacher than by a low rating teacher?

Data Sources

For this study, we used data from four individual school districts that collected achievement, ability and TRS data on all students in at least one grade level. Although the specific instruments varied across districts, all participating districts provided ability, math achievement, reading achievement, TRS, teacher and school ID variables, and one or more demographic variables. See Table 1 for a list of the assessment instruments that were used in each of the participating districts. Participating districts varied in terms of their size, number of

schools, geographic location, and student diversity. All the districts included in this study come from states with a legal mandate to identify and/or serve gifted students. Districts include urban and suburban districts from the East coast, the South, and the Midwest. Given the differences across districts in terms of TRS, we conducted analyses separately across districts. Two districts provided data from multiple cohorts of students. District M provided data for three cohorts; however, one cohort was pre- Covid -19, one was during Covid, and one was post- Covid-19. Therefore, we decided to analyze the three cohorts separately. District O provided data for one cohort of second graders and one cohort of third graders. However, because the two grade levels were different, we decided to analyze those two datasets separately as well. Therefore, our analyses for this paper include separate analyses of seven different data files from four different districts. Table 2 describes the sample sizes and sample demographics for all seven data sets. Replicating findings across multiple datasets that vary in terms of TRS, time period, grade level, and district context provides stronger evidence of the generalizability of our results.

Teacher Rating Scales (TRS)

Each of the four districts used a different TRS. District C used the Gifted Rating Scales (GRS; Pfeiffer & Jarosewich, 2003; 2007), a nationally normed teacher rating scale of giftedness for students ages 6-13. The GRS contains 72 items that assess giftedness in six domains: intellectual, academic, motivation, creativity, leadership, and artistic (Pfeiffer & Jarosewich, 2003). The developers report internal consistency reliabilities of .97 or higher, and test-retest reliabilities ranging from .83 to .97. However, reported inter-rater reliabilities were substantially lower, ranging from .64 to .79, indicating a fair amount if inconsistency across teachers in ratings of the same student. District M used the HOPE Scale (Gentry et al., 2015), a brief (11-item) teacher rating scale designed to measure academic and social aspects of giftedness primarily in

elementary-aged students. Reported Cronbach's alpha internal consistency reliability was .96 for the academic subscale and .92 for the academic and social subscales. Peters and Gentry (2012) reported gender and racial/ethnic invariance. There is no research on inter-rater reliability of the HOPE scale. However, Pereira (2021) reported ICCs of .19 and .20 for the Academic and Social subscales, suggesting that nearly 20% of the variance on the HOPE scale lies between teachers. District-made or adapted versions of published TRS are quite common in gifted education. Districts H and O created or substantially adapted their TRS, therefore, no reliability or validity data are available.

Data Analysis

To better examine the behavior of TRS, we fit a series of three-level multilevel models in which students (level-1) were nested within teachers (level-2), who were nested within schools (level-3). Predictor variables included ability, achievement, and demographics (race/ethnicity, FRL, EL, and reported gender). Race/ethnicity contained six categories. Models with race/ethnicity included five dummy coded variables: Black, Latinx, Asian, Native American, and Other (1 = Black, Latinx, Asian, Native American, and Other). FRL status, EL status, and reported gender were also dummy coded (1 = FRL, EL, and female). As requested by the reviewers during the review of the registered report, we ran each model using two different ethnicity codes. In the first set of analyses, we included all five (with White being the omitted group) dummy coded variables for race/ethnicity. In the second set of analyses, we grouped Black, Latinx, and Native Americans into one dummy coded variable called underserved (0 = White, Asian, and Other), given that Black, Latinx, and Native American have been traditionally under-identified as gifted. The results of the two analyses were similar. In this paper, we report the results of the models that include all five race/ethnicity dummy variables, but the models

with the 2-category race/ethnicity variable (underserved) are available in the online appendix (see Table S1).

Centering

As requested by a reviewer during the review of the registered report, for all models that included covariates, we used two separate centering strategies. In strategy one, we group mean centered ability and achievement at level 1 (student level; $X_{1i} = X_{ijk} - \bar{X}_{\cdot jk}$), aggregated them within teachers and centered around the school mean at level 2 (teacher level; $X_{2j} = \bar{X}_{\cdot jk} - \bar{X}_{\cdot\cdot k}$), and aggregated them within schools centered around the grand (sample) mean at level 3 (school level; $X_{3k} = \bar{X}_{\cdot\cdot k} - \bar{X}_{\cdot\cdot\cdot}$). Demographic variables were centered and included at each level in the same way. In strategy two, we grand mean centered all continuous variables at level 1 (student level), did not include them at level 2, and included the grand mean centered school-level aggregates of those grand mean centered variables at level 3 (school level). In this model, the level-1 demographic variables were uncentered at level 1, not included at level 2, and included as grand mean centered school-level aggregates (to create centered proportions of students for each subgroup) at level 3 (the school level). The results section summarizes the results of the group mean centered analyses, as this is the preferred strategy (Hoffman & Walters, 2022; Raudenbush & Bryk, 2002; Rights, 2022; Yaremych et al., 2022).

Because we conducted all analyses using group mean centering and grand mean centering and with the six-category race/ethnicity variable and the two-category race/ethnicity variable, there are four sets of multilevel tables (each containing all five estimated multilevel models) for each of the seven district data sets. Our results were quite similar across the two centering techniques and across the two coding systems for race/ethnicity. All tables are available in the online appendix.

Evaluating Between-Teacher Variance in the TRS

How did we determine whether there was a non-negligible amount of between-teacher variance in GRS after controlling for ability and achievement? Evaluating whether variance components are needed in a multilevel model is less straightforward than evaluating fixed effect parameters (McCoach & Cintron, 2022; McCoach et al., 2022). In addition to examining modified likelihood ratio test described in Snijders and Bosker (2012), we computed the proportion of the total TRS variance that is residual between-teacher variance, after controlling for ability, achievement and student demographics using the following formula: $\frac{\tau_{\pi 00(c)}}{\sigma^2_{(u)}+\tau_{\pi 00(c)}+\tau_{\beta 00(u)}}$ where $\tau_{\pi 00(c)}$ is the between-teacher variance from the conditional model, $\sigma^2_{(u)}$ is the between-student variance from the unconditional (random effects) model, $\tau_{\pi 00(u)}$ is the between-teacher variance from the unconditional (random effects) model, and $\tau_{\beta 00(u)}$ is the between-school variance from the unconditional (random effects) model. This formula resembles the formula for the ICC. The denominator is identical to the denominator from the ICC; however, the numerator is the residual between-student variance from the conditional model. Residual between-teacher variance from the fully specified conditional model representing at least 10% of the total variance from the unconditional model represents a substantial amount of residual between-teacher variance.

Statistical Models

We estimated a set of five multilevel models. Below we describe our model building approach in more detail.

Model 1. Unconditional Three-Level Model

$$TRS_{ijk} = \gamma_{000} + e_{ijk} + r_{0jk} + u_{00k} \quad (1)$$

Model 1, an unconditional model, partitioned the variance into three orthogonal components. Model 1 estimates the proportion of between-student, -teacher, and -school variance on TRS. In addition, we estimated unconditional three-level models for ability, mathematics achievement, and reading achievement to determine whether the proportion of between-teacher variance in the rating scale data was similar to or greater than the proportion of between-teacher variance in achievement and ability scores.

Model 2. Ability and Achievement (Student, Teacher, and School) Predict Rating Scales

$$TRS_{ijk} = \gamma_{000} + \gamma_{100}(\text{Ability}_{1i}) + \gamma_{200}(\text{Ach}_{1i}) + \gamma_{010}(\text{Ability}_{2j}) + \gamma_{020}(\text{Ach}_{2j}) + \gamma_{001}(\text{Ability}_{3k}) + \gamma_{002}(\text{Ach}_{3k}) + e_{ijk} + r_{0jk} + u_{00k} \quad (2)$$

Model 2 included ability and achievement. Because TRS are designed to measure students' abilities and potential for high academic performance, ability and/or achievement should explain a substantial proportion of the between-student variance in the TRS. The residual between-student variance component (after controlling for ability and achievement) is σ^2. $\tau_{\pi 00}$ represents the residual between-teacher variance, after controlling for ability and achievement. If $\tau_{\pi 00}$ is due to between-class differences in achievement or ability, class aggregated ability and achievement should explain the between-teacher variance in the TRS.

Using Model 2, we computed the proportion of between-teacher variance in TRS that was explained by ability and achievement and the proportion of teacher variance in TRS remained after controlling for ability and achievement. The Snijders and Bosker (2012) Modified Likelihood Ratio Test (MLRT) compares the conditional TRS model conditional with a randomly varying intercept to the conditional TRS model with a fixed teacher intercept. A non-statistically significant MLRT indicates that a randomly varying intercept is not needed at the teacher level. In all seven datasets, the MLRT was statistically significant ($p < 0.001$), favoring the model with randomly varying intercepts at the teacher level.

Model 3. Includes Students' Demographic Characteristics

$$TRS_{ijk} = \gamma_{000} + \sum_{p=1}^{P} \gamma_{p00} \left(\mathrm{Var}_{p_{1i}}\right) + \sum_{p=1}^{P} \gamma_{0p0} \left(\mathrm{Var}_{p_{2j}}\right) + \sum_{p=1}^{P} \gamma_{00p} \left(\mathrm{Var}_{p_{3k}}\right) + e_{ijk} + r_{0jk} + u_{00k} \quad (3)$$

Examining Between-Student Differences on Demographic Variables

We conducted a three-model sequence to add demographic variables to Model 2. Model 3a included race/ethnicity. Level-1 demographic slopes represent the differential between TRS of one group of students, as compared to the reference group (holding all other variables constant). A demographic slope effect indicates that students from different demographic groups with the same ability and achievement scores are rated differently by teachers. For instance, if an ethnicity variable explains residual *student-level* variance in student ratings, then teachers tend to rate students of that ethnicity systematically higher or lower, even after controlling for ability and achievement, suggesting potential teacher bias. It could be positive, indicating that teachers tend to give higher scores to students from the subgroup, or it could be negative, indicating that teachers tend to give lower scores to students from the subgroup. For any statistically significant demographic effects, we computed Cohen's *d* effect sizes to evaluate the magnitude and practical importance of the effect.

Model 3b included ethnicity by ability and ethnicity by achievement interactions. Ethnicity by ability and ethnicity by achievement interactions that were not statistically significant were not retained. Ethnicity-by-ability (or –achievement) interactions indicate that changes in student ability or achievement do not impact teacher ratings equally across racial/ethnic groups. For instance, a positive underserved by ability interaction would indicate that that the ability slope is more positive for underserved students, whereas a negative

underserved by ability interaction would indicate that the ability slope is less positive (more negative) for underserved students than for reference students.[2]

Model 3c included the ethnicity variables, any ethnicity by ability and ethnicity by achievement interactions that were statistically significant (p<.01) in Model 3b, and FRL status and EL status for three of the four districts (District C did not provide data on FRL or EL status). Across the seven district datasets, none of the ethnicity by ability and ethnicity by achievement interactions were statistically significant. Therefore, Model 3c, the final model, includes ability, achievement, and all available demographic variables (race/ethnicity, FRL status, EL status, and reported gender), but does not include any interactions.

Additional Robustness Checks and Alternative Modeling Strategies

As requested by one of the reviewers, we used bootstrapping to empirically derive standard errors and confidence intervals and compared these results to the completely analytic results as a sort of robustness check. We also fit the conditional model that included ability, math achievement, and reading achievement using Bayesian estimation (using the brms package in R) as an additional robustness check. The results were quite similar to the original multilevel results. In addition, we fit a mixed effects location scale (MELS) model (Hedeker & Mermelstein, 2022), which allows the residual between-student variance to vary across teachers.

How Much Is the Observed Teacher Variance Likely to Influence Identification Outcomes?

Using the estimate of the between-teacher variance in TRS, we created a set of plausible values for TRS (Raudenbush & Bryk, 2002). Using $\tau_{\pi 00}$ (the residual between-teacher variance

[2] Theoretically, teachers could also differ from each other in terms of the degree to which they systematically give higher or lower ratings to students from different demographic groups. Allowing the demographic slopes (at the student level) to randomly vary at the teacher level would enable estimation of between-teacher variance in ethnicity slopes or in interactions between ethnicity and ability/achievement. Unfortunately, the data are unable to support the estimation of these random effects at the teacher level, especially given that many classrooms are likely to be demographically homogenous.

in the rating scale) from the full multilevel model, we estimated a range of likely values across teachers for student rating scale scores, given a particular prototypical profile. For instance, $\gamma_{000} \pm \sqrt{\tau_{\pi 00}}$ provides a plausible range of TRS scores for average students with teachers who span the range from 1 SD below the average teacher on TRS (16th percentile) to 1 SD above the average teacher on TRS (84th percentile), and provides a 68% plausible interval; $\gamma_{000} \pm 1.645\sqrt{\tau_{\pi 00}}$ provides a plausible range of TRS scores for average students with very low scoring (5th percentile) teachers to very high scoring teachers (95th percentile) and provides a 90% plausible interval. This range of plausible values indicates how much between-teacher differences contribute to variability in TRS and indicates how much TRS scores for a prototypical student might differ across teachers within the same school. We also computed a Cohen's *d* effect size that indicates the expected standard deviation unit change in students' TRS scores per between-teacher standard deviation.

To determine if teacher-level variance in ratings is likely to substantively change identification decisions, we used the lowest between-teacher effect size to create high and low TRS for each student. We then "simulated" conditions in which the student's TRS was completed by a teacher whose scores were 1 SD lower than their current teacher. We computed the mean of ability, achievement, and TRS using the original and modified TRS scores. Using a cut-off score derived from the original data, we determined the 90th percentile for the mean of ability, achievement, and TRS in the original dataset, and used this as the cut-off score for identification as gifted. Using that cut-off score, we determined which students in the district were identified as gifted using the original and modified TRS scores, and we calculated the proportion of overlap across two groups. In this manner, we were able to evaluate the degree to which between-teacher differences in TRS may undermine the overall identification process.

Finally, some districts use TRS as an initial screener to determine who should move to Phase 2 of the identification process. To estimate how many students with high cognitive ability would be missed if TRS were used as the Phase I screener, we computed the percentage of students who scored in the top 10% of the district on cognitive ability but did not score in the top 10, 20, 25, or 30% of the district on TRS.

Results

Fixed Effects

Across the seven datasets, after controlling for ability and achievement, there were no *consistent* effects of demographic variables such as race/ethnicity, FRL, or EL status on students' TRS scores.[3] Prior to controlling for ability and achievement, there were large differences in the TRS means across demographic groups, but these differences were generally explained by between-group differences in ability and achievement. Table 3 contains the Cohen's *d* effect sizes for statistically significant demographic predictors and the proportion of residual (and total) level-1 variance explained by all included demographics. None of the demographic variables were statistically significant in the majority of district datasets, and none of the Cohen's *d* effect sizes for statistically significant slopes exceeded +/- .18 standard deviation units. The FRL slope was statistically significant and negative in three of the six datasets (O2, O3, and H) that contained information on FRL status. Even so, the Cohen's *d* effect size for this effect was less than small, ranging from -.11 (H) to -.15 (O3). O2 and O3 were the only datafiles containing gender. After controlling for the other variables in the model, females had slightly higher TRS than males, but the effect size for this difference was very small (*d* = .13 in O2 and *d* = .11 in

[3] These are the results for analyses that contained all 6 race/ethnicity categories and group mean centered all predictors at levels 1 and 2.

O3). In addition, we computed the proportion of residual level-1 variance and the proportion of total level-1 variance that was explained by the entire collection of demographic variables (race/ethnicity, FRL, EL, and gender). The entire collection of demographic variables explained between .35% (H) and 1.9% (M1) of the residual between-student variance and less than 1% of the total between-student variance in TRS. Therefore, overall, after controlling for ability and achievement, the very small effect sizes and the very small proportion of between-student variance that is uniquely explained by the demographic variables suggest no robust and consistent pattern of teacher bias for or against any particular demographic group. Table 3 contains Cohen's *d* effect sizes and proportions of residual and total between-student variance explained for the final group mean centered model that includes all six race/ethnicity categories (see the online appendix the full multilevel results).

Between-Teacher Variance in TRS

In contrast to our findings above, a consistent pattern emerged across the seven datasets for between-teacher variance in TRS. As expected, most of the variability in TRS (and ability, math achievement, and reading achievement) was between students within classes/schools. However, a substantial percentage of variance in TRS was between teachers (within schools), ranging from a low of 10.5% in C to a high of 24.9% in M2 (see Table 4). The proportion of between-teacher variance in TRS was much larger than the proportion of between-teacher variance in student ability, reading achievement, or math achievement. In fact, as shown in Table 5, the proportion of between-teacher variance in TRS was over twice as large as the proportion of between-teacher variance in student ability or achievement in districts H and C, over three times as large as the proportion of between-teacher variance in student ability or achievement in M1, over five times as large as the between-teacher variance in O2, and well over 10 times as

large in M2 and O3. In addition, the proportion of between-teacher variance in the TRS was always higher than the proportion of between-school variance, and often it was substantially higher. For ability and achievement, this trend was reversed: School generally explained a greater proportion of ability and achievement variance than teacher did.

Residual Between-Teacher Variance in TRS

Across all seven datasets, student ability and achievement explained little to none of the between-teacher variance in TRS; the residual between-teacher variance in TRS was nearly identical to the unconditional between-teacher variance in TRS. After controlling for ability and achievement, the residual between-teacher variance in TRS ranged from .105 to .235 (see Table 5). The percentage of between-teacher variance in TRS that was explained by both ability and achievement ranged from 0% (C and O2) to 14% (in M1). Therefore, little to none of the between-teacher variability in TRS could be explained by between-classroom differences in ability or achievement.

To recap, 10%-25% of the total variance in TRS was between teachers within schools, and between-class differences in ability and achievement explained very little of this between-teacher variance in TRS. Substantial residual between-teacher variance in our full model indicates that ratings of different students within the same classroom share common variance, even after accounting for ability, achievement, and student demographics. The residual between-teacher variance that is not explained by ability and achievement measures (such as class average ability, math achievement, and reading achievement) implies teacher rating effects, indicating that teachers differ from each other in terms of their use of TRS. Taken as a whole, this pattern of results suggests that teachers do contribute a substantial amount of what appears to be

construct irrelevant variance into the TRS, suggesting that the TRS is a measure of (something about) the teacher as well as the student.

How Much Is the Between-Teacher Variance in TRS Likely to Influence Students' Scores on the TRS and the Subsequent Identification Decisions?

Plausible Values. Using the between-teacher SDs, we created 68% plausible values around the mean TRS (see Table 6). For example, the mean of TRS in district C is 64.65, with an overall standard deviation of 20.76. So, we would expect a student with average ability and average achievement to score around 64.65 on the TRS if they had an "average" scoring teacher. However, the between-teacher standard deviation was 6.75. [68% PI: 57.90, 71.45], meaning an average student with a low scoring teacher (a teacher whose class mean TRS scores are one standard deviation below the overall teacher mean), would be expected to score 57.90 points (6.75 points lower). The expected TRS with a high scoring teacher (a teacher whose class mean scores are one standard deviation above the overall teacher mean), would be 71.45 points (6.75 points higher). In other words, the expected score for a student in a low scoring teacher's class would be 6.75 points lower than the expected score for a similar student in an average scoring teacher's class and almost 12.5 points lower than the expected score for the same student in a high scoring teacher's class. The results are similar for the other six data sets (see Table 6).

The between-teacher differences on TRS are substantial. They translate into a .33 SD unit effect size in C and a .50 effect size in M3. So, students of high scoring teachers are advantaged by ⅓ to ½ SD over students of average scoring teachers and by ⅔–1 SD over students of low scoring teachers.

How much do these between-teacher differences really matter? As an illustration, using each student's district standardized TRS, we created a modified TRS by decreasing the observed

TRS by .33 standard deviations. We then evaluated whether each student would have been in the 90th percentile or above on the mean of ability, achievement, and TRS using their current score as well as their lowered score (see Table 7). Overall, 20.44% of the students who are currently identified as gifted (using the mean rule and the current 90% cut-off value) would not be identified as gifted if their TRS were .33 SD lower (see Table 7). Over 30% of the students who are currently identified as gifted would not be identified as gifted if their TRS were .50 SD lower.

As a final illustration, we conducted a descriptive analysis to determine what percentage of students who scored in the top 10% of their district on the cognitive ability test also scored in the top 10%, 20%, 25%, or 30% of their district on the TRS. Table 8 contains these results, both overall, and broken out by district. Overall, less than one-third (32.37%) of the students who score in the top 10% of their district on ability also score in the top 10% of the district on TRS. In fact, over 16% of the students who scored in the top 10% on the TRS scored below their district's average on the cognitive ability test. Just over half (56.18%) of these high ability students score in the top 20% of the district on the TRS. In other words, in our sample, using those who score in the top 20% on the TRS as a screener for further testing with a cognitive ability test misses nearly half of the students who score in the top 10% on cognitive ability. Even when considering the top 30% on TRS, we would miss almost 30% of students who score in the top 10% on ability, as only 70.29% of students in the top 10% of their district on ability score in the top 30% of their district on the TRS.

Correlations among TRS, Ability, and Achievement

Tables 9 and 10 report the correlations among ability, math achievement, reading achievement, and TRS overall (Table 9), and at the between-student, between-teacher, and

between-school levels (Table 10). As expected, the correlations among ability and achievement are larger than the correlations between TRS and the other variables. Disaggregating these correlations by level, the correlation between TRS and ability and achievement is higher between students than it is between classes, even though the between-teacher correlation among the ability and achievement measures is generally as high as or higher than the between-student correlations. Within a given classroom, the relationship between TRS and ability and/or achievement scores is fairly strong: Teachers do tend to rate higher ability and higher achieving students higher on the TRS. However, at the teacher level, teachers' mean TRS scores are far more weakly correlated with their class average ability and achievement. In other words, within schools, teachers with higher ability/higher achieving students (on average) do tend to have higher TRS, but these relationships are far weaker than they are at the between-student level. This is another indication that something other than the average ability and/or average achievement of the class is influencing relative class standing on the TRS.

The results of the mixed-effects location scale (MELS) models were quite similar to the results of the standard multilevel models in terms of the magnitudes of the between-teacher variances. However, the MELS model did indicate that even after controlling for ability and achievement there are between-teacher differences in residual between-student variances. This means that teachers differ not only in terms of their overall level on the TRS, but also in terms of the degree to which they distinguish between students (distribute scores) on the TRS. After controlling for ability and achievement, the relationship between teacher means and teacher variances was positive in four of the datasets (H = .19, O2 = .38, O3 = .64, and M1 = .51), indicating that after controlling for ability and achievement, teachers with higher average TRS also had more variable TRS. In the other three datasets, this correlation was not statistically

significant, indicating no detectable relationship between teachers' average TRS and their variability on the TRS.

Discussion

Between-Teacher Variance in TRS

Hypothesis 1

Although the proportion of variance explained by teacher varied widely (from 10.5% to almost 25%), as we hypothesized, a substantial proportion of the variance in TRS (10%-25%) fell between teachers within schools.

Hypothesis 2

Compared to achievement and ability scores, a substantially greater proportion of the variance in TRS was between teachers within schools. In addition, there was more between-teacher variability than between-school variability in the TRS.

Hypothesis 3

Controlling for ability and achievement did not eliminate between-teacher variance in TRS. In fact, ability, achievement, and demographics explained little to none of the between-teacher variance in TRS (although these variables did explain substantial proportions of between-student variance and between-school variance in the TRS). Even after controlling ability, achievement, and demographics, the between-teacher variance in TRS was so large that it represented an effect size of a third to one half standard deviation. Taken together, these findings provide compelling evidence that TRS are a measure of the teacher as well as the student, and that students' scores on the TRS are influenced by the teacher who completes the scale. These differences are large enough to have real world ramifications in terms of who is identified as gifted, even when the TRS is used as one of multiple identification measures. Although we found

no strong evidence of consistent teacher bias against a particular population, differences between teachers in their use of TRS advantage students in some classrooms and disadvantage students in others.

Students' ratings are partially a function of the teacher to which they were assigned (i.e., the same student would be rated higher by one teacher and lower by another), calling into question the comparability of ratings across teachers and making it impossible to set a cut score for admission into a program (or for further screening) that functions equitably across teachers[4]. The between-teacher standard deviations are large: they represent one-third to one-half of the overall standard deviation on the TRS. Between-teacher differences of this magnitude can have real consequences for who gets identified as gifted. Obviously, the real-world effect depends not just on the magnitude of the differences between teachers, but also on the way in which the TRS is used as part of the overall identification system. We provided two demonstrations to illustrate the potential consequences.

If teachers detected important components of giftedness that both ability and achievement tests miss and those qualities vary across students, this would manifest as residual ***student-level*** variance in models that condition on ability and achievement, ***not*** residual between-teacher variance. Between-teacher variance indicates that any unobserved factor explaining such variability functions at the ***teacher-level***—meaning that it is a function of the teacher (or the classroom), not the student.

---

[4] Unless the scores are standardized or centered within teacher. However, this is not commonly done. Our data are unusual in that teachers have completed rating scales on all students in their classes. However, more commonly, districts have teachers complete rating scales only for students who have been (or are being) referred for identification as gifted. Such a system makes it impossible to standardize or center teacher rating scale scores under normal conditions.

To be clear, we cannot definitively establish that the residual between-teacher variance in the rating scales is solely a result of teachers' differential usage of the rating scale. There are other possibilities: perhaps some teachers are more engaging, so students in their classes are more enthusiastic and academically engaged, leading to higher ratings on the gifted rating scale (holding ability/achievement constant). However, all plausible explanations must reside at the classroom (teacher) level and cannot be explained by between-class differences in ability or achievement. Teacher-level variance suggests that a student's rating is at least partially a function of the teacher to whom a student was assigned (and who completed the rating scale). In other words, the same student, assigned to a different teacher, would have likely received a different rating.

The degree to which the teacher mediates the rating scale score represents both a reliability and a validity issue. Teacher-specific variance adds "noise" (non-target specific variance) to the rating scale score, which results in a score that is less precise. Even more importantly, if certain students receive higher ratings than other students solely as a function of the teacher to whom they were assigned, the validity of the inferences that educators make from the rating scale data is compromised.

Interestingly, using TRS introduces a type of "bias" (inequity) into the identification process, but it is not based on race/ethnicity/demographics. Instead, the "bias" in student identification is a function of the teacher. Students whose teachers tend to give higher scores on the TRS are more likely to be identified as gifted than students whose teachers give lower scores on the TRS. As we have demonstrated, these effects are large enough to influence who is ultimately identified as gifted.

Given these findings, we have several recommendations for school districts that use TRS as part of the identification process and for researchers in this area:

1. Never use TRS as the sole universal screening instrument to determine which students move forward to a second stage gifted identification process.
2. When selecting a TRS, consider the proportion of between-teacher variance as an additional source of error. Reported reliability estimates do not take between-teacher variance into account. In fact, between-teacher variance in TRS is likely to artificially increase the reported reliability coefficient for TRS. In our study, the GRS had the lowest between-teacher variance. However, future researchers should examine whether this result is replicated in other datasets. In addition, future research on TRS should always report the proportion of between-teacher variance on the TRS, as this is an important, but generally overlooked aspect of its psychometric adequacy.
3. Districts using TRS as part of their identification process should provide frequent professional development for teachers to try to standardize their usage of the TRS as much as possible, which should help to decrease the proportion of between-teacher variance on the TRS.
4. Districts using TRS as part of their screening/identification process should have a clear idea about the purpose of including a TRS in the identification process, and they should examine the TRS to ensure that it is designed to elicit the kind of information they are seeking.

Centering students' scores around their class/teacher means, thereby creating a teacher centered (or teacher normed) TRS score might seem to provide one potential solution to the problem of between-teacher variance. However, we are hesitant to recommend such an option

for three reasons. First, the very small number of students per teacher makes such a process unstable/unreliable. Second, using teacher centered TRS does not control for ability and achievement, and there are real between-teacher differences in class composition. The combination of the within-class sample size and the between-class differences in classroom composition could make classroom-normed TRS equally problematic. Finally, teacher centering the TRS does nothing to address the issue of between-teacher variability in between-student variance on the TRS.

Limitations

There are several limitations of the current study. These include (a) the lack of item-level data, (b) the lack of detailed school-level data on such as the assignment of students to classes, and (c) the lack of teacher-level covariates that could help to explain between-teacher variability in the use of the TRS. Future research should explore whether certain teacher characteristics help to explain the between-teacher variance in TRS. In addition, future research should determine the degree to which professional development on the TRS can decrease the between-teacher variance on TRS.

References

Anthony, C. J., Styck, K. M., Cooke, E., Martel, J. R., & Frye, K. E. (2022). Evaluating the impact of rater effects on behavior rating scale score validity and utility. *School Psychology Review, 51*(1), 25–39. https://doi.org/10.1080/2372966X.2020.1827681.

American Education Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.

Callahan, C. M., Moon, T. R., & Oh, S. (2013). *Status of elementary gifted programs: 2013*. https://www.nagc.org/sites/default/files/key%20reports/ELEM%20school%20GT%20Survey%20Report.pdf

Council for Exceptional Children. (2020). *Initial practice based professional preparation standards for special educators*. https://exceptionalchildren.org/standards/initial-practice-based-professional-preparation-standards-special-educators

Dai, D. Y., & Chen, F. (2013). Three paradigms of gifted education: In search of conceptual clarity in research and practice. *Gifted Child Quarterly, 57*(3), 151–168. https://doi.org/10.1177/0016986213490020

Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education, 77*(4), 267–282. https://doi.org/10.1177/003804070407700401

Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *ILR Review, 48*(3), 547–561. https://doi.org/10.1177/001979399504800312

Eklund, K., Kilgus, S., von der Embse, N., Beardmore, M., & Tanner, N. (2017). Use of universal screening scores to predict distal academic and behavioral outcomes: A multilevel approach. *Psychological Assessment, 29*(5), 486–499. https://doi.org/10.1037/pas0000355

Ford, D. Y., & King, R. A. (2014, Fall). Desegregating gifted education for under-represented Black and Hispanic students: Equity promotes equality. *Teaching for High Potential, 1*, 13-16.

Fultz, M., Lara-Alecio, R., Irby, B. J., & Tong, F. (2013). The Hispanic Bilingual Gifted Screening Instrument: A validation study. *National Forum of Multicultural Issues Journal, 10*(1), 1-26. http://www.nationalforum.com/Electronic%20Journal%20Volumes/Fultz,%20Monica%20The%20Hispanic%20Bilingual%20Gifted%20Screening%20Instrument%20NFMIJ%20V10%20N1%202013.pdf

Gentry, M., Pereira, N., Peters, S. J., McIntosh, J., & Fugate, C. M. *HOPE Teacher Rating Scale*: *Administration manual*. (2015). Prufrock Press.

Gentry, M., Whiting, G., & Gray, A. M. (2022). Systemic inequities in identification and representation of Black youth with gifts and talents: Access, equity, and missingness in urban and other school locales. *Urban Education*. Advance online publication. https://doi.org/10.1177/00420859221095000

Grissom, J. A., & Redding, C. (2016). Discretion and disproportionality: Explaining the underrepresentation of high-achieving student of color in gifted programs. *AERA Open, 2*(1), 1-25. https://doi.org/10.1177/2332858415622175

Grissom, J. A., Redding, C., & Bleiberg, J. F. (2019). Money over merit?: Socioeconomic gaps in receipt of gifted services. *Harvard Education Review, 89*(3), 337-369. https://doi.org/10.17763/1943-5045-89.3.337

Hamilton, R., McCoach, D. B., Tutwiler, M. S., Siegle, D., Gubbins, E. J., Callahan, C. M., Brodersen, A. V., & Mun, R. U. (2018). Disentangling the roles of institutional and individual poverty in the identification of gifted students. *Gifted Child Quarterly, 62*(1), 6–24. https://doi.org/10.1177/0016986217738053

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245-270. https://doi.org/10.1080/02671520500193744

Harradine, C. C., Coleman, M. R. B., & Winn, D.-M. C. (2014). Recognizing academic potential in students of color: Findings of U-STARS~PLUS. *Gifted Child Quarterly, 58*(1), 24–34. https://doi.org/10.1177/0016986213506040

Hedeker, D., & Mermelstein, R.J. (2022). Modeling variation in intensive longitudinal data. In A. A. O'Connell, D. B. McCoach, & B. A. Bell (Eds.) *Multilevel modeling methods with introductory and advanced applications* (pp. 421-457). Charlotte: Information Age Publishing.

Hertzog, N. B. (2009). The arbitrary nature of giftedness. In L. V. Shavinina (Ed.), *International handbook on giftedness* (pp. 205-214). Springer Science.

Hoffman, L., & Walters, R. W. (2022). Catching up on multilevel modeling. *Annual Review of Psychology, 73*, 659–689. https://doi.org/10.1146/annurev-psych-020821-103525

Hosp, J. L., & Reschly, D. J. (2003). Referral rates for intervention or assessment: A meta-analysis of racial differences. *Journal of Special Education, 37*(2), 67–80. https://doi.org/10.1177/00224669030370020201

Long, D., McCoach, D. B., Siegle, D., Callahan, C. M., & Gubbins, E. J. (2023). Inequality at the starting line: Under-representation in gifted identification and disparities in early achievement. *AERA Open, 9*(1), 1–25. https://doi.org/10.1177/23328584231171535

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*(2), 78-102. https://doi.org/10.1080/10627190903039429

Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated with teachers' ratings of prekindergartners' relationships and behaviors. *Journal of Psychoeducational Assessment, 24*(4), 367-380. https://doi.org/10.1177/0734282906290594

Matthews, M. S. (2018). Utilizing non-test assessments in identifying gifted and talented learners. In C. M. Callahan & H. L. Hertberg-Davis (Eds.) *Fundamentals of gifted education: Considering multiple perspectives*. (pp. 135-145). Routledge.

McBee, M. T., Peters, S. J., & Miller, E. M. (2016). The impact of the nomination stage on gifted program identification: A comprehensive psychometric analysis. *Gifted Child Quarterly, 60*(4), 258-278. https://doi.org/10.1177/0016986216656256

McCoach, D., B. & Cintron, D. W. (2022). *Quantitative toolkit series: Vol. 9. Introduction to modern modeling methods*. SAGE.

McCoach, D. B., Newton, S. D., & Gambino, A. J. (2022). Evaluating the fit and adequacy of multilevel models. In A.A. O'Connell, D. B. McCoach, & B. A. Bell (Eds.) *Multilevel modeling methods with introductory and advanced applications.* Information Age Press.

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Exploring student characteristics that influence teacher judgements of student ability. *Teaching and Teacher Education 65*, 48-60. https://doi.org/10.1016/j.tate.2017.02.021

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher 44*(5), 278–292. https://doi.org/10.3102/0013189X15591157

Miller, E. M. (2009). The effect of training in gifted education on elementary classroom teachers' theory-based reasoning about the concept of giftedness. *Journal for the Education of the Gifted, 33*(1), 65–105. https://doi.org/10.1177/016235320903300104

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-421. https://pubmed.ncbi.nlm.nih.gov/14523257/

Pedersen, B., Makel, M. C., Rambo-Hernandez, K. E., Peters, S. J., & Plucker, J. (2023). Most mathematics classrooms contain wide-ranging achievement levels. *Gifted Child Quarterly, 67*(3), 220-234. https://doi.org/10.1177/00169862231166074

Pereira, N. (2021). Finding talent among elementary English learners: A validity study of the HOPE Teacher Rating Scale. *Gifted Child Quarterly, 65*(2), 153-166. https://doi.org/10.1177/0016986220985942

Peters, S. J., & Gentry, M. (2010). Multigroup construct validity evidence of the HOPE Scale: Instrumentation to identify low-income elementary students for gifted programs. *Gifted Child Quarterly, 54*(4), 298–313. https://doi.org/10.1177/0016986210378332

Peters, S. J., & Gentry, M. (2012). Additional validity evidence and across-group equivalency of the *HOPE Teacher Rating Scale*. *Gifted Child Quarterly,* 57(2), 85-100. https://doi.org/10.1177/0016986212469253

Peters, S. J., Gentry, M., Whiting, G. W., & McBee, M. T. (2019). Who gets served in gifted education? Demographic proportionality and a call for action. *Gifted Child Quarterly, 63*(4), 273-287. https://doi.org/10.1177/0016986219833738

Pfeiffer, S. I., & Jarosewich, T. (2003). *Gifted Rating Scales*. Harcourt Assessment.

Pfeiffer, S. I., & Jarosewich, T. (2007). The Gifted Rating Scales-School Form: An analysis of the standardization sample based on age, gender, race, and diagnostic efficiency. *Gifted Child Quarterly, 51*(1), 39–50. https://doi.org/10.1177/0016986206296658

Rambo-Hernandez, K. E., Brigandi, C., Amin, S., & Spillane, N. (2023). Who gets identified? The consequences of variability in teacher ratings and combination rules for determining eligibility for gifted services for young children. *Journal for the Education of the Gifted, 46*(2), 111-139. https://doi.org/10.1177/01623532231162612

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). Sage Publications.

Rights, J. D. (2022). Aberrant distortion of variance components in multilevel models under conflation of level-specific effects. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000514.supp

Ryser, G. R., & McConnell, K. (2004). *Scales for Identifying Gifted Students*. Prufrock Press.

Siegle, D., McCoach, D.B., Gubbins, E.J., Long, D., & Hamilton, R. (2018, January 9-10). *Gifted identification gap: When just as good is not good enough.* Relevance & Rigor: Creating the Future of Education Research—Annual Principal Investigators Meeting of the Institute for Education Science. Arlington, VA. https://ncrge.uconn.edu/wp-content/uploads/sites/982/2018/01/Gifted-Identification-Gap_Siegle_Jan-10_1330-1.pdf

Splett, J. W., Smith-Millman, M., Raborn, A., Brann, K. L., Flaspohler, P.D., & Maras, M. A. (2018). Student, teacher, and classroom predictors of between-teacher variance of students' teacher-rated behavior. *School Psychology Quarterly, 33*(3), 460-468. https://doi.org/10.1037/spq0000241

Starck, J. G., Riddle, T., Sinclair, S., & Warikoo, N. (2020). Teachers are people too: Examining the racial bias of teachers compared to other American adults. *Educational Researcher, 49*(4), 273–284. https://doi.org/10.3102/0013189X20912758

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd Edition). Sage

VanTassel-Baska, J. L. (2008). Epilogue: What do we know about identifying and assessing the learning of gifted students? In J. L. VanTassel-Baska (Ed.), *Alternative assessment with gifted and talented students* (pp. 309-319). Prufrock Press.

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review, 34*(4), 454-474. https://doi.org/10.1080/02796015.2005.12088009

Wind, S. A. (2020). Monotonicity as a nonparametric approach to evaluating rater fit in performance assessments. *Measurement: Interdisciplinary Research and Perspectives*, *18*(3), 124–141. https://doi.org/10.1080/15366367.2020.1711679

Yaremych, H. E., Preacher, K. J., & Hedeker, D. (2023). Centering categorical predictors in multilevel models: Best practices and interpretation. *Psychological Methods, 28(3),* 613–630. https://doi.org/10.1037/met0000434.supp

Table 1

Assessments Used in Each District

District	Grade	School Year	Ability	Achievement	TRS
C	2nd	2019-2020	CogAT	MAP Test	GRS
H	K	2018-2019	CogAT-Nonverbal	IOWA Test	District-made Scale
M	2nd	2019-2020, 2021-2022, 2022-2023	CogAT	MAP Test	HOPE Scale
O	2nd & 3rd	2021-2022	InView	MAP Test	District-made Scale

Table 2

Descriptive Statistics for Analysis Sample by District Datafile

Variable	District (*N*)						
	C (8,685)	H (11,892)	M1 (2,036)	M2 (1,859)	M3 (1,832)	O2 (2,618)	O3 (2,176)
TRS	64.65 (20.76)	57.52 (25.09)	35.33 (10.94)	35.64 (11.37)	25.30 (11.00)	3.28 (2.80)	2.87 (2.76)
Ability	101.30 (14.54)	92.61 (15.34)	109.51 (14.34)	103.46 (14.07)	101.78 (12.48)	334.75 (61.91)	368.39 (54.16)
Math	0.06 (0.92)	-0.29 (1.05)	180.76 (12.42)	179.54 (12.50)	180.75 (13.27)	0.03 (0.95)	0.04 (0.94)
Reading	0.06 (0.92)	-0.25 (0.93)	183.65 (16.27)	174.49 (17.42)	176.32 (17.93)	0.03 (0.93)	0.04 (0.92)
White	0.29	0.11	0.63	0.61	0.59	0.26	0.23
Black	0.32	0.20	0.12	0.13	0.13	0.20	0.24
Latinx	0.27	0.64	0.10	0.09	0.11	0.37	0.36
Asian	0.09	0.05	0.08	0.07	0.07	0.08	0.09
Native Am.	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Other Race	0.03	0.01	0.08	0.09	0.10	0.07	0.07
Underserved	0.59	0.83	0.21	0.22	0.24	0.58	0.61
FRL Status	–	0.80	0.24	0.24	0.28	0.71	0.75
EL Status	–	0.45	0.14	0.13	0.14	0.28	0.25
Gender	–	–	–	–	–	0.50	0.50

Note. Numbers in parentheses under district names are their sample sizes. Numbers in table are means (with standard deviations in parentheses). TRS = Teacher Rating Scale; Native Am. = Native American; FRL = Free/Reduced Lunch; EL = English Learner.

Table 3

*Cohen's d Effect Sizes for Statistically Significant Demographic Predictors and the Proportion of Level-1 Total (Residual) Variance Explained by **All** Included Demographics*

District	Effect Size (Demographics)	Percentage of Residual (Total) Level-1 Variance Explained
C	-0.11 (Black)	0.52% (0.21%)
H	0.08 (Black)[a], -0.11 (FRL)	0.35% (0.21%)
M1	-0.17 (Asian)	1.90% (0.80%)
M2	-0.14 (EL)[a]	0.75% (0.37%)
M3	N/A	1.06% (0.54%)
O2	-0.12 (FRL), -0.18 (EL), 0.13 (Gender)	1.16% (0.60%)
O3	0.15 (Latinx), -0.14 (FRL), 0.11 (Gender)	1.20% (0.80%)

Note. [a] indicates that these two slopes were statistically significant ($p<.01$) in using the standard errors from the standard multilevel model but were not statistically significant when using the bootstrapped results.

Table 4

Intraclass Correlations for Each Outcome by District

District	Level	Teacher Rating Scale	Ability	Math Achievement	Reading Achievement
C	Student (N = 8,685)	0.809	0.780	0.729	0.723
	Teacher (J = 587)	0.104	0.023	0.042	0.038
	School (K = 109)	0.087	0.197	0.229	0.239
H	Student (N = 11,892)	0.617	0.811	0.766	0.763
	Teacher (J = 1,013)	0.246	0.070	0.120	0.117
	School (K = 166)	0.137	0.119	0.114	0.120
M1	Student (N = 2,036)	0.778	0.859	0.914	0.920
	Teacher (J = 92)	0.222	0.063	0.045	0.050
	School (K = 19)	0.000	0.078	0.041	0.030
M2	Student (N = 1,859)	0.751	0.942	0.922	0.950
	Teacher (J = 90)	0.249	0.006	0.006	0.018
	School (K = 19)	0.000	0.052	0.072	0.032
M3	Student (N = 1,832)	0.866	0.942	0.935	0.969
	Teacher (J = 89)	0.119	0.000	0.002	0.000
	School (K = 20)	0.015	0.058	0.063	0.031
O2	Student (N = 2,618)	0.803	0.846	0.836	0.838
	Teacher (J = 171)	0.137	0.018	0.020	0.024
	School (K = 60)	0.060	0.136	0.144	0.138
O3	Student (N = 2,176)	0.797	0.856	0.818	0.817
	Teacher (J = 153)	0.160	0.004	0.003	0.004
	School (K = 56)	0.043	0.140	0.179	0.179

Note. The difference in the proportion of between teacher variance in the TRS is statistically significantly higher than the proportion of between teacher variance in ability, math achievement, and reading achievement in every district ($p < .01$).

Table 5

Proportion of Teacher Rating Scale Variance that was Unexplained Between Teacher Variance Across Models Using Group Mean Centering Strategy and 6-Category Race

District	Model 1	Model 2	Model 3a	Model 3b	Model 3c
C	0.104	0.106	0.106	0.106	0.106
H	0.246	0.242	0.239	0.239	0.239
M1	0.222	0.196	0.186	0.186	0.192
M2	0.249	0.231	0.235	0.235	0.224
M3	0.119	0.133	0.137	0.137	0.132
O2	0.137	0.156	0.142	0.142	0.143
O3	0.160	0.167	0.144	0.144	0.140

Table 6

TRS Mean and SD, Between-Teacher SD, Teacher Effect Size, and 68% Plausible Values for an Average Student as a Function of Teacher

District	TRS Mean	TRS SD	Between-Teacher SD	Effect Size (Teacher)	68% Plausible Values
C	64.65	20.76	6.75	0.33	[57.90, 71.45]
H	57.52	25.09	12.39	0.49	[45.13, 69.91]
M1	35.33	10.94	4.77	0.44	[30.59, 40.33]
M2	35.41	11.41	5.63	0.49	[30.09, 41.36]
M3	25.34	11.05	5.54	0.50	[30.14, 41.23]
O2	3.28	2.80	1.07	0.38	[2.21, 4.34]
O3	2.87	2.76	1.06	0.38	[1.85, 3.97]

Table 7

Comparison of Identified Students when Students' TRS is Decreased by 0.33 SD Units

District	Still ID	Not ID	Current	% Not ID
C	708	163	871	18.71%
H	950	272	1,222	22.26%
M1	164	43	207	20.77%
M2	157	31	188	16.49%
M3	145	43	188	22.87%
O2	209	53	262	20.23%
O3	178	40	218	18.35%
Total	2,511	645	3,156	20.44%

Note. The Current column contains the number of students who would currently be identified if the district were to identify the top 10% of students on the mean of ability, achievement, and TRS. The Still ID column is the number of students who would still be identified if their TRS were decreased by 0.33 SD units. The Not ID column contains the number of students who would no longer be identified if their TRS were decreased by 0.33 SD units.

Table 8

Percentage of Students Who Are in the Top 10% of Their Districts on Cognitive Ability Who Score in the Top 10, 20, 25, and 30% of Their Districts on the TRS

District	Top 10% TRS	Top 20% TRS	Top 25% TRS	Top 30% TRS
C	39.11%	64.56%	72.56%	78.33%
H	26.50%	51.54%	57.94%	64.34%
M1	35.41%	58.37%	67.46%	74.64%
M2	35.00%	51.67%	58.89%	71.67%
M3	36.22%	55.14%	62.70%	71.89%
O2	36.26%	62.60%	72.52%	78.63%
O3	24.77%	42.66%	49.08%	54.13%
Overall	32.37%	56.18%	63.61%	70.29%

Table 9

Level-Specific Correlations between Assessments by District

District	Variable	Student Level				Teacher Level				School Level			
		TRS	Ability	Math	Reading	TRS	Ability	Math	Reading	TRS	Ability	Math	Reading
C	TRS	1				1				1			
	Ability	0.594	1			0.402	1			0.774	1		
	Math	0.735	0.691	1		0.443	0.781	1		0.743	0.920	1	
	Reading	0.760	0.666	0.924	1	0.432	0.768	0.938	1	0.764	0.931	0.987	1
		(N = 8,685)				(J = 587)				(K = 109)			
H	TRS	1				1				1			
	Ability	0.402	1			0.213	1			0.611	1		
	Math	0.553	0.498	1		0.225	0.530	1		0.632	0.820	1	
	Reading	0.579	0.468	0.679	1	0.269	0.553	0.769	1	0.655	0.822	0.913	1
		(N = 11,892)				(J = 1,013)				(K = 166)			
M1	TRS	1				1				1			
	Ability	0.569	1			0.463	1			0.194	1		
	Math	0.662	0.717	1		0.442	0.853	1		0.473	0.764	1	
	Reading	0.636	0.612	0.726	1	0.459	0.863	0.859	1	0.214	0.866	0.840	1
		(N = 2,036)				(J = 92)				(K = 19)			
M2	TRS	1				1				1			
	Ability	0.565	1			0.218	1			0.265	1		
	Math	0.660	0.717	1		0.394	0.711	1		0.263	0.915	1	
	Reading	0.638	0.568	0.730	1	0.225	0.647	0.747	1	0.285	0.879	0.892	1
		(N = 1,859)				(J = 90)				(K = 19)			
M3	TRS	1				1				1			
	Ability	0.587	1			0.132	1			0.189	1		
	Math	0.670	0.737	1		0.245	0.634	1		0.192	0.854	1	
	Reading	0.659	0.603	0.754	1	0.165	0.396	0.700	1	0.201	0.801	0.902	1
		(N = 1,832)				(J = 89)				(K = 20)			
O2	TRS	1				1				1			
	Ability	0.625	1			0.242	1			0.252	1		
	Math	0.647	0.841	1		0.141	0.773	1		0.334	0.948	1	
	Reading	0.662	0.834	0.947	1	0.149	0.772	0.954	1	0.346	0.956	0.992	1
		(N = 2,618)				(J = 171)				(K = 60)			
O3	TRS	1				1				1			
	Ability	0.494	1			0.239	1			0.195	1		
	Math	0.563	0.798	1		0.205	0.664	1		0.190	0.938	1	
	Reading	0.575	0.795	0.940	1	0.193	0.675	0.935	1	0.178	0.943	0.988	1
		(N = 2,176)				(J = 153)				(K = 56)			

Note. TRS = Teacher Rating Scale. Student-level variables are centered within teachers. Teacher-level variables are teacher means centered within schools. School-level variables are school means. N = number of students; J = number of teachers; K = number of schools.

Table 10

Marginal Correlations between Assessments by District

District	Variable	TRS	Ability	Math	Reading
C	TRS	1			
	Ability	0.597	1		
	Math	0.697	0.747	1	
	Reading	0.716	0.732	0.941	1
H	TRS	1			
	Ability	0.400	1		
	Math	0.500	0.550	1	
	Reading	0.529	0.533	0.729	1
M1	TRS	1			
	Ability	0.518	1		
	Math	0.610	0.729	1	
	Reading	0.581	0.647	0.742	1
M2	TRS	1			
	Ability	0.488	1		
	Math	0.577	0.730	1	
	Reading	0.550	0.587	0.738	1
M3	TRS	1			
	Ability	0.526	1		
	Math	0.604	0.741	1	
	Reading	0.597	0.605	0.759	1
O2	TRS	1			
	Ability	0.539	1		
	Math	0.558	0.855	1	
	Reading	0.572	0.849	0.955	1
O3	TRS	1			
	Ability	0.428	1		
	Math	0.473	0.816	1	
	Reading	0.479	0.815	0.949	1

Note. TRS = Teacher Rating Scale.

Appendix A

Table A1

Three-Level Unconditional Model Results for Each Assessment in District C

District C	TRS	CogAT	Math	Reading
Random Effects				
Level 1				
Sigma-square				
Coefficient	349.00	165.61	0.63	0.62
Std. error	5.48	2.60	0.01	0.01
Intercept				
Coefficient	64.01**	100.29**	-0.01	-0.01
Std. error	0.69	0.65	0.04	0.05
Level 2				
Tau				
Coefficient	44.72	4.84	0.04	0.03
Std. error	4.52	1.06	0.01	0.01
Level 3				
Tau				
Coefficient	37.35	41.80	0.20	0.20
Std. error	7.04	6.21	0.03	0.03
Model fit				
N	8685	8685	8685	8685
Loglik. H0	.	.	.	.
LogLikelihood	-38126.57	-34757.87	-10612.49	-10533.95
df	4	4	4	4
AIC	76261.14	69523.75	21232.97	21075.9
BIC	76289.41	69552.03	21261.25	21104.18
R^2: Prop. of Variance Explained by...				
level-2 random intercept	0.104	0.023	0.042	0.038
level-3 random intercept	0.087	0.197	0.229	0.239
all random intercepts	0.190	0.220	0.271	0.277

*$p < 0.01$. **$p < 0.001$.

Table A2

Three-Level Unconditional Model Results for Each Assessment in District H

District H	TRS	CogAT	Math	Reading
Random Effects				
Level 1				
Sigma-square				
Coefficient	386.86	190.73	0.84	0.66
Std. error	5.22	2.56	0.01	0.01
Intercept				
Coefficient	56.47**	91.69**	-0.35**	-0.29**
Std. error	0.88	0.46	0.03	0.03
Level 2				
Tau				
Coefficient	154.57	16.36	0.13	0.10
Std. error	10.34	1.75	0.01	0.01
Level 3				
Tau				
Coefficient	86.15	27.87	0.13	0.10
Std. error	13.57	3.84	0.02	0.02
Model fit				
N	11892	11892	11892	11892
Loglik. H0	.	.	.	.
LogLikelihood	-53147.53	-48546.89	-16416.22	-14993.86
df	4	4	4	4
AIC	106303.1	97101.79	32840.43	29995.72
BIC	106332.6	97131.32	32869.97	30025.26
R^2: Prop. of Variance Explained by...				
level-2 random intercept	0.246	0.070	0.120	0.117
level-3 random intercept	0.137	0.119	0.114	0.120
all random intercepts	0.384	0.188	0.235	0.237

*$p < 0.01$. **$p < 0.001$.

Table A3

Three-Level Unconditional Model Results for Each Assessment in District M1

District M1	TRS	CogAT	Math	Reading
Random Effects				
Level 1				
Sigma-square				
Coefficient	92.94	177.39	141.41	244.07
Std. error	3.81	5.71	4.55	7.85
Intercept				
Coefficient	35.33**	109.49**	180.70**	183.65**
Std. error	0.58	1.04	0.69	0.83
Level 2				
Tau				
Coefficient	26.59	13.10	7.04	13.33
Std. error	4.61	3.92	2.50	4.37
Level 3				
Tau				
Coefficient	0.00	16.09	6.32	8.05
Std. error	0.00	6.84	3.09	4.46
Model fit				
N	2036	2036	2036	2036
Loglik. H0	.	.	.	.
LogLikelihood	-7592.318	-8217.768	-7973.723	-8529.277
df	4	4	4	4
AIC	15192.64	16443.54	15955.45	17066.55
BIC	15215.11	16466.01	15977.92	17089.03
R^2: Prop. of Variance Explained by...				
level-2 random intercept	0.222	0.063	0.045	0.050
level-3 random intercept	0.000	0.078	0.041	0.030
all random intercepts	0.222	0.141	0.086	0.081

*$p < 0.01$. **$p < 0.001$.

Table A4

Three-Level Unconditional Model Results for Each Assessment in District M2

District M2	TRS	CogAT	Math	Reading
Random Effects				
Level 1				
Sigma-square				
Coefficient	96.27	186.69	144.32	289.06
Std. error	3.24	6.28	4.85	9.72
Intercept				
Coefficient	35.71**	103.43**	179.52**	174.47**
Std. error	0.64	0.81	0.83	0.86
Level 2				
Tau				
Coefficient	31.94	1.13	0.98	5.44
Std. error	5.46	1.72	1.37	3.34
Level 3				
Tau				
Coefficient	0.00	10.32	11.19	9.70
Std. error	0.00	4.16	4.29	4.74
Model fit				
N	1859	1859	1859	1859
Loglik. H0	.	.	.	.
LogLikelihood	-6974.97	-7519.231	-7283.157	-7929.818
df	4	4	4	4
AIC	13957.94	15046.46	14574.31	15867.64
BIC	13980.05	15068.57	14596.42	15889.75
R^2: Prop. of Variance Explained by...				
level-2 random intercept	0.249	0.006	0.006	0.018
level-3 random intercept	0.000	0.052	0.072	0.032
all random intercepts	0.249	0.058	0.078	0.050

*$p < 0.01$. **$p < 0.001$.

Table A5

Three-Level Unconditional Model Results for Each Assessment in District M3

District M3	TRS	CogAT	Math	Reading
Random Effects				
Level 1				
Sigma-square				
Coefficient	105.23	147.03	166.30	312.73
Std. error	3.57	4.89	5.64	10.41
Intercept				
Coefficient	25.34**	102.00**	181.06**	176.53**
Std. error	0.57	0.75	0.83	0.84
Level 2				
Tau				
Coefficient	14.52	0.02	0.31	0.00
Std. error	3.40	0.05	1.44	0.00
Level 3				
Tau				
Coefficient	1.84	9.11	11.26	10.04
Std. error	2.20	3.60	4.83	4.79
Model fit				
N	1832	1832	1832	1832
Loglik. H0	.	.	.	.
LogLikelihood	-6926.672	-7189.41	-7302.981	-7874.315
df	4	4	4	4
AIC	13861.34	14386.82	14613.96	15756.63
BIC	13883.4	14408.87	14636.02	15778.68
R^2: Prop. of Variance Explained by...				
level-2 random intercept	0.119	0.000	0.002	0.000
level-3 random intercept	0.015	0.058	0.063	0.031
all random intercepts	0.135	0.058	0.065	0.031

*$p < 0.01$. **$p < 0.001$.

Table A6

Three-Level Unconditional Model Results for Each Assessment in District O2

District O2	TRS	InView	Math	Reading
Random Effects				
Level 1				
Sigma-square				
Coefficient	6.36	3280.72	0.76	0.74
Std. error	0.18	93.61	0.02	0.02
Intercept				
Coefficient	3.37**	337.00**	0.06	0.06
Std. error	0.14	3.37	0.05	0.05
Level 2				
Tau				
Coefficient	1.09	68.98	0.02	0.02
Std. error	0.22	37.78	0.01	0.01
Level 3				
Tau				
Coefficient	0.47	527.98	0.13	0.12
Std. error	0.22	129.51	0.03	0.03
Model fit				
N	2618	2618	2618	2618
Loglik. H0	.	.	.	.
LogLikelihood	-6259.313	-14383.45	-3430.714	-3401.63
df	4	4	4	4
AIC	12526.63	28774.91	6869.428	6811.26
BIC	12550.11	28798.39	6892.909	6834.741
R^2: Prop. of Variance Explained by...				
level-2 random intercept	0.137	0.018	0.020	0.024
level-3 random intercept	0.060	0.136	0.144	0.138
all random intercepts	0.197	0.154	0.164	0.162

*$p < 0.01$. **$p < 0.001$.

Table A7

Three-Level Unconditional Model Results for Each Assessment in District O3

District O3	TRS	InView	Math	Reading
Random Effects				
Level 1				
Sigma-square				
Coefficient	6.20	2537.55	0.73	0.70
Std. error	0.20	79.63	0.02	0.02
Intercept				
Coefficient	2.92**	369.45**	0.07	0.07
Std. error	0.13	3.01	0.06	0.06
Level 2				
Tau				
Coefficient	1.25	13.28	0.00	0.00
Std. error	0.25	26.96	0.01	0.01
Level 3				
Tau				
Coefficient	0.33	414.98	0.16	0.15
Std. error	0.25	99.30	0.04	0.04
Model fit				
N	2176	2176	2176	2176
Loglik. H0	.	.	.	.
LogLikelihood	-5186.216	-11670.62	-2809.783	-2770.584
df	4	4	4	4
AIC	10380.43	23349.24	5627.567	5549.169
BIC	10403.17	23371.99	5650.308	5571.91
R^2: Prop. of Variance Explained by...				
level-2 random intercept	0.160	0.004	0.003	0.004
level-3 random intercept	0.043	0.140	0.179	0.179
all random intercepts	0.203	0.144	0.183	0.183

*$p < 0.01$. **$p < 0.001$.

Table A8

Model Results for District C Using Group Mean Centering and 6-Category Race

District C	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.20**	0.18**	0.18**	0.18**
Std. error		0.01	0.01	0.01	0.01
Math					
Coefficient		3.60**	3.26**	3.38**	3.26**
Std. error		0.45	0.46	0.46	0.46
Reading					
Coefficient		12.60**	12.95**	12.99**	12.95**
Std. error		0.44	0.45	0.45	0.45
Black					
Coefficient			-2.29**	-2.22**	-2.29**
Std. error			0.43	0.44	0.43
Latinx					
Coefficient			-0.76	-0.81	-0.76
Std. error			0.45	0.45	0.45
Asian					
Coefficient			0.98	0.98	0.98
Std. error			0.54	0.57	0.54
Nat. Amer.					
Coefficient			-1.42	-2.20	-1.42
Std. error			3.31	3.47	3.31
Other Race					
Coefficient			-1.02	-1.29	-1.02
Std. error			0.80	0.81	0.80
CogAT # Black					
Coefficient				-0.07	
Std. error				0.05	
CogAT # Latinx					
Coefficient				0.02	
Std. error				0.05	
CogAT # Asian					
Coefficient				-0.00	
Std. error				0.06	
CogAT # Nat. Amer.					
Coefficient				0.00	
Std. error				0.44	
CogAT # Other Race					
Coefficient				-0.11	
Std. error				0.09	
Math # Black					
Coefficient				0.96	
Std. error				1.55	
Math # Latinx					

Coefficient				-2.27	
Std. error				1.61	
Math # Asian					
Coefficient				2.07	
Std. error				1.87	
Math # Nat. Amer.					
Coefficient				-2.14	
Std. error				12.61	
Math # Other Race					
Coefficient				-0.64	
Std. error				3.06	
Reading # Black					
Coefficient				0.10	
Std. error				1.53	
Reading # Latinx					
Coefficient				0.43	
Std. error				1.59	
Reading # Asian					
Coefficient				-3.50	
Std. error				1.84	
Reading # Nat. Amer.					
Coefficient				3.70	
Std. error				11.33	
Reading # Other Race					
Coefficient				3.52	
Std. error				2.96	
Intercept					
Coefficient	64.01**	64.69**	64.67**	64.59**	64.67**
Std. error	0.69	0.45	0.44	0.45	0.44
Level 2					
CogAT					
Coefficient		0.29	0.23	0.22	0.23
Std. error		0.13	0.14	0.14	0.14
Math					
Coefficient		5.90	5.16	5.20	5.16
Std. error		3.51	3.59	3.59	3.59
Reading					
Coefficient		4.27	5.08	5.18	5.08
Std. error		3.56	3.64	3.64	3.64
Black					
Coefficient			-6.77	-6.74	-6.77
Std. error			4.11	4.11	4.11
Latinx					
Coefficient			-5.58	-5.65	-5.58
Std. error			4.29	4.29	4.29
Asian					
Coefficient			0.22	0.19	0.22
Std. error			5.36	5.36	5.36
Nat. Amer.					
Coefficient			-2.20	-2.87	-2.20
Std. error			30.61	30.65	30.61

Other Race					
Coefficient			-3.00	-3.04	-3.00
Std. error			7.32	7.32	7.32
Level 3					
CogAT					
Coefficient		0.49*	0.34	0.34	0.34
Std. error		0.18	0.20	0.20	0.20
Math					
Coefficient		-3.89	0.80	0.89	0.80
Std. error		6.00	6.31	6.32	6.31
Reading					
Coefficient		8.73	2.91	2.79	2.91
Std. error		6.17	6.81	6.81	6.81
Black					
Coefficient			-2.68	-2.66	-2.68
Std. error			3.44	3.44	3.44
Latinx					
Coefficient			-8.09	-8.20	-8.09
Std. error			4.13	4.14	4.13
Asian					
Coefficient			9.78	9.98	9.78
Std. error			6.14	6.15	6.14
Nat. Amer.					
Coefficient			12.74	11.65	12.74
Std. error			92.39	92.42	92.39
Other Race					
Coefficient			-33.07	-33.17	-33.07
Std. error			20.53	20.54	20.53
Random Effects					
Level 1					
Sigma-square					
Coefficient	349.00	141.31	140.57	140.22	140.57
Std. error	5.48	2.22	2.21	2.21	2.21
Level 2					
Tau					
Coefficient	44.72	45.81	45.92	45.95	45.92
Std. error	4.52	3.66	3.68	3.68	3.68
Level 3					
Tau					
Coefficient	37.35	10.64	9.81	9.81	9.81
Std. error	7.04	3.13	3.05	3.05	3.05
Model fit					
N	8685	8685	8685	8685	8685
Loglik. H0	.	.	.	.	.
LogLikelihood	-38126.57	-34360.98	-34297.11	-34274.08	-34297.11
df	4	13	28	43	28
AIC	76261.14	68747.95	68650.21	68634.16	68650.21
BIC	76289.41	68839.85	68848.16	68938.14	68848.16
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.448	0.446	0.447	0.446
level-2 fixed slopes	0.000	0.027	0.027	0.028	0.027

level-3 fixed slopes	0.000	0.068	0.075	0.075	0.075
all fixed slopes	0.000	0.542	0.549	0.550	0.549
level-2 random intercept	0.104	0.106	0.106	0.106	0.106
level-3 random intercept	0.087	0.025	0.023	0.023	0.023
all random intercepts	0.190	0.131	0.128	0.128	0.128
Proportion Reduction in...					
level-1 var. of previous model	.	0.869	0.005	0.008	0.005
level-2 var. of previous model	.	-0.024	-0.002	-0.003	-0.002
level-3 var. of previous model	.	0.715	0.078	0.078	0.078

*$p < 0.01$. **$p < 0.001$.

Table A9

Model Results for District C Using Group Mean Centering and 2-Category Race

District C	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.20**	0.19**	0.19**	0.19**
Std. error		0.01	0.01	0.01	0.01
Math					
Coefficient		3.60**	3.47**	3.46**	3.47**
Std. error		0.45	0.45	0.45	0.45
Reading					
Coefficient		12.60**	12.62**	12.63**	12.62**
Std. error		0.44	0.44	0.44	0.44
Underserved					
Coefficient			-1.74**	-1.75**	-1.74**
Std. error			0.35	0.35	0.35
CogAT # Underserved					
Coefficient				-0.02	
Std. error				0.04	
Math # Underserved					
Coefficient				-0.80	
Std. error				1.25	
Reading # Underserved					
Coefficient				0.60	
Std. error				1.22	
Intercept					
Coefficient	64.01**	64.69**	64.71**	64.67**	64.71**
Std. error	0.69	0.45	0.45	0.45	0.45
Level 2					
CogAT					
Coefficient		0.29	0.25	0.25	0.25
Std. error		0.13	0.13	0.13	0.13
Math					
Coefficient		5.90	5.02	5.02	5.02
Std. error		3.51	3.53	3.53	3.53
Reading					
Coefficient		4.27	5.01	5.02	5.01
Std. error		3.56	3.57	3.58	3.57
Underserved					
Coefficient			-5.93	-5.90	-5.93
Std. error			3.24	3.25	3.24
Level 3					
CogAT					
Coefficient		0.49*	0.38	0.38	0.38
Std. error		0.18	0.20	0.20	0.20
Math					
Coefficient		-3.89	-3.62	-3.66	-3.62

Std. error		6.00	6.00	6.00	6.00
Reading					
Coefficient		8.73	7.78	7.79	7.78
Std. error		6.17	6.22	6.22	6.22
Underserved					
Coefficient			-3.88	-3.89	-3.88
Std. error			3.32	3.32	3.32
Random Effects					
Level 1					
Sigma-square					
Coefficient	349.00	141.31	140.90	140.92	140.90
Std. error	5.48	2.22	2.21	2.22	2.21
Level 2					
Tau					
Coefficient	44.72	45.81	45.55	45.61	45.55
Std. error	4.52	3.66	3.64	3.65	3.64
Level 3					
Tau					
Coefficient	37.35	10.64	10.65	10.64	10.65
Std. error	7.04	3.13	3.13	3.13	3.13
Model fit					
N	8685	8685	8685	8685	8685
Loglik. H0	.	.	.	.	.
LogLikelihood	-38126.57	-34360.98	-34342.23	-34342.63	-34342.23
df	4	13	16	19	16
AIC	76261.14	68747.95	68716.45	68723.26	68716.45
BIC	76289.41	68839.85	68829.56	68857.57	68829.56
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.448	0.448	0.448	0.448
level-2 fixed slopes	0.000	0.027	0.028	0.028	0.028
level-3 fixed slopes	0.000	0.068	0.070	0.070	0.070
all fixed slopes	0.000	0.542	0.545	0.545	0.545
level-2 random intercept	0.104	0.106	0.105	0.105	0.105
level-3 random intercept	0.087	0.025	0.025	0.025	0.025
all random intercepts	0.190	0.131	0.130	0.130	0.130
Proportion Reduction in...					
level-1 var. of previous model	.	0.869	0.003	0.003	0.003
level-2 var. of previous model	.	-0.024	0.006	0.004	0.006
level-3 var. of previous model	.	0.715	-0.001	0.000	-0.001

*$p < 0.01$. **$p < 0.001$.

Table A10

Model Results for District C Using Grand Mean Centering and 6-Category Race

District C	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.20**	0.18**	0.18**	0.18**
Std. error		0.01	0.01	0.03	0.01
Math					
Coefficient		3.62**	3.27**	2.82**	3.27**
Std. error		0.45	0.45	0.80	0.45
Reading					
Coefficient		12.47**	12.83**	13.11**	12.83**
Std. error		0.44	0.45	0.78	0.45
Black					
Coefficient			-2.35**	-2.21**	-2.35**
Std. error			0.43	0.46	0.43
Latinx					
Coefficient			-0.82	-1.23*	-0.82
Std. error			0.45	0.48	0.45
Asian					
Coefficient			0.97	1.07	0.97
Std. error			0.54	0.61	0.54
Nat. Amer.					
Coefficient			-1.43	-2.91	-1.43
Std. error			3.29	3.54	3.29
Other Race					
Coefficient			-1.03	-1.57	-1.03
Std. error			0.79	0.83	0.79
CogAT # Black					
Coefficient				-0.05	
Std. error				0.04	
CogAT # Latinx					
Coefficient				0.06	
Std. error				0.04	
CogAT # Asian					
Coefficient				-0.02	
Std. error				0.05	
CogAT # Nat. Amer.					
Coefficient				-0.45	
Std. error				0.44	
CogAT # Other Race					
Coefficient				-0.10	
Std. error				0.08	
Math # Black					
Coefficient				2.12	
Std. error				1.13	
Math # Latinx					

Coefficient				-0.72	
Std. error				1.17	
Math # Asian					
Coefficient				2.20	
Std. error				1.61	
Math # Nat. Amer.					
Coefficient				10.35	
Std. error				12.81	
Math # Other Race					
Coefficient				-1.70	
Std. error				2.74	
Reading # Black					
Coefficient				0.03	
Std. error				1.11	
Reading # Latinx					
Coefficient				-0.68	
Std. error				1.16	
Reading # Asian					
Coefficient				-2.55	
Std. error				1.57	
Reading # Nat. Amer.					
Coefficient				2.41	
Std. error				11.31	
Reading # Other Race					
Coefficient				4.73	
Std. error				2.67	
Intercept					
Coefficient	64.01**	64.68**	65.59**	65.69**	65.59**
Std. error	0.69	0.45	0.52	0.54	0.52
Level 3					
School Mean CogAT					
Coefficient		0.28	0.15	0.16	0.15
Std. error		0.18	0.20	0.20	0.20
School Mean Math					
Coefficient		-7.17	-2.13	-2.29	-2.13
Std. error		6.00	6.31	6.31	6.31
School Mean Reading					
Coefficient		-3.99	-10.18	-10.51	-10.18
Std. error		6.17	6.81	6.81	6.81
School Prop. Black					
Coefficient			-0.34	-0.53	-0.34
Std. error			3.46	3.46	3.46
School Prop. Latinx					
Coefficient			-7.28	-8.05	-7.28
Std. error			4.15	4.15	4.15
School Prop. Asian					
Coefficient			8.94	9.07	8.94
Std. error			6.15	6.15	6.15
School Prop. Nat. Amer.					
Coefficient			11.78	9.82	11.78
Std. error			92.26	92.27	92.26

School Prop. Other Race					
Coefficient			-31.56	-31.57	-31.56
Std. error			20.50	20.50	20.50
Random Effects					
Level 1					
Sigma-square					
Coefficient	349.00	141.31	140.56	139.77	140.56
Std. error	5.48	2.22	2.21	2.20	2.21
Level 2					
Tau					
Coefficient	44.72	47.84	47.56	47.71	47.56
Std. error	4.52	3.79	3.76	3.77	3.76
Level 3					
Tau					
Coefficient	37.35	10.13	9.39	9.36	9.39
Std. error	7.04	3.12	3.04	3.03	3.04
Model fit					
N	8685	8685	8685	8685	8685
Loglik. H0	.	.	.	.	.
LogLikelihood	-38126.57	-34372.79	-34323.79	-34290.64	-34323.79
df	4	10	20	35	20
AIC	76261.14	68765.57	68687.58	68651.29	68687.58
BIC	76289.41	68836.26	68828.97	68898.72	68828.97
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.434	0.434	0.435	0.434
level-2 fixed slopes	0.000	0.047	0.047	0.047	0.047
level-3 fixed slopes	0.000	0.067	0.074	0.074	0.074
all fixed slopes	0.000	0.548	0.554	0.556	0.554
level-2 random intercept	0.104	0.108	0.107	0.108	0.107
level-3 random intercept	0.087	0.023	0.021	0.021	0.021
all random intercepts	0.190	0.131	0.128	0.129	0.128
Proportion Reduction in...					
level-1 var. of previous model	.	0.863	0.005	0.011	0.005
level-2 var. of previous model	.	-0.070	0.006	0.003	0.006
level-3 var. of previous model	.	0.729	0.073	0.076	0.073

*$p < 0.01$. **$p < 0.001$.

Table A11

Model Results for District C Using Grand Mean Centering and 2-Category Race

District C	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.20**	0.19**	0.18**	0.19**
Std. error		0.01	0.01	0.02	0.01
Math					
Coefficient		3.62**	3.47**	3.17**	3.47**
Std. error		0.45	0.45	0.68	0.45
Reading					
Coefficient		12.47**	12.50**	12.71**	12.50**
Std. error		0.44	0.44	0.66	0.44
Underserved					
Coefficient			-1.80**	-1.84**	-1.80**
Std. error			0.35	0.35	0.35
CogAT # Underserved					
Coefficient				0.03	
Std. error				0.03	
Math # Underserved					
Coefficient				0.51	
Std. error				0.89	
Reading # Underserved					
Coefficient				-0.38	
Std. error				0.87	
Intercept					
Coefficient	64.01**	64.68**	65.77**	65.90**	65.77**
Std. error	0.69	0.45	0.49	0.50	0.49
Level 3					
School Mean CogAT					
Coefficient		0.28	0.19	0.19	0.19
Std. error		0.18	0.20	0.20	0.20
School Mean Math					
Coefficient		-7.17	-6.71	-6.85	-6.71
Std. error		6.00	6.00	6.01	6.00
School Mean Reading					
Coefficient		-3.99	-5.00	-4.93	-5.00
Std. error		6.17	6.23	6.24	6.23
School Prop. Underserved					
Coefficient			-2.11	-2.20	-2.11
Std. error			3.33	3.34	3.33
Random Effects					
Level 1					
Sigma-square					
Coefficient	349.00	141.31	140.90	140.89	140.90
Std. error	5.48	2.22	2.21	2.22	2.21
Level 2					

Tau					
Coefficient	44.72	47.84	47.63	47.66	47.63
Std. error	4.52	3.79	3.77	3.77	3.77
Level 3					
Tau					
Coefficient	37.35	10.13	10.13	10.18	10.13
Std. error	7.04	3.12	3.12	3.13	3.12
Model fit					
N	8685	8685	8685	8685	8685
Loglik. H0	.	.	.	.	.
LogLikelihood	-38126.57	-34372.79	-34356.81	-34357.42	-34356.81
df	4	10	12	15	12
AIC	76261.14	68765.57	68737.62	68744.84	68737.62
BIC	76289.41	68836.26	68822.46	68850.88	68822.46
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.434	0.435	0.435	0.435
level-2 fixed slopes	0.000	0.047	0.047	0.047	0.047
level-3 fixed slopes	0.000	0.067	0.069	0.068	0.069
all fixed slopes	0.000	0.548	0.550	0.550	0.550
level-2 random intercept	0.104	0.108	0.108	0.108	0.108
level-3 random intercept	0.087	0.023	0.023	0.023	0.023
all random intercepts	0.190	0.131	0.131	0.131	0.131
Proportion Reduction in...					
level-1 var. of previous model	.	0.863	0.003	0.003	0.003
level-2 var. of previous model	.	-0.070	0.004	0.004	0.004
level-3 var. of previous model	.	0.729	-0.000	-0.005	-0.000

*$p < 0.01$. **$p < 0.001$.

Table A12

Model Results for District H Using Group Mean Centering and 6-Category Race

District H	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.15**	0.15**	0.15**	0.15**
Std. error		0.01	0.01	0.01	0.01
Math					
Coefficient		5.59**	5.57**	5.58**	5.54**
Std. error		0.23	0.23	0.23	0.23
Reading					
Coefficient		8.50**	8.46**	8.47**	8.45**
Std. error		0.25	0.25	0.25	0.26
Black					
Coefficient			1.28	1.14	2.03*
Std. error			0.72	0.73	0.73
Latinx					
Coefficient			-0.72	-0.89	-0.20
Std. error			0.64	0.64	0.65
Asian					
Coefficient			1.20	0.88	0.95
Std. error			0.87	0.89	0.88
Nat. Amer.					
Coefficient			1.81	0.09	2.23
Std. error			4.42	4.85	4.42
Other Race					
Coefficient			1.34	0.86	1.44
Std. error			1.32	1.38	1.32
CogAT # Black					
Coefficient				-0.04	
Std. error				0.06	
CogAT # Latinx					
Coefficient				-0.06	
Std. error				0.05	
CogAT # Asian					
Coefficient				0.07	
Std. error				0.07	
CogAT # Nat. Amer.					
Coefficient				0.42	
Std. error				0.55	
CogAT # Other Race					
Coefficient				0.06	
Std. error				0.13	
Math # Black					
Coefficient				-0.46	
Std. error				1.18	
Math # Latinx					

Coefficient				-0.50	
Std. error				1.02	
Math # Asian					
Coefficient				-2.19	
Std. error				1.41	
Math # Nat. Amer.					
Coefficient				4.89	
Std. error				7.48	
Math # Other Race					
Coefficient				0.30	
Std. error				2.06	
Reading # Black					
Coefficient				2.62	
Std. error				1.22	
Reading # Latinx					
Coefficient				1.79	
Std. error				1.04	
Reading # Asian					
Coefficient				1.62	
Std. error				1.44	
Reading # Nat. Amer.					
Coefficient				-6.57	
Std. error				10.77	
Reading # Other Race					
Coefficient				1.06	
Std. error				2.36	
FRL Status					
Coefficient					-2.75**
Std. error					0.54
EL Status					
Coefficient					1.06
Std. error					0.52
Intercept					
Coefficient	56.47**	57.52**	57.43**	57.42**	57.26**
Std. error	0.88	0.67	0.66	0.66	0.64
Level 2					
CogAT					
Coefficient		0.14	0.15	0.15	0.14
Std. error		0.07	0.07	0.07	0.07
Math					
Coefficient		2.69	2.84	2.82	2.78
Std. error		1.25	1.26	1.26	1.27
Reading					
Coefficient		7.74**	7.91**	7.92**	7.96**
Std. error		1.43	1.44	1.44	1.44
Black					
Coefficient			7.26	7.25	7.77
Std. error			5.17	5.17	5.32
Latinx					
Coefficient			5.72	5.73	5.88
Std. error			4.70	4.70	4.94

Asian					
Coefficient			13.19	13.04	13.15
Std. error			6.98	6.98	7.01
Nat. Amer.					
Coefficient			119.05	117.06	118.23
Std. error			68.80	68.93	68.87
Other Race					
Coefficient			13.14	12.98	13.72
Std. error			11.81	11.82	11.82
FRL Status					
Coefficient					-1.32
Std. error					4.23
EL Status					
Coefficient					0.24
Std. error					1.77
Level 3					
CogAT					
Coefficient		0.24	0.22	0.22	0.06
Std. error		0.20	0.22	0.22	0.22
Math					
Coefficient		7.70	9.33	9.39	9.35
Std. error		4.05	4.10	4.10	4.00
Reading					
Coefficient		8.17	2.46	2.42	2.11
Std. error		4.53	4.82	4.82	4.68
Black					
Coefficient			-17.43	-17.77	4.48
Std. error			7.68	7.68	13.72
Latinx					
Coefficient			-20.44*	-20.78*	9.00
Std. error			7.56	7.56	13.09
Asian					
Coefficient			-33.51	-34.22	-9.48
Std. error			14.08	14.08	15.27
Nat. Amer.					
Coefficient			-104.10	-110.89	-54.01
Std. error			167.30	167.31	163.20
Other Race					
Coefficient			-13.42	-15.02	-9.62
Std. error			48.02	48.02	46.61
FRL Status					
Coefficient					-17.10
Std. error					10.49
EL Status					
Coefficient					-11.08
Std. error					5.15
Random Effects					
Level 1					
Sigma-square					
Coefficient	386.86	236.08	235.81	235.84	235.26
Std. error	5.22	3.19	3.19	3.19	3.18

Level 2					
Tau					
Coefficient	154.57	154.12	153.32	153.44	153.63
Std. error	10.34	9.57	9.55	9.55	9.57
Level 3					
Tau					
Coefficient	86.15	37.33	34.46	34.42	30.35
Std. error	13.57	8.10	7.86	7.86	7.41
Model fit					
N	11892	11892	11892	11892	11892
Loglik. H0	.	.	.	.	.
LogLikelihood	-53147.53	-50350.22	-50292.57	-50279.73	-50262.38
df	4	13	28	43	34
AIC	106303.1	100726.4	100641.1	100645.5	100592.8
BIC	106332.6	100822.4	100847.9	100963	100843.8
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.217	0.216	0.217	0.216
level-2 fixed slopes	0.000	0.027	0.028	0.028	0.028
level-3 fixed slopes	0.000	0.084	0.095	0.095	0.105
all fixed slopes	0.000	0.328	0.339	0.339	0.349
level-2 random intercept	0.246	0.242	0.239	0.239	0.239
level-3 random intercept	0.137	0.059	0.054	0.054	0.047
all random intercepts	0.384	0.301	0.293	0.293	0.286
Proportion Reduction in...					
level-1 var. of previous model	.	0.602	0.001	0.001	0.003
level-2 var. of previous model	.	0.003	0.005	0.004	0.003
level-3 var. of previous model	.	0.567	0.077	0.078	0.187

*$p < 0.01$. **$p < 0.001$.

Table A13

Model Results for District H Using Group Mean Centering and 2-Category Race

District H	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.15**	0.15**	0.15**	0.15**
Std. error		0.01	0.01	0.01	0.01
Math					
Coefficient		5.59**	5.58**	5.58**	5.54**
Std. error		0.23	0.23	0.23	0.23
Reading					
Coefficient		8.50**	8.49**	8.51**	8.46**
Std. error		0.25	0.25	0.25	0.26
FRL Status					
Coefficient					-2.72**
Std. error					0.55
EL Status					
Coefficient					0.67
Std. error					0.50
Underserved					
Coefficient			-0.55	-0.58	0.09
Std. error			0.55	0.55	0.56
CogAT # Underserved					
Coefficient				-0.09	
Std. error				0.05	
Math # Underserved					
Coefficient				0.07	
Std. error				0.88	
Reading # Underserved					
Coefficient				1.66	
Std. error				0.90	
Intercept					
Coefficient	56.47**	57.52**	57.52**	57.53**	57.26**
Std. error	0.88	0.67	0.66	0.66	0.64
Level 2					
CogAT					
Coefficient		0.14	0.14	0.14	0.15
Std. error		0.07	0.07	0.07	0.07
Math					
Coefficient		2.69	2.68	2.68	2.72
Std. error		1.25	1.25	1.25	1.27
Reading					
Coefficient		7.74**	7.72**	7.73**	7.69**
Std. error		1.43	1.43	1.43	1.43
FRL Status					
Coefficient					-1.52
Std. error					4.22

EL Status					
Coefficient					-0.46
Std. error					1.36
Underserved					
Coefficient			-0.30	-0.23	0.28
Std. error			3.67	3.67	3.91
Level 3					
CogAT					
Coefficient		0.24	0.07	0.07	0.07
Std. error		0.20	0.21	0.21	0.21
Math					
Coefficient		7.70	7.44	7.46	9.31
Std. error		4.05	4.00	4.00	3.94
Reading					
Coefficient		8.17	5.82	5.83	2.12
Std. error		4.53	4.58	4.58	4.55
FRL Status					
Coefficient					-23.37
Std. error					9.31
EL Status					
Coefficient					-7.37
Std. error					3.21
Underserved					
Coefficient			-10.46	-10.51	16.68
Std. error			4.37	4.37	9.68
Random Effects					
Level 1					
Sigma-square					
Coefficient	386.86	236.08	236.09	236.01	235.59
Std. error	5.22	3.19	3.19	3.19	3.18
Level 2					
Tau					
Coefficient	154.57	154.12	154.09	154.20	154.02
Std. error	10.34	9.57	9.57	9.57	9.55
Level 3					
Tau					
Coefficient	86.15	37.33	35.38	35.40	29.94
Std. error	13.57	8.10	7.89	7.90	7.23
Model fit					
N	11892	11892	11892	11892	11892
Loglik. H0	.	.	.	.	.
LogLikelihood	-53147.53	-50350.22	-50341.95	-50339.73	-50311.44
df	4	13	16	19	22
AIC	106303.1	100726.4	100715.9	100717.5	100666.9
BIC	106332.6	100822.4	100834	100857.7	100829.3
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.217	0.217	0.217	0.216
level-2 fixed slopes	0.000	0.027	0.026	0.026	0.026
level-3 fixed slopes	0.000	0.084	0.090	0.090	0.104
all fixed slopes	0.000	0.328	0.333	0.334	0.347
level-2 random intercept	0.246	0.242	0.241	0.241	0.240

level-3 random intercept	0.137	0.059	0.055	0.055	0.047
all random intercepts	0.384	0.301	0.297	0.297	0.286
Proportion Reduction in...					
level-1 var. of previous model	.	0.602	-0.000	0.000	0.002
level-2 var. of previous model	.	0.003	0.000	-0.000	0.001
level-3 var. of previous model	.	0.567	0.052	0.052	0.198

*$p < 0.01$. **$p < 0.001$.

Table A14

Model Results for District H Using Grand Mean Centering and 6-Category Race

District H	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.15**	0.15**	0.15**	0.15**
Std. error		0.01	0.01	0.04	0.01
Math					
Coefficient		5.49**	5.48**	5.52**	5.41**
Std. error		0.22	0.22	0.68	0.22
Reading					
Coefficient		8.46**	8.43**	6.43**	6.25**
Std. error		0.25	0.25	0.68	0.49
Black					
Coefficient			1.35	0.26	0.83
Std. error			0.71	0.77	0.77
Latinx					
Coefficient			-0.70	-1.83*	-1.31
Std. error			0.63	0.69	0.70
Asian					
Coefficient			1.38	0.18	0.23
Std. error			0.86	1.01	0.93
Nat. Amer.					
Coefficient			2.36	1.79	3.07
Std. error			4.41	5.01	4.94
Other Race					
Coefficient			1.47	1.44	1.45
Std. error			1.31	1.65	1.60
CogAT # Black					
Coefficient				0.02	
Std. error				0.05	
CogAT # Latinx					
Coefficient				-0.01	
Std. error				0.04	
CogAT # Asian					
Coefficient				0.08	
Std. error				0.06	
CogAT # Nat. Amer.					
Coefficient				0.59	
Std. error				0.60	
CogAT # Other Race					
Coefficient				0.02	
Std. error				0.11	
Math # Black					
Coefficient				0.24	
Std. error				0.85	
Math # Latinx					

Coefficient				-0.02	
Std. error				0.73	
Math # Asian					
Coefficient				-1.78	
Std. error				1.23	
Math # Nat. Amer.					
Coefficient				5.12	
Std. error				6.94	
Math # Other Race					
Coefficient				-0.97	
Std. error				1.93	
Reading # Black					
Coefficient				1.77	2.34**
Std. error				0.88	0.61
Reading # Latinx					
Coefficient				2.74**	2.83**
Std. error				0.75	0.51
Reading # Asian					
Coefficient				0.96	0.44
Std. error				1.22	0.78
Reading # Nat. Amer.					
Coefficient				-14.69	-1.31
Std. error				12.32	4.81
Reading # Other Race					
Coefficient				0.36	-0.30
Std. error				2.11	1.52
FRL Status					
Coefficient					-2.96**
Std. error					0.54
EL Status					
Coefficient					0.59
Std. error					0.49
Intercept					
Coefficient	56.47**	57.64**	57.61**	58.86**	60.35**
Std. error	0.88	0.67	0.85	0.90	0.96
Level 3					
School Mean CogAT					
Coefficient		0.09	0.08	0.08	-0.08
Std. error		0.20	0.22	0.22	0.22
School Mean Math					
Coefficient		2.36	3.98	3.85	3.93
Std. error		4.05	4.10	4.11	3.99
School Mean Reading					
Coefficient		-0.48	-6.17	-6.25	-6.61
Std. error		4.53	4.81	4.82	4.67
School Prop. Black					
Coefficient			-18.56	-19.19	2.78
Std. error			7.70	7.72	13.69
School Prop. Latinx					
Coefficient			-19.45	-20.15*	9.88
Std. error			7.57	7.58	13.05

School Prop. Asian					
Coefficient			-35.12	-35.70	-10.55
Std. error			14.08	14.10	15.23
School Prop. Nat. Amer.					
Coefficient			-119.91	-123.84	-70.32
Std. error			166.57	166.78	162.07
School Prop. Other Race					
Coefficient			-12.20	-12.46	-7.29
Std. error			47.91	47.97	46.38
School Prop. FRL					
Coefficient					-14.86
Std. error					10.46
School Prop. EL					
Coefficient					-12.01
Std. error					5.15
Random Effects					
Level 1					
Sigma-square					
Coefficient	386.86	236.01	235.73	235.08	234.39
Std. error	5.22	3.19	3.19	3.18	3.17
Level 2					
Tau					
Coefficient	154.57	158.22	157.47	157.84	158.16
Std. error	10.34	9.74	9.69	9.71	9.72
Level 3					
Tau					
Coefficient	86.15	36.23	33.42	33.53	28.94
Std. error	13.57	8.05	7.80	7.83	7.31
Model fit					
N	11892	11892	11892	11892	11892
Loglik. H0	.	.	.	.	.
LogLikelihood	-53147.53	-50358.48	-50318.47	-50293.81	-50267.85
df	4	10	20	35	29
AIC	106303.1	100737	100676.9	100657.6	100593.7
BIC	106332.6	100810.8	100824.6	100916.1	100807.8
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.209	0.209	0.210	0.210
level-2 fixed slopes	0.000	0.045	0.044	0.045	0.045
level-3 fixed slopes	0.000	0.082	0.094	0.093	0.103
all fixed slopes	0.000	0.337	0.347	0.348	0.358
level-2 random intercept	0.246	0.244	0.241	0.241	0.241
level-3 random intercept	0.137	0.056	0.051	0.051	0.044
all random intercepts	0.384	0.300	0.292	0.293	0.285
Proportion Reduction in...					
level-1 var. of previous model	.	0.591	0.001	0.004	0.007
level-2 var. of previous model	.	-0.024	0.005	0.002	0.000
level-3 var. of previous model	.	0.579	0.078	0.074	0.201

*$p < 0.01$. **$p < 0.001$.

Table A15

Model Results for District H Using Grand Mean Centering and 2-Category Race

District H	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.15**	0.15**	0.18**	0.15**
Std. error		0.01	0.01	0.03	0.01
Math					
Coefficient		5.49**	5.48**	4.95**	5.42**
Std. error		0.22	0.22	0.54	0.22
Reading					
Coefficient		8.46**	8.45**	6.71**	6.32**
Std. error		0.25	0.25	0.54	0.41
FRL Status					
Coefficient					-2.91**
Std. error					0.54
EL Status					
Coefficient					0.05
Std. error					0.47
Underserved					
Coefficient			-0.59	-1.41	-0.85
Std. error			0.54	0.57	0.57
CogAT # Underserved					
Coefficient				-0.03	
Std. error				0.03	
Math # Underserved					
Coefficient				0.62	
Std. error				0.59	
Reading # Underserved					
Coefficient				2.28**	2.66**
Std. error				0.60	0.41
Intercept					
Coefficient	56.47**	57.64**	58.14**	59.05**	60.64**
Std. error	0.88	0.67	0.80	0.82	0.89
Level 3					
School Mean CogAT					
Coefficient		0.09	-0.08	-0.08	-0.08
Std. error		0.20	0.21	0.21	0.21
School Mean Math					
Coefficient		2.36	2.12	2.03	4.01
Std. error		4.05	4.00	4.00	3.93
School Mean Reading					
Coefficient		-0.48	-2.77	-2.88	-6.66
Std. error		4.53	4.58	4.59	4.54
School Prop. FRL					
Coefficient					-21.73
Std. error					9.29

School Prop. EL					
Coefficient					-7.57
Std. error					3.22
School Prop. Underserved					
Coefficient			-9.66	-10.25	17.52
Std. error			4.39	4.40	9.66
Random Effects					
Level 1					
Sigma-square					
Coefficient	386.86	236.01	236.03	235.24	234.66
Std. error	5.22	3.19	3.19	3.18	3.17
Level 2					
Tau					
Coefficient	154.57	158.22	157.94	158.33	157.95
Std. error	10.34	9.74	9.72	9.74	9.70
Level 3					
Tau					
Coefficient	86.15	36.23	34.41	34.64	28.75
Std. error	13.57	8.05	7.85	7.88	7.16
Model fit					
N	11892	11892	11892	11892	11892
Loglik. H0	.	.	.	.	.
LogLikelihood	-53147.53	-50358.48	-50352.45	-50335.82	-50305.33
df	4	10	12	15	17
AIC	106303.1	100737	100728.9	100701.6	100644.7
BIC	106332.6	100810.8	100817.5	100812.4	100770.2
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.209	0.209	0.209	0.209
level-2 fixed slopes	0.000	0.045	0.045	0.046	0.046
level-3 fixed slopes	0.000	0.082	0.089	0.088	0.102
all fixed slopes	0.000	0.337	0.342	0.344	0.357
level-2 random intercept	0.246	0.244	0.242	0.243	0.241
level-3 random intercept	0.137	0.056	0.053	0.053	0.044
all random intercepts	0.384	0.300	0.295	0.296	0.285
Proportion Reduction in...					
level-1 var. of previous model	.	0.591	-0.000	0.003	0.006
level-2 var. of previous model	.	-0.024	0.002	-0.001	0.002
level-3 var. of previous model	.	0.579	0.050	0.044	0.206

*$p < 0.01$. **$p < 0.001$.

Table A16

Model Results for District M1 Using Group Mean Centering and 6-Category Race

District M1	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.10**	0.10**	0.09**	0.10**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.28**	0.28**	0.28**	0.27**
Std. error		0.02	0.02	0.02	0.02
Reading					
Coefficient		0.19**	0.19**	0.19**	0.18**
Std. error		0.01	0.01	0.01	0.01
Black					
Coefficient			-0.55	-0.82	0.20
Std. error			0.52	0.58	0.58
Latinx					
Coefficient			-0.34	-0.24	0.39
Std. error			0.56	0.59	0.61
Asian					
Coefficient			-1.83*	-1.71*	-1.37
Std. error			0.59	0.61	0.62
Nat. Amer.					
Coefficient			-0.34	-0.86	-0.29
Std. error			2.50	4.65	2.49
Other Race					
Coefficient			-1.06	-1.15	-1.02
Std. error			0.59	0.60	0.60
CogAT # Black					
Coefficient				-0.02	
Std. error				0.06	
CogAT # Latinx					
Coefficient				-0.03	
Std. error				0.07	
CogAT # Asian					
Coefficient				0.03	
Std. error				0.07	
CogAT # Nat. Amer.					
Coefficient				0.14	
Std. error				0.31	
CogAT # Other Race					
Coefficient				-0.02	
Std. error				0.07	
Math # Black					
Coefficient				0.06	
Std. error				0.08	
Math # Latinx					

Coefficient				0.14	
Std. error				0.09	
Math # Asian					
Coefficient				-0.20	
Std. error				0.09	
Math # Nat. Amer.					
Coefficient				-0.56	
Std. error				0.90	
Math # Other Race					
Coefficient				0.10	
Std. error				0.09	
Reading # Black					
Coefficient				-0.08	
Std. error				0.05	
Reading # Latinx					
Coefficient				-0.08	
Std. error				0.05	
Reading # Asian					
Coefficient				0.10	
Std. error				0.07	
Reading # Nat. Amer.					
Coefficient				0.21	
Std. error				0.60	
Reading # Other Race					
Coefficient				-0.06	
Std. error				0.06	
FRL Status					
Coefficient					-0.72
Std. error					0.46
EL Status					
Coefficient					-1.42
Std. error					0.58
Intercept					
Coefficient	35.33**	35.43**	35.45**	35.45**	35.46**
Std. error	0.58	0.53	0.52	0.53	0.54
Level 2					
CogAT					
Coefficient		0.17	0.05	0.03	0.04
Std. error		0.28	0.28	0.28	0.30
Math					
Coefficient		0.04	-0.08	-0.07	-0.07
Std. error		0.33	0.35	0.35	0.36
Reading					
Coefficient		0.23	0.35	0.36	0.36
Std. error		0.25	0.27	0.27	0.28
Black					
Coefficient			-17.92	-18.10	-19.36
Std. error			7.03	7.04	9.22
Latinx					
Coefficient			-4.94	-4.25	-6.46
Std. error			7.08	7.08	8.90

Asian					
Coefficient			7.44	7.57	6.92
Std. error			8.89	8.90	10.71
Nat. Amer.					
Coefficient			18.32	18.07	17.84
Std. error			43.12	43.29	44.00
Other Race					
Coefficient			3.46	4.05	2.47
Std. error			10.62	10.62	11.32
FRL Status					
Coefficient					1.88
Std. error					7.32
EL Status					
Coefficient					0.18
Std. error					7.27
Level 3					
CogAT					
Coefficient		-0.10	-0.12	-0.11	-0.16
Std. error		0.24	0.26	0.26	0.28
Math					
Coefficient		0.80	0.48	0.47	0.27
Std. error		0.33	0.43	0.43	0.51
Reading					
Coefficient		-0.29	-0.15	-0.17	-0.09
Std. error		0.36	0.39	0.39	0.42
Black					
Coefficient			12.12	11.68	20.95
Std. error			13.09	13.10	17.85
Latinx					
Coefficient			-4.63	-4.46	20.55
Std. error			14.01	14.01	31.28
Asian					
Coefficient			1.53	2.41	9.94
Std. error			11.93	11.94	18.02
Nat. Amer.					
Coefficient			62.58	58.46	-24.24
Std. error			87.55	87.72	145.30
Other Race					
Coefficient			-54.08	-53.85	-54.31
Std. error			26.97	26.98	27.94
FRL Status					
Coefficient					-5.30
Std. error					13.72
EL Status					
Coefficient					-17.78
Std. error					26.12
Random Effects					
Level 1					
Sigma-square					
Coefficient	92.94	46.75	46.61	46.59	46.41
Std. error	3.81	1.52	1.61	1.50	1.49

Level 2					
Tau					
Coefficient	26.59	23.46	22.72	22.71	23.71
Std. error	4.61	3.98	4.48	4.15	4.63
Level 3					
Tau					
Coefficient	0.00	0.00	0.00	0.00	0.16
Std. error	0.00	0.00	0.00	0.02	2.92
Model fit					
N	2036	2036	2036	2036	2036
Loglik. H0	.	.	.	.	.
LogLikelihood	-7592.318	-6925.004	-6873.768	-6889.806	-6854.401
df	4	13	28	43	34
AIC	15192.64	13876.01	13803.54	13865.61	13776.8
BIC	15215.11	13949.05	13960.86	14107.22	13967.84
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.370	0.365	0.367	0.362
level-2 fixed slopes	0.000	0.026	0.042	0.042	0.042
level-3 fixed slopes	0.000	0.017	0.025	0.025	0.027
all fixed slopes	0.000	0.413	0.432	0.434	0.431
level-2 random intercept	0.222	0.196	0.186	0.186	0.192
level-3 random intercept	0.000	0.000	0.000	0.000	0.001
all random intercepts	0.222	0.196	0.186	0.186	0.193
Proportion Reduction in...					
level-1 var. of previous model	.	0.748	0.003	0.003	0.007
level-2 var. of previous model	.	0.118	0.032	0.032	-0.011
level-3 var. of previous model	.	1.000	-1.3e+08	-1.0e+09	-1.2e+12

*$p < 0.01$. **$p < 0.001$.

Table A17

Model Results for District M1 Using Group Mean Centering and 2-Category Race

District M1	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.10**	0.10**	0.10**	0.10**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.28**	0.28**	0.28**	0.27**
Std. error		0.02	0.02	0.02	0.02
Reading					
Coefficient		0.19**	0.19**	0.19**	0.18**
Std. error		0.01	0.01	0.01	0.01
FRL Status					
Coefficient					-0.77
Std. error					0.45
EL Status					
Coefficient					-1.65*
Std. error					0.55
Underserved					
Coefficient			-0.20	-0.21	0.60
Std. error			0.41	0.43	0.46
CogAT # Underserved					
Coefficient				-0.01	
Std. error				0.05	
Math # Underserved					
Coefficient				0.11	
Std. error				0.06	
Reading # Underserved					
Coefficient				-0.09	
Std. error				0.04	
Intercept					
Coefficient	35.33**	35.43**	35.44**	35.45**	35.45**
Std. error	0.58	0.53	0.52	0.53	0.49
Level 2					
CogAT					
Coefficient		0.17	0.16	0.16	0.17
Std. error		0.28	0.27	0.27	0.26
Math					
Coefficient		0.04	-0.07	-0.06	-0.00
Std. error		0.33	0.33	0.33	0.32
Reading					
Coefficient		0.23	0.24	0.24	0.21
Std. error		0.25	0.25	0.25	0.24
FRL Status					
Coefficient					1.53
Std. error					6.62

EL Status					
Coefficient					2.87
Std. error					5.72
Underserved					
Coefficient			-10.28	-10.26	-13.22
Std. error			4.82	4.83	6.48
Level 3					
CogAT					
Coefficient		-0.10	-0.09	-0.08	-0.10
Std. error		0.24	0.26	0.26	0.25
Math					
Coefficient		0.80	0.78	0.78	0.78
Std. error		0.33	0.34	0.34	0.37
Reading					
Coefficient		-0.29	-0.29	-0.31	-0.32
Std. error		0.36	0.36	0.36	0.35
FRL Status					
Coefficient					-5.73
Std. error					11.55
EL Status					
Coefficient					-5.36
Std. error					12.48
Underserved					
Coefficient			-0.61	-0.44	9.66
Std. error			8.19	8.21	14.46
Random Effects					
Level 1					
Sigma-square					
Coefficient	92.94	46.75	46.77	46.71	46.33
Std. error	3.81	1.52	1.54	1.71	1.49
Level 2					
Tau					
Coefficient	26.59	23.46	22.74	22.83	20.10
Std. error	4.61	3.98	4.23	3.99	3.30
Level 3					
Tau					
Coefficient	0.00	0.00	0.00	0.00	0.00
Std. error	0.00	0.00	0.00	0.00	0.00
Model fit					
N	2036	2036	2036	2036	2036
Loglik. H0	.	.	.	.	.
LogLikelihood	-7592.318	-6925.004	-6917.089	-6921.401	-6901.33
df	4	13	16	19	22
AIC	15192.64	13876.01	13866.18	13880.8	13846.66
BIC	15215.11	13949.05	13956.08	13987.56	13970.27
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.370	0.368	0.368	0.378
level-2 fixed slopes	0.000	0.026	0.038	0.038	0.041
level-3 fixed slopes	0.000	0.017	0.016	0.016	0.018
all fixed slopes	0.000	0.413	0.422	0.423	0.437
level-2 random intercept	0.222	0.196	0.189	0.190	0.170

level-3 random intercept	0.000	0.000	0.000	0.000	0.000
all random intercepts	0.222	0.196	0.189	0.190	0.170
Proportion Reduction in...					
level-1 var. of previous model	.	0.748	-0.000	0.001	0.009
level-2 var. of previous model	.	0.118	0.031	0.027	0.143
level-3 var. of previous model	.	1.000	-2.4e+04	-2.2e+03	-1.0e+05

*$p < 0.01$. **$p < 0.001$.

Table A18

Model Results for District M1 Using Grand Mean Centering and 6-Category Race

District M1	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.10**	0.10**	0.10**	0.09**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.28**	0.28**	0.28**	0.27**
Std. error		0.02	0.02	0.03	0.02
Reading					
Coefficient		0.19**	0.19**	0.19**	0.18**
Std. error		0.01	0.01	0.02	0.01
Black					
Coefficient			-0.65	-0.95	0.11
Std. error			0.52	0.62	0.58
Latinx					
Coefficient			-0.37	-0.42	0.36
Std. error			0.56	0.62	0.61
Asian					
Coefficient			-1.79*	-1.63*	-1.34
Std. error			0.59	0.62	0.61
Nat. Amer.					
Coefficient			-0.39	1.71	-0.34
Std. error			2.49	4.30	2.49
Other Race					
Coefficient			-1.06	-1.06	-1.01
Std. error			0.59	0.60	0.59
CogAT # Black					
Coefficient				-0.04	
Std. error				0.05	
CogAT # Latinx					
Coefficient				-0.07	
Std. error				0.07	
CogAT # Asian					
Coefficient				0.01	
Std. error				0.06	
CogAT # Nat. Amer.					
Coefficient				0.30	
Std. error				0.23	
CogAT # Other Race					
Coefficient				-0.00	
Std. error				0.07	
Math # Black					
Coefficient				0.05	
Std. error				0.07	
Math # Latinx					

Coefficient				0.10	
Std. error				0.08	
Math # Asian					
Coefficient				-0.20*	
Std. error				0.08	
Math # Nat. Amer.					
Coefficient				-0.26	
Std. error				0.81	
Math # Other Race					
Coefficient				0.04	
Std. error				0.08	
Reading # Black					
Coefficient				-0.05	
Std. error				0.05	
Reading # Latinx					
Coefficient				-0.04	
Std. error				0.05	
Reading # Asian					
Coefficient				0.12	
Std. error				0.06	
Reading # Nat. Amer.					
Coefficient				0.01	
Std. error				0.46	
Reading # Other Race					
Coefficient				-0.04	
Std. error				0.06	
FRL Status					
Coefficient					-0.72
Std. error					0.46
EL Status					
Coefficient					-1.39
Std. error					0.57
Intercept					
Coefficient	35.33**	35.46**	35.80**	35.80**	35.98**
Std. error	0.58	0.52	0.52	0.55	0.57
Level 3					
School Mean CogAT					
Coefficient		-0.19	-0.22	-0.21	-0.25
Std. error		0.24	0.25	0.27	0.28
School Mean Math					
Coefficient		0.54	0.25	0.24	0.04
Std. error		0.33	0.41	0.44	0.52
School Mean Reading					
Coefficient		-0.48	-0.37	-0.39	-0.32
Std. error		0.35	0.37	0.40	0.43
School Prop. Black					
Coefficient			11.43	11.07	19.21
Std. error			12.52	13.34	18.30
School Prop. Latinx					
Coefficient			-4.54	-4.83	19.81
Std. error			13.39	14.30	32.09

School Prop. Asian					
Coefficient			3.55	4.18	11.82
Std. error			11.42	12.17	18.54
School Prop. Nat. Amer.					
Coefficient			61.45	59.16	-27.31
Std. error			83.72	89.39	149.28
School Prop. Other Race					
Coefficient			-49.25	-49.10	-50.00
Std. error			25.75	27.46	28.66
School Prop. FRL					
Coefficient					-4.01
Std. error					14.05
School Prop. EL					
Coefficient					-16.96
Std. error					26.87
Random Effects					
Level 1					
Sigma-square					
Coefficient	92.94	46.77	46.42	46.57	46.41
Std. error	3.81	1.50	1.49	1.50	1.49
Level 2					
Tau					
Coefficient	26.59	22.73	20.60	23.07	23.32
Std. error	4.61	3.81	3.41	4.15	4.23
Level 3					
Tau					
Coefficient	0.00	0.00	0.00	0.13	0.54
Std. error	0.00	0.00	0.00	2.00	2.93
Model fit					
N	2036	2036	2036	2036	2036
Loglik. H0	.	.	.	.	.
LogLikelihood	-7592.318	-6923.425	-6904.143	-6909.944	-6879.381
df	4	10	20	35	24
AIC	15192.64	13866.85	13848.29	13889.89	13806.76
BIC	15215.11	13923.04	13960.66	14086.54	13941.61
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.367	0.373	0.366	0.363
level-2 fixed slopes	0.000	0.036	0.037	0.037	0.038
level-3 fixed slopes	0.000	0.017	0.026	0.025	0.027
all fixed slopes	0.000	0.421	0.436	0.428	0.428
level-2 random intercept	0.222	0.189	0.173	0.189	0.190
level-3 random intercept	0.000	0.000	0.000	0.001	0.004
all random intercepts	0.222	0.189	0.173	0.190	0.194
Proportion Reduction in...					
level-1 var. of previous model	.	0.755	0.007	0.004	0.008
level-2 var. of previous model	.	0.145	0.094	-0.015	-0.026
level-3 var. of previous model	.	1.000	-13.048	-1.4e+11	-5.6e+11

*$p < 0.01$. **$p < 0.001$.

Table A19

Model Results for District M1 Using Grand Mean Centering and 2-Category Race

District M1	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.10**	0.10**	0.10**	0.10**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.28**	0.28**	0.26**	0.27**
Std. error		0.02	0.02	0.02	0.02
Reading					
Coefficient		0.19**	0.19**	0.20**	0.18**
Std. error		0.01	0.01	0.02	0.01
FRL Status					
Coefficient					-0.77
Std. error					0.45
EL Status					
Coefficient					-1.61*
Std. error					0.55
Underserved					
Coefficient			-0.27	-0.38	0.52
Std. error			0.41	0.45	0.46
CogAT # Underserved					
Coefficient				-0.03	
Std. error				0.04	
Math # Underserved					
Coefficient				0.07	
Std. error				0.05	
Reading # Underserved					
Coefficient				-0.06	
Std. error				0.04	
Intercept					
Coefficient	35.33**	35.46**	35.52**	35.52**	35.77**
Std. error	0.58	0.52	0.53	0.53	0.55
Level 3					
School Mean CogAT					
Coefficient		-0.19	-0.21	-0.20	-0.20
Std. error		0.24	0.26	0.26	0.27
School Mean Math					
Coefficient		0.54	0.52	0.52	0.53
Std. error		0.33	0.34	0.34	0.41
School Mean Reading					
Coefficient		-0.48	-0.48	-0.48	-0.51
Std. error		0.35	0.36	0.36	0.38
School Prop. FRL					
Coefficient					-4.66
Std. error					12.67

School Prop. EL					
Coefficient					-3.92
Std. error					13.73
School Prop. Underserved					
Coefficient			-0.85	-0.81	8.54
Std. error			8.23	8.24	15.90
Random Effects					
Level 1					
Sigma-square					
Coefficient	92.94	46.77	46.79	46.78	46.50
Std. error	3.81	1.50	1.53	1.51	1.49
Level 2					
Tau					
Coefficient	26.59	22.73	22.97	23.03	23.30
Std. error	4.61	3.81	4.21	3.89	4.23
Level 3					
Tau					
Coefficient	0.00	0.00	0.00	0.00	0.29
Std. error	0.00	0.00	0.00	0.00	2.35
Model fit					
N	2036	2036	2036	2036	2036
Loglik. H0	.	.	.	.	.
LogLikelihood	-7592.318	-6923.425	-6920.16	-6925.791	-6905.744
df	4	10	12	15	16
AIC	15192.64	13866.85	13864.32	13881.58	13843.49
BIC	15215.11	13923.04	13931.74	13965.86	13933.39
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.367	0.366	0.367	0.366
level-2 fixed slopes	0.000	0.036	0.036	0.036	0.038
level-3 fixed slopes	0.000	0.017	0.017	0.017	0.019
all fixed slopes	0.000	0.421	0.420	0.420	0.422
level-2 random intercept	0.222	0.189	0.191	0.191	0.192
level-3 random intercept	0.000	0.000	0.000	0.000	0.002
all random intercepts	0.222	0.189	0.191	0.191	0.194
Proportion Reduction in...					
level-1 var. of previous model	.	0.755	-0.000	-0.000	0.006
level-2 var. of previous model	.	0.145	-0.010	-0.013	-0.025
level-3 var. of previous model	.	1.000	-2.2e+06	-9.9e+06	-3.0e+11

*$p < 0.01$. **$p < 0.001$.

Table A20

Model Results for District M2 Using Group Mean Centering and 6-Category Race

District M2	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.11**	0.11**	0.11**	0.10**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.26**	0.25**	0.25**	0.24**
Std. error		0.02	0.02	0.02	0.02
Reading					
Coefficient		0.18**	0.19**	0.19**	0.19**
Std. error		0.01	0.01	0.01	0.01
Black					
Coefficient			-0.85	-0.57	-0.19
Std. error			0.53	0.58	0.58
Latinx					
Coefficient			-0.95	-1.05	-0.25
Std. error			0.59	0.63	0.63
Asian					
Coefficient			-1.51	-2.20*	-0.94
Std. error			0.67	0.72	0.69
Nat. Amer.					
Coefficient			-5.21	12.59	-5.12
Std. error			3.58	14.78	3.58
Other Race					
Coefficient			-0.41	-0.44	-0.29
Std. error			0.59	0.60	0.60
CogAT # Black					
Coefficient				0.12	
Std. error				0.06	
CogAT # Latinx					
Coefficient				0.01	
Std. error				0.07	
CogAT # Asian					
Coefficient				-0.11	
Std. error				0.07	
CogAT # Nat. Amer.					
Coefficient				1.51	
Std. error				1.22	
CogAT # Other Race					
Coefficient				-0.06	
Std. error				0.06	
Math # Black					
Coefficient				-0.13	
Std. error				0.08	
Math # Latinx					

Coefficient				-0.05	
Std. error				0.09	
Math # Asian					
Coefficient				0.03	
Std. error				0.09	
Math # Nat. Amer.					
Coefficient				-2.68	
Std. error				2.64	
Math # Other Race					
Coefficient				-0.03	
Std. error				0.09	
Reading # Black					
Coefficient				-0.05	
Std. error				0.05	
Reading # Latinx					
Coefficient				-0.01	
Std. error				0.06	
Reading # Asian					
Coefficient				0.13	
Std. error				0.06	
Reading # Nat. Amer.					
Coefficient				0.08	
Std. error				1.40	
Reading # Other Race					
Coefficient				0.02	
Std. error				0.05	
FRL Status					
Coefficient					-0.59
Std. error					0.46
EL Status					
Coefficient					-1.59*
Std. error					0.58
Intercept					
Coefficient	35.71**	35.74**	35.72**	35.70**	35.68**
Std. error	0.64	0.66	0.70	0.69	0.69
Level 2					
CogAT					
Coefficient		-0.09	-0.15	-0.15	-0.09
Std. error		0.30	0.32	0.32	0.32
Math					
Coefficient		1.14*	1.12	1.12	1.02
Std. error		0.41	0.45	0.45	0.45
Reading					
Coefficient		-0.23	-0.20	-0.19	-0.34
Std. error		0.23	0.25	0.25	0.25
Black					
Coefficient			-0.41	-0.45	7.66
Std. error			9.41	9.43	10.76
Latinx					
Coefficient			5.48	5.44	12.72
Std. error			11.21	11.24	13.01

Asian					
Coefficient			-4.93	-5.59	-5.11
Std. error			13.10	13.12	13.60
Nat. Amer.					
Coefficient			-69.66	-47.41	-64.66
Std. error			59.64	61.85	58.85
Other Race					
Coefficient			-5.86	-5.72	-5.50
Std. error			9.06	9.06	9.05
FRL Status					
Coefficient					-23.61
Std. error					11.63
EL Status					
Coefficient					5.26
Std. error					9.16
Level 3					
CogAT					
Coefficient		0.03	-0.13	-0.12	-0.56
Std. error		0.50	0.66	0.66	0.74
Math					
Coefficient		0.07	-0.23	-0.24	0.00
Std. error		0.51	0.78	0.78	0.91
Reading					
Coefficient		0.15	0.66	0.66	0.86
Std. error		0.43	0.55	0.55	0.57
Black					
Coefficient			-8.78	-8.70	2.29
Std. error			19.31	19.18	21.52
Latinx					
Coefficient			-7.57	-8.26	8.62
Std. error			25.81	25.61	30.65
Asian					
Coefficient			13.18	12.94	38.14
Std. error			21.42	21.27	29.08
Nat. Amer.					
Coefficient			223.61	239.01	231.85
Std. error			176.57	175.81	203.60
Other Race					
Coefficient			46.90	46.65	33.41
Std. error			37.71	37.44	39.37
FRL Status					
Coefficient					8.32
Std. error					21.98
EL Status					
Coefficient					-48.78
Std. error					35.76
Random Effects					
Level 1					
Sigma-square					
Coefficient	96.27	48.25	48.12	47.82	47.89
Std. error	3.24	1.62	1.62	1.62	1.61

Level 2					
Tau					
Coefficient	31.94	30.19	31.75	31.86	30.74
Std. error	5.46	5.59	6.05	6.07	5.97
Level 3					
Tau					
Coefficient	0.00	1.20	1.84	1.68	2.01
Std. error	0.00	3.14	4.15	4.09	4.62
Model fit					
N	1859	1859	1859	1859	1859
Loglik. H0	.	.	.	.	.
LogLikelihood	-6974.97	-6366.31	-6314.701	-6322.049	-6291.042
df	4	13	28	43	34
AIC	13957.94	12758.62	12685.4	12730.1	12650.08
BIC	13980.05	12830.48	12840.18	12967.79	12838.03
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.351	0.341	0.345	0.338
level-2 fixed slopes	0.000	0.034	0.038	0.038	0.049
level-3 fixed slopes	0.000	0.006	0.017	0.017	0.024
all fixed slopes	0.000	0.391	0.396	0.400	0.412
level-2 random intercept	0.249	0.231	0.235	0.235	0.224
level-3 random intercept	0.000	0.009	0.014	0.012	0.015
all random intercepts	0.249	0.240	0.248	0.247	0.239
Proportion Reduction in...					
level-1 var. of previous model	.	0.686	0.003	0.009	0.008
level-2 var. of previous model	.	0.055	-0.052	-0.055	-0.018
level-3 var. of previous model	.	-5.6e+08	-0.537	-0.402	-0.679

*$p < 0.01$. **$p < 0.001$.

Table A21

Model Results for District M2 Using Group Mean Centering and 2-Category Race

District M2	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.11**	0.11**	0.11**	0.10**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.26**	0.25**	0.25**	0.24**
Std. error		0.02	0.02	0.02	0.02
Reading					
Coefficient		0.18**	0.19**	0.19**	0.19**
Std. error		0.01	0.01	0.01	0.01
FRL Status					
Coefficient					-0.64
Std. error					0.45
EL Status					
Coefficient					-1.73*
Std. error					0.56
Underserved					
Coefficient			-0.79	-0.82	-0.09
Std. error			0.42	0.44	0.47
CogAT # Underserved					
Coefficient				0.10	
Std. error				0.05	
Math # Underserved					
Coefficient				-0.12	
Std. error				0.06	
Reading # Underserved					
Coefficient				-0.04	
Std. error				0.04	
Intercept					
Coefficient	35.71**	35.74**	35.73**	35.70**	35.69**
Std. error	0.64	0.66	0.68	0.67	0.68
Level 2					
CogAT					
Coefficient		-0.09	-0.09	-0.10	-0.04
Std. error		0.30	0.31	0.31	0.30
Math					
Coefficient		1.14*	1.14*	1.15*	1.04
Std. error		0.41	0.42	0.42	0.43
Reading					
Coefficient		-0.23	-0.23	-0.23	-0.35
Std. error		0.23	0.23	0.23	0.24
FRL Status					
Coefficient					-22.88
Std. error					11.44

EL Status					
Coefficient					4.60
Std. error					8.44
Underserved					
Coefficient			0.14	-0.12	7.55
Std. error			7.20	7.20	9.02
Level 3					
CogAT					
Coefficient		0.03	0.01	0.03	-0.10
Std. error		0.50	0.56	0.56	0.58
Math					
Coefficient		0.07	0.13	0.11	0.43
Std. error		0.51	0.73	0.72	0.78
Reading					
Coefficient		0.15	0.14	0.14	0.40
Std. error		0.43	0.46	0.45	0.50
FRL Status					
Coefficient					21.77
Std. error					18.61
EL Status					
Coefficient					-22.23
Std. error					24.98
Underserved					
Coefficient			1.43	1.19	0.38
Std. error			12.04	11.95	18.27
Random Effects					
Level 1					
Sigma-square					
Coefficient	96.27	48.25	48.19	47.95	47.89
Std. error	3.24	1.62	1.62	1.62	1.61
Level 2					
Tau					
Coefficient	31.94	30.19	30.68	30.66	29.78
Std. error	5.46	5.59	5.71	5.71	5.65
Level 3					
Tau					
Coefficient	0.00	1.20	1.66	1.52	1.93
Std. error	0.00	3.14	3.44	3.39	3.72
Model fit					
N	1859	1859	1859	1859	1859
Loglik. H0	.	.	.	.	.
LogLikelihood	-6974.97	-6366.31	-6358.211	-6359.201	-6334.13
df	4	13	16	19	22
AIC	13957.94	12758.62	12748.42	12756.4	12712.26
BIC	13980.05	12830.48	12836.87	12861.43	12833.87
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.351	0.349	0.351	0.347
level-2 fixed slopes	0.000	0.034	0.034	0.034	0.046
level-3 fixed slopes	0.000	0.006	0.006	0.006	0.011
all fixed slopes	0.000	0.391	0.388	0.391	0.404
level-2 random intercept	0.249	0.231	0.233	0.233	0.223

level-3 random intercept	0.000	0.009	0.013	0.012	0.014
all random intercepts	0.249	0.240	0.246	0.245	0.238
Proportion Reduction in...					
level-1 var. of previous model	.	0.686	0.001	0.006	0.008
level-2 var. of previous model	.	0.055	-0.016	-0.016	0.013
level-3 var. of previous model	.	-5.6e+08	-0.385	-0.275	-0.615

*$p < 0.01$. **$p < 0.001$.

Table A22

Model Results for District M2 Using Grand Mean Centering and 6-Category Race

District M2	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.11**	0.11**	0.09**	0.10**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.26**	0.25**	0.30**	0.29**
Std. error		0.02	0.02	0.03	0.03
Reading					
Coefficient		0.18**	0.19**	0.18**	0.19**
Std. error		0.01	0.01	0.02	0.01
Black					
Coefficient			-0.85	-0.95	-0.64
Std. error			0.53	0.60	0.60
Latinx					
Coefficient			-0.96	-1.20	-0.54
Std. error			0.59	0.65	0.67
Asian					
Coefficient			-1.52	-2.30*	-1.27
Std. error			0.67	0.73	0.74
Nat. Amer.					
Coefficient			-5.52	0.95	-5.60
Std. error			3.57	7.54	3.91
Other Race					
Coefficient			-0.43	-0.49	-0.32
Std. error			0.59	0.60	0.60
CogAT # Black					
Coefficient				0.11	
Std. error				0.05	
CogAT # Latinx					
Coefficient				0.05	
Std. error				0.06	
CogAT # Asian					
Coefficient				-0.10	
Std. error				0.07	
CogAT # Nat. Amer.					
Coefficient				2.38	
Std. error				1.45	
CogAT # Other Race					
Coefficient				-0.00	
Std. error				0.06	
Math # Black					
Coefficient				-0.18*	-0.14**
Std. error				0.07	0.04
Math # Latinx					

Coefficient				-0.12	-0.12
Std. error				0.08	0.05
Math # Asian					
Coefficient				0.04	0.04
Std. error				0.09	0.05
Math # Nat. Amer.					
Coefficient				1.86	-0.37
Std. error				3.22	2.90
Math # Other Race					
Coefficient				-0.06	-0.08
Std. error				0.08	0.05
Reading # Black					
Coefficient				-0.02	
Std. error				0.05	
Reading # Latinx					
Coefficient				-0.00	
Std. error				0.05	
Reading # Asian					
Coefficient				0.11	
Std. error				0.06	
Reading # Nat. Amer.					
Coefficient				-2.44	
Std. error				1.39	
Reading # Other Race					
Coefficient				-0.01	
Std. error				0.05	
FRL Status					
Coefficient					-0.69
Std. error					0.46
EL Status					
Coefficient					-1.93*
Std. error					0.59
Intercept					
Coefficient	35.71**	35.74**	36.07**	36.01**	36.14**
Std. error	0.64	0.65	0.71	0.70	0.70
Level 3					
School Mean CogAT					
Coefficient		-0.07	-0.22	-0.23	-0.67
Std. error		0.50	0.66	0.65	0.73
School Mean Math					
Coefficient		-0.19	-0.49	-0.49	-0.24
Std. error		0.50	0.78	0.78	0.89
School Mean Reading					
Coefficient		-0.03	0.46	0.46	0.69
Std. error		0.43	0.55	0.54	0.56
School Prop. Black					
Coefficient			-7.38	-7.42	1.89
Std. error			19.23	19.07	21.23
School Prop. Latinx					
Coefficient			-8.01	-8.93	5.91
Std. error			25.73	25.51	30.20

School Prop. Asian					
Coefficient			14.38	15.16	39.36
Std. error			21.33	21.16	28.67
School Prop. Nat. Amer.					
Coefficient			233.77	232.54	240.29
Std. error			176.09	174.59	200.99
School Prop. Other Race					
Coefficient			44.78	44.97	32.09
Std. error			37.50	37.19	38.72
School Prop. FRL					
Coefficient					10.32
Std. error					21.65
School Prop. EL					
Coefficient					-46.41
Std. error					35.26
Random Effects					
Level 1					
Sigma-square					
Coefficient	96.27	48.25	48.13	47.68	47.47
Std. error	3.24	1.62	1.62	1.61	1.60
Level 2					
Tau					
Coefficient	31.94	31.20	30.93	30.70	30.72
Std. error	5.46	5.64	5.58	5.54	5.55
Level 3					
Tau					
Coefficient	0.00	0.93	1.97	1.87	1.78
Std. error	0.00	3.11	4.10	4.04	4.49
Model fit					
N	1859	1859	1859	1859	1859
Loglik. H0	.	.	.	.	.
LogLikelihood	-6974.97	-6367.45	-6334.729	-6339.447	-6315.413
df	4	10	20	35	29
AIC	13957.94	12754.9	12709.46	12748.89	12688.83
BIC	13980.05	12810.18	12820.01	12942.37	12849.13
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.355	0.350	0.356	0.354
level-2 fixed slopes	0.000	0.018	0.018	0.018	0.018
level-3 fixed slopes	0.000	0.006	0.018	0.018	0.025
all fixed slopes	0.000	0.378	0.386	0.392	0.397
level-2 random intercept	0.249	0.241	0.235	0.233	0.232
level-3 random intercept	0.000	0.007	0.015	0.014	0.013
all random intercepts	0.249	0.248	0.249	0.247	0.245
Proportion Reduction in...					
level-1 var. of previous model	.	0.676	0.003	0.012	0.016
level-2 var. of previous model	.	0.023	0.009	0.016	0.015
level-3 var. of previous model	.	-4.3e+08	-1.127	-1.015	-0.919

*$p < 0.01$. **$p < 0.001$.

Table A23

Model Results for District M2 Using Grand Mean Centering and 2-Category Race

District M2	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.11**	0.11**	0.09**	0.10**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.26**	0.25**	0.30**	0.28**
Std. error		0.02	0.02	0.03	0.03
Reading					
Coefficient		0.18**	0.19**	0.19**	0.18**
Std. error		0.01	0.01	0.02	0.01
FRL Status					
Coefficient					-0.68
Std. error					0.45
EL Status					
Coefficient					-2.24**
Std. error					0.57
Underserved					
Coefficient			-0.79	-1.03	-0.50
Std. error			0.42	0.45	0.48
CogAT # Underserved					
Coefficient				0.09	
Std. error				0.04	
Math # Underserved					
Coefficient				-0.16*	-0.14**
Std. error				0.05	0.03
Reading # Underserved					
Coefficient				-0.02	
Std. error				0.04	
Intercept					
Coefficient	35.71**	35.74**	35.92**	35.87**	36.08**
Std. error	0.64	0.65	0.68	0.68	0.69
Level 3					
School Mean CogAT					
Coefficient		-0.07	-0.09	-0.09	-0.21
Std. error		0.50	0.56	0.56	0.57
School Mean Math					
Coefficient		-0.19	-0.13	-0.13	0.19
Std. error		0.50	0.72	0.72	0.77
School Mean Reading					
Coefficient		-0.03	-0.05	-0.05	0.23
Std. error		0.43	0.46	0.45	0.50
School Prop. FRL					
Coefficient					23.19
Std. error					18.49

School Prop. EL					
Coefficient					-17.97
Std. error					24.90
School Prop. Underserved					
Coefficient			2.27	2.12	-1.25
Std. error			12.00	11.92	18.23
Random Effects					
Level 1					
Sigma-square					
Coefficient	96.27	48.25	48.19	47.80	47.41
Std. error	3.24	1.62	1.62	1.61	1.60
Level 2					
Tau					
Coefficient	31.94	31.20	31.19	31.09	30.99
Std. error	5.46	5.64	5.64	5.62	5.61
Level 3					
Tau					
Coefficient	0.00	0.93	1.48	1.39	1.60
Std. error	0.00	3.11	3.41	3.36	3.68
Model fit					
N	1859	1859	1859	1859	1859
Loglik. H0	.	.	.	.	.
LogLikelihood	-6974.97	-6367.45	-6362.241	-6360.69	-6339.389
df	4	10	12	15	17
AIC	13957.94	12754.9	12748.48	12751.38	12712.78
BIC	13980.05	12810.18	12814.81	12834.3	12806.75
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.355	0.354	0.357	0.358
level-2 fixed slopes	0.000	0.018	0.018	0.019	0.019
level-3 fixed slopes	0.000	0.006	0.006	0.006	0.011
all fixed slopes	0.000	0.378	0.378	0.382	0.388
level-2 random intercept	0.249	0.241	0.240	0.239	0.237
level-3 random intercept	0.000	0.007	0.011	0.011	0.012
all random intercepts	0.249	0.248	0.251	0.250	0.249
Proportion Reduction in...					
level-1 var. of previous model	.	0.676	0.001	0.009	0.017
level-2 var. of previous model	.	0.023	0.000	0.004	0.007
level-3 var. of previous model	.	-4.3e+08	-0.598	-0.497	-0.728

*$p < 0.01$. **$p < 0.001$.

Table A24

Model Results for District M3 Using Group Mean Centering and 6-Category Race

District M3	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.14**	0.13**	0.12**	0.13**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.23**	0.22**	0.23**	0.22**
Std. error		0.02	0.02	0.02	0.02
Reading					
Coefficient		0.20**	0.20**	0.20**	0.20**
Std. error		0.01	0.01	0.01	0.01
Black					
Coefficient			-1.70*	-1.98**	-0.88
Std. error			0.55	0.60	0.60
Latinx					
Coefficient			-1.11	-0.95	-0.24
Std. error			0.57	0.61	0.63
Asian					
Coefficient			-1.39	-1.00	-0.92
Std. error			0.70	0.73	0.71
Nat. Amer.					
Coefficient			-0.61	-4.35	-0.60
Std. error			2.98	3.66	2.97
Other Race					
Coefficient			-1.04	-0.96	-0.93
Std. error			0.60	0.61	0.61
CogAT # Black					
Coefficient				-0.12	
Std. error				0.08	
CogAT # Latinx					
Coefficient				0.06	
Std. error				0.07	
CogAT # Asian					
Coefficient				0.01	
Std. error				0.09	
CogAT # Nat. Amer.					
Coefficient				-1.94	
Std. error				1.12	
CogAT # Other Race					
Coefficient				0.04	
Std. error				0.07	
Math # Black					
Coefficient				-0.04	
Std. error				0.08	
Math # Latinx					

Coefficient				-0.03	
Std. error				0.08	
Math # Asian					
Coefficient				-0.22	
Std. error				0.10	
Math # Nat. Amer.					
Coefficient				0.31	
Std. error				0.43	
Math # Other Race					
Coefficient				-0.06	
Std. error				0.08	
Reading # Black					
Coefficient				0.09	
Std. error				0.05	
Reading # Latinx					
Coefficient				-0.02	
Std. error				0.06	
Reading # Asian					
Coefficient				0.11	
Std. error				0.07	
Reading # Nat. Amer.					
Coefficient				0.62	
Std. error				0.36	
Reading # Other Race					
Coefficient				0.00	
Std. error				0.05	
FRL Status					
Coefficient					-0.97
Std. error					0.47
EL Status					
Coefficient					-1.28
Std. error					0.61
Intercept					
Coefficient	25.34**	25.31**	25.29**	25.22**	25.30**
Std. error	0.57	0.63	0.52	0.52	0.46
Level 2					
CogAT					
Coefficient		-0.05	0.03	0.03	-0.11
Std. error		0.25	0.27	0.27	0.27
Math					
Coefficient		0.43	0.35	0.35	0.31
Std. error		0.30	0.32	0.32	0.33
Reading					
Coefficient		-0.01	0.03	0.03	0.05
Std. error		0.19	0.20	0.20	0.20
Black					
Coefficient			1.94	1.12	9.46
Std. error			6.69	6.70	7.53
Latinx					
Coefficient			-4.63	-5.21	0.41
Std. error			9.32	9.34	9.44

Asian					
Coefficient			1.15	1.18	5.68
Std. error			8.25	8.27	8.51
Nat. Amer.					
Coefficient			-66.46	-73.52	-53.40
Std. error			53.43	53.58	52.93
Other Race					
Coefficient			4.40	3.89	6.40
Std. error			9.30	9.31	9.20
FRL Status					
Coefficient					-1.12
Std. error					6.40
EL Status					
Coefficient					-14.02
Std. error					7.17
Level 3					
CogAT					
Coefficient		-0.15	0.39	0.40	0.37
Std. error		0.36	0.34	0.34	0.32
Math					
Coefficient		0.21	0.08	0.04	-0.19
Std. error		0.49	0.52	0.52	0.50
Reading					
Coefficient		0.21	0.04	0.07	0.29
Std. error		0.40	0.42	0.42	0.40
Black					
Coefficient			4.83	4.00	-11.34
Std. error			10.38	10.31	19.19
Latinx					
Coefficient			22.85	22.25	14.01
Std. error			11.50	11.44	19.66
Asian					
Coefficient			23.34	24.23	23.85
Std. error			11.77	11.70	11.16
Nat. Amer.					
Coefficient			204.88	202.77	335.26*
Std. error			99.69	99.00	115.40
Other Race					
Coefficient			-5.17	-4.43	-15.60
Std. error			19.83	19.69	19.82
FRL Status					
Coefficient					-16.29
Std. error					14.49
EL Status					
Coefficient					45.48
Std. error					22.75
Random Effects					
Level 1					
Sigma-square					
Coefficient	105.23	50.90	50.62	50.57	50.36
Std. error	3.57	1.73	1.72	1.72	2.17

Level 2					
Tau					
Coefficient	14.52	16.55	17.11	17.14	16.44
Std. error	3.40	3.33	3.55	3.56	3.56
Level 3					
Tau					
Coefficient	1.84	3.39	0.90	0.82	0.00
Std. error	2.20	2.92	2.39	2.36	0.00
Model fit					
N	1832	1832	1832	1832	1832
Loglik. H0	.	.	.	.	.
LogLikelihood	-6926.672	-6302.489	-6248.48	-6263.684	-6225.063
df	4	13	28	43	34
AIC	13861.34	12630.98	12552.96	12613.37	12518.13
BIC	13883.4	12702.65	12707.33	12850.43	12705.57
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.415	0.418	0.420	0.419
level-2 fixed slopes	0.000	0.008	0.012	0.012	0.020
level-3 fixed slopes	0.000	0.009	0.019	0.019	0.027
all fixed slopes	0.000	0.431	0.449	0.451	0.466
level-2 random intercept	0.119	0.133	0.137	0.137	0.132
level-3 random intercept	0.015	0.027	0.007	0.007	0.000
all random intercepts	0.135	0.160	0.145	0.144	0.132
Proportion Reduction in...					
level-1 var. of previous model	.	0.843	0.006	0.006	0.011
level-2 var. of previous model	.	-0.140	-0.034	-0.036	0.006
level-3 var. of previous model	.	-0.845	0.734	0.758	1.000

*$p < 0.01$. **$p < 0.001$.

Table A25

Model Results for District M3 Using Group Mean Centering and 2-Category Race

District M3	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.14**	0.14**	0.13**	0.13**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.23**	0.22**	0.22**	0.22**
Std. error		0.02	0.02	0.02	0.02
Reading					
Coefficient		0.20**	0.20**	0.20**	0.20**
Std. error		0.01	0.01	0.01	0.01
FRL Status					
Coefficient					-1.11
Std. error					0.47
EL Status					
Coefficient					-1.24
Std. error					0.60
Underserved					
Coefficient			-1.27*	-1.37*	-0.37
Std. error			0.43	0.45	0.50
CogAT # Underserved					
Coefficient				-0.06	
Std. error				0.05	
Math # Underserved					
Coefficient				-0.01	
Std. error				0.06	
Reading # Underserved					
Coefficient				0.03	
Std. error				0.04	
Intercept					
Coefficient	25.34**	25.31**	25.27**	25.22**	25.25**
Std. error	0.57	0.63	0.60	0.61	0.62
Level 2					
CogAT					
Coefficient		-0.05	-0.06	-0.07	-0.20
Std. error		0.25	0.26	0.26	0.26
Math					
Coefficient		0.43	0.42	0.42	0.42
Std. error		0.30	0.31	0.31	0.31
Reading					
Coefficient		-0.01	-0.01	-0.01	-0.00
Std. error		0.19	0.19	0.19	0.19
FRL Status					
Coefficient					0.26
Std. error					6.18

EL Status					
Coefficient					-12.07
Std. error					7.05
Underserved					
Coefficient			-3.27	-3.55	1.97
Std. error			5.93	5.92	6.59
Level 3					
CogAT					
Coefficient		-0.15	0.08	0.08	0.24
Std. error		0.36	0.37	0.37	0.42
Math					
Coefficient		0.21	0.67	0.67	0.49
Std. error		0.49	0.54	0.54	0.58
Reading					
Coefficient		0.21	-0.19	-0.18	-0.22
Std. error		0.40	0.44	0.44	0.46
FRL Status					
Coefficient					11.14
Std. error					14.73
EL Status					
Coefficient					12.92
Std. error					23.77
Underserved					
Coefficient			14.17	13.95	-6.51
Std. error			7.87	7.89	21.36
Random Effects					
Level 1					
Sigma-square					
Coefficient	105.23	50.90	50.66	50.70	50.36
Std. error	3.57	1.73	1.72	1.72	1.71
Level 2					
Tau					
Coefficient	14.52	16.55	16.62	16.57	16.29
Std. error	3.40	3.33	3.35	3.35	3.35
Level 3					
Tau					
Coefficient	1.84	3.39	2.75	2.80	3.22
Std. error	2.20	2.92	2.69	2.70	3.06
Model fit					
N	1832	1832	1832	1832	1832
Loglik. H0	.	.	.	.	.
LogLikelihood	-6926.672	-6302.489	-6290.625	-6296.606	-6268.554
df	4	13	16	19	22
AIC	13861.34	12630.98	12613.25	12631.21	12581.11
BIC	13883.4	12702.65	12701.46	12735.96	12702.4
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.415	0.417	0.417	0.416
level-2 fixed slopes	0.000	0.008	0.008	0.008	0.014
level-3 fixed slopes	0.000	0.009	0.012	0.012	0.014
all fixed slopes	0.000	0.431	0.437	0.437	0.444
level-2 random intercept	0.119	0.133	0.134	0.133	0.130

level-3 random intercept	0.015	0.027	0.022	0.022	0.026
all random intercepts	0.135	0.160	0.156	0.155	0.155
Proportion Reduction in...					
level-1 var. of previous model	.	0.843	0.005	0.004	0.010
level-2 var. of previous model	.	-0.140	-0.005	-0.001	0.016
level-3 var. of previous model	.	-0.845	0.188	0.176	0.051

*$p < 0.01$. **$p < 0.001$.

Table A26

Model Results for District M3 Using Grand Mean Centering and 6-Category Race

District M3	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.14**	0.13**	0.12**	0.13**
Std. error		0.02	0.02	0.03	0.02
Math					
Coefficient		0.23**	0.22**	0.28**	0.25**
Std. error		0.02	0.02	0.03	0.03
Reading					
Coefficient		0.20**	0.20**	0.18**	0.19**
Std. error		0.01	0.01	0.02	0.01
Black					
Coefficient			-1.67*	-1.88*	-0.51
Std. error			0.55	0.62	0.64
Latinx					
Coefficient			-1.13	-1.22	-0.33
Std. error			0.57	0.63	0.67
Asian					
Coefficient			-1.34	-0.77	-0.16
Std. error			0.70	0.74	0.75
Nat. Amer.					
Coefficient			-0.88	-4.71	-0.54
Std. error			2.97	4.44	3.17
Other Race					
Coefficient			-1.02	-1.02	-0.88
Std. error			0.60	0.61	0.61
CogAT # Black					
Coefficient				-0.07	
Std. error				0.07	
CogAT # Latinx					
Coefficient				0.07	
Std. error				0.07	
CogAT # Asian					
Coefficient				0.08	
Std. error				0.08	
CogAT # Nat. Amer.					
Coefficient				-1.90	
Std. error				1.64	
CogAT # Other Race					
Coefficient				0.09	
Std. error				0.07	
Math # Black					
Coefficient				-0.10	-0.03
Std. error				0.07	0.04
Math # Latinx					

Coefficient				-0.05	-0.09
Std. error				0.08	0.04
Math # Asian					
Coefficient				-0.26*	-0.14*
Std. error				0.10	0.05
Math # Nat. Amer.					
Coefficient				0.41	-0.01
Std. error				0.57	0.17
Math # Other Race					
Coefficient				-0.11	-0.07
Std. error				0.08	0.04
Reading # Black					
Coefficient				0.11	
Std. error				0.05	
Reading # Latinx					
Coefficient				-0.05	
Std. error				0.05	
Reading # Asian					
Coefficient				0.08	
Std. error				0.06	
Reading # Nat. Amer.					
Coefficient				0.60	
Std. error				0.42	
Reading # Other Race					
Coefficient				-0.00	
Std. error				0.05	
FRL Status					
Coefficient					-1.13
Std. error					0.47
EL Status					
Coefficient					-1.64*
Std. error					0.62
Intercept					
Coefficient	25.34**	25.32**	25.83**	25.76**	25.99**
Std. error	0.57	0.63	0.54	0.54	0.49
Level 3					
School Mean CogAT					
Coefficient		-0.29	0.26	0.27	0.24
Std. error		0.37	0.34	0.34	0.32
School Mean Math					
Coefficient		-0.03	-0.16	-0.19	-0.41
Std. error		0.49	0.52	0.52	0.50
School Mean Reading					
Coefficient		0.02	-0.15	-0.11	0.13
Std. error		0.40	0.42	0.42	0.40
School Prop. Black					
Coefficient			6.42	5.43	-10.47
Std. error			10.38	10.33	19.02
School Prop. Latinx					
Coefficient			23.95	24.30	15.93
Std. error			11.52	11.46	19.50

School Prop. Asian					
Coefficient			24.82	26.05	26.12
Std. error			11.78	11.72	11.10
School Prop. Nat. Amer.					
Coefficient			207.63	216.76	350.17*
Std. error			99.66	99.11	114.41
School Prop. Other Race					
Coefficient			-4.73	-3.16	-13.06
Std. error			19.82	19.71	19.66
School Prop. FRL					
Coefficient					-16.08
Std. error					14.37
School Prop. EL					
Coefficient					47.02
Std. error					22.54
Random Effects					
Level 1					
Sigma-square					
Coefficient	105.23	50.90	50.61	50.38	50.18
Std. error	3.57	1.73	1.72	1.72	1.80
Level 2					
Tau					
Coefficient	14.52	16.23	16.14	16.39	16.10
Std. error	3.40	3.21	3.17	3.21	2.98
Level 3					
Tau					
Coefficient	1.84	3.50	1.10	0.99	0.00
Std. error	2.20	2.93	2.36	2.33	0.00
Model fit					
N	1832	1832	1832	1832	1832
Loglik. H0	.	.	.	.	.
LogLikelihood	-6926.672	-6301.488	-6265.711	-6278.672	-6254.018
df	4	10	20	35	29
AIC	13861.34	12622.98	12571.42	12627.34	12566.04
BIC	13883.4	12678.11	12681.69	12820.31	12725.92
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.412	0.419	0.421	0.423
level-2 fixed slopes	0.000	0.015	0.015	0.016	0.015
level-3 fixed slopes	0.000	0.009	0.019	0.019	0.028
all fixed slopes	0.000	0.435	0.453	0.456	0.466
level-2 random intercept	0.119	0.130	0.130	0.132	0.130
level-3 random intercept	0.015	0.028	0.009	0.008	0.000
all random intercepts	0.135	0.158	0.139	0.139	0.130
Proportion Reduction in...					
level-1 var. of previous model	.	0.846	0.006	0.010	0.014
level-2 var. of previous model	.	-0.118	0.006	-0.010	0.008
level-3 var. of previous model	.	-0.902	0.684	0.717	1.000

*$p < 0.01$. **$p < 0.001$.

Table A27

Model Results for District M3 Using Grand Mean Centering and 2-Category Race

District M3	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
CogAT					
Coefficient		0.14**	0.13**	0.14**	0.13**
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		0.23**	0.22**	0.24**	0.22**
Std. error		0.02	0.02	0.03	0.02
Reading					
Coefficient		0.20**	0.20**	0.19**	0.20**
Std. error		0.01	0.01	0.02	0.01
FRL Status					
Coefficient					-1.12
Std. error					0.47
EL Status					
Coefficient					-1.32
Std. error					0.59
Underserved					
Coefficient			-1.29*	-1.45*	-0.36
Std. error			0.43	0.46	0.50
CogAT # Underserved					
Coefficient				-0.03	
Std. error				0.05	
Math # Underserved					
Coefficient				-0.04	
Std. error				0.05	
Reading # Underserved					
Coefficient				0.04	
Std. error				0.04	
Intercept					
Coefficient	25.34**	25.32**	25.58**	25.56**	25.85**
Std. error	0.57	0.63	0.61	0.61	0.64
Level 3					
School Mean CogAT					
Coefficient		-0.29	-0.05	-0.06	0.12
Std. error		0.37	0.37	0.37	0.43
School Mean Math					
Coefficient		-0.03	0.45	0.44	0.28
Std. error		0.49	0.54	0.54	0.58
School Mean Reading					
Coefficient		0.02	-0.38	-0.38	-0.42
Std. error		0.40	0.44	0.44	0.46
School Prop. FRL					
Coefficient					12.21
Std. error					14.82

School Prop. EL					
Coefficient					14.37
Std. error					23.93
School Prop. Underserved					
Coefficient			15.49	15.39	-6.12
Std. error			7.89	7.90	21.49
Random Effects					
Level 1					
Sigma-square					
Coefficient	105.23	50.90	50.66	50.69	50.36
Std. error	3.57	1.73	1.72	1.72	1.71
Level 2					
Tau					
Coefficient	14.52	16.23	16.06	16.06	15.86
Std. error	3.40	3.21	3.16	3.16	3.13
Level 3					
Tau					
Coefficient	1.84	3.50	2.92	2.92	3.40
Std. error	2.20	2.93	2.69	2.69	3.08
Model fit					
N	1832	1832	1832	1832	1832
Loglik. H0	.	.	.	.	.
LogLikelihood	-6926.672	-6301.488	-6292.36	-6298.506	-6277.09
df	4	10	12	15	16
AIC	13861.34	12622.98	12608.72	12627.01	12586.18
BIC	13883.4	12678.11	12674.88	12709.71	12674.39
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.412	0.415	0.415	0.415
level-2 fixed slopes	0.000	0.015	0.014	0.014	0.015
level-3 fixed slopes	0.000	0.009	0.012	0.012	0.014
all fixed slopes	0.000	0.435	0.442	0.442	0.445
level-2 random intercept	0.119	0.130	0.129	0.129	0.126
level-3 random intercept	0.015	0.028	0.023	0.023	0.027
all random intercepts	0.135	0.158	0.152	0.152	0.154
Proportion Reduction in...					
level-1 var. of previous model	.	0.846	0.005	0.004	0.010
level-2 var. of previous model	.	-0.118	0.011	0.011	0.023
level-3 var. of previous model	.	-0.902	0.166	0.164	0.028

*$p < 0.01$. **$p < 0.001$.

Table A28

Model Results for District O2 Using Group Mean Centering and 6-Category Race

District O2	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
InView					
Coefficient		0.01**	0.01**	0.01**	0.01**
Std. error		0.00	0.00	0.00	0.00
Math					
Coefficient		0.23	0.22	0.24	0.39*
Std. error		0.14	0.14	0.14	0.14
Reading					
Coefficient		1.17**	1.17**	1.15**	1.04**
Std. error		0.14	0.14	0.14	0.14
Black					
Coefficient			-0.15	-0.15	-0.07
Std. error			0.12	0.12	0.12
Latinx					
Coefficient			-0.18	-0.15	0.03
Std. error			0.12	0.12	0.12
Asian					
Coefficient			-0.06	-0.01	0.26
Std. error			0.16	0.16	0.17
Nat. Amer.					
Coefficient			-0.27	-0.86	-0.37
Std. error			0.56	0.68	0.56
Other Race					
Coefficient			-0.08	-0.07	-0.05
Std. error			0.16	0.16	0.16
InView # Black					
Coefficient				-0.00	
Std. error				0.00	
InView # Latinx					
Coefficient				-0.00	
Std. error				0.00	
InView # Asian					
Coefficient				0.00	
Std. error				0.01	
InView # Nat. Amer.					
Coefficient				-0.00	
Std. error				0.02	
InView # Other Race					
Coefficient				0.01	
Std. error				0.01	
Math # Black					
Coefficient				-0.14	
Std. error				0.49	
Math # Latinx					

Coefficient				-0.03	
Std. error				0.47	
Math # Asian					
Coefficient				-0.82	
Std. error				0.64	
Math # Nat. Amer.					
Coefficient				0.30	
Std. error				2.35	
Math # Other Race					
Coefficient				-0.76	
Std. error				0.67	
Reading # Black					
Coefficient				-0.16	
Std. error				0.47	
Reading # Latinx					
Coefficient				-0.17	
Std. error				0.46	
Reading # Asian					
Coefficient				0.33	
Std. error				0.63	
Reading # Nat. Amer.					
Coefficient				-1.80	
Std. error				2.93	
Reading # Other Race					
Coefficient				-0.14	
Std. error				0.64	
FRL Status					
Coefficient					-0.34**
Std. error					0.10
EL Status					
Coefficient					-0.49**
Std. error					0.11
Gender					
Coefficient					0.35**
Std. error					0.08
Intercept					
Coefficient	3.37**	3.34**	3.32**	3.29**	3.27**
Std. error	0.14	0.13	0.13	0.13	0.12
Level 2					
InView					
Coefficient		0.02	0.02	0.02	0.01
Std. error		0.01	0.01	0.01	0.01
Math					
Coefficient		-0.88	-0.72	-0.76	-0.49
Std. error		1.41	1.37	1.38	1.40
Reading					
Coefficient		0.44	0.46	0.52	0.17
Std. error		1.36	1.33	1.34	1.37
Black					
Coefficient			-2.25	-2.26	-2.07
Std. error			1.53	1.54	1.55

Latinx				
Coefficient		-2.69	-2.71	-2.32
Std. error		1.29	1.29	1.36
Asian				
Coefficient		-6.11*	-6.07*	-6.07*
Std. error		2.17	2.18	2.19
Nat. Amer.				
Coefficient		-8.69	-9.44	-7.83
Std. error		6.12	6.16	6.19
Other Race				
Coefficient		-1.09	-1.14	-1.34
Std. error		1.74	1.75	1.75
FRL Status				
Coefficient				-0.77
Std. error				0.90
EL Status				
Coefficient				-0.57
Std. error				1.12
Gender				
Coefficient				0.61
Std. error				1.02
Level 3				
InView				
Coefficient	-0.01	0.00	0.00	-0.00
Std. error	0.02	0.02	0.02	0.02
Math				
Coefficient	-1.74	1.96	1.99	2.06
Std. error	2.13	2.40	2.42	2.42
Reading				
Coefficient	3.84	0.50	0.45	0.71
Std. error	2.40	2.58	2.61	2.64
Black				
Coefficient		4.61*	4.72*	2.11
Std. error		1.56	1.57	1.92
Latinx				
Coefficient		3.78*	3.88*	2.78
Std. error		1.30	1.31	1.98
Asian				
Coefficient		3.47	3.53	2.63
Std. error		1.68	1.70	2.43
Nat. Amer.				
Coefficient		-9.53	-9.96	-12.56
Std. error		13.51	13.71	13.44
Other Race				
Coefficient		4.64	4.73	1.43
Std. error		3.78	3.82	4.04
FRL Status				
Coefficient				2.57
Std. error				1.37
EL Status				
Coefficient				-1.64

Std. error					1.84
Gender					
Coefficient					1.28
Std. error					1.55
Random Effects					
Level 1					
Sigma-square					
Coefficient	6.36	3.46	3.46	3.45	3.39
Std. error	0.18	0.10	0.10	0.10	0.10
Level 2					
Tau					
Coefficient	1.09	1.26	1.14	1.15	1.16
Std. error	0.22	0.22	0.21	0.21	0.21
Level 3					
Tau					
Coefficient	0.47	0.37	0.37	0.39	0.32
Std. error	0.22	0.21	0.21	0.21	0.20
Model fit					
N	2618	2618	2618	2618	2618
Loglik. H0	.	.	.	.	.
LogLikelihood	-6259.313	-5517.668	-5492.687	-5504.131	-5459.466
df	4	13	28	43	37
AIC	12526.63	11061.34	11041.37	11094.26	10992.93
BIC	12550.11	11137.65	11205.74	11346.68	11210.13
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.335	0.336	0.337	0.344
level-2 fixed slopes	0.000	0.008	0.016	0.016	0.017
level-3 fixed slopes	0.000	0.026	0.032	0.031	0.037
all fixed slopes	0.000	0.369	0.383	0.384	0.398
level-2 random intercept	0.137	0.156	0.142	0.142	0.143
level-3 random intercept	0.060	0.046	0.046	0.048	0.040
all random intercepts	0.197	0.202	0.188	0.190	0.183
Proportion Reduction in...					
level-1 var. of previous model	.	0.802	-0.000	0.003	0.020
level-2 var. of previous model	.	-0.157	0.091	0.084	0.081
level-3 var. of previous model	.	0.210	0.013	-0.043	0.142

*$p < 0.01$. **$p < 0.001$.

Table A29

Model Results for District O2 Using Group Mean Centering and 2-Category Race

District O2	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
InView					
Coefficient		0.01**	0.01**	0.01**	0.01**
Std. error		0.00	0.00	0.00	0.00
Math					
Coefficient		0.23	0.22	0.22	0.40*
Std. error		0.14	0.14	0.14	0.14
Reading					
Coefficient		1.17**	1.17**	1.17**	1.05**
Std. error		0.14	0.14	0.14	0.14
FRL Status					
Coefficient					-0.33**
Std. error					0.09
EL Status					
Coefficient					-0.40**
Std. error					0.09
Gender					
Coefficient					0.35**
Std. error					0.08
Underserved					
Coefficient			-0.14	-0.14	-0.07
Std. error			0.09	0.09	0.09
InView # Underserved					
Coefficient				-0.00	
Std. error				0.00	
Math # Underserved					
Coefficient				0.26	
Std. error				0.35	
Reading # Underserved					
Coefficient				-0.22	
Std. error				0.34	
Intercept					
Coefficient	3.37**	3.34**	3.34**	3.33**	3.26**
Std. error	0.14	0.13	0.13	0.13	0.12
Level 2					
InView					
Coefficient		0.02	0.02	0.02	0.02
Std. error		0.01	0.01	0.01	0.01
Math					
Coefficient		-0.88	-0.90	-0.87	-0.58
Std. error		1.41	1.41	1.41	1.42
Reading					
Coefficient		0.44	0.49	0.46	0.09
Std. error		1.36	1.36	1.36	1.38

FRL Status					
Coefficient					-0.94
Std. error					0.90
EL Status					
Coefficient					-1.12
Std. error					1.11
Gender					
Coefficient					0.80
Std. error					1.03
Underserved					
Coefficient			-1.24	-1.21	-0.65
Std. error			1.04	1.04	1.09
Level 3					
InView					
Coefficient		-0.01	-0.01	-0.01	-0.00
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		-1.74	-0.75	-0.82	0.99
Std. error		2.13	2.18	2.19	2.17
Reading					
Coefficient		3.84	2.86	2.89	1.95
Std. error		2.40	2.45	2.46	2.39
FRL Status					
Coefficient					3.14*
Std. error					1.03
EL Status					
Coefficient					-0.71
Std. error					1.03
Gender					
Coefficient					1.12
Std. error					1.49
Underserved					
Coefficient			1.38	1.39	1.15
Std. error			0.73	0.73	0.93
Random Effects					
Level 1					
Sigma-square					
Coefficient	6.36	3.46	3.46	3.45	3.39
Std. error	0.18	0.10	0.10	0.10	0.10
Level 2					
Tau					
Coefficient	1.09	1.26	1.24	1.26	1.22
Std. error	0.22	0.22	0.21	0.22	0.21
Level 3					
Tau					
Coefficient	0.47	0.37	0.38	0.38	0.28
Std. error	0.22	0.21	0.20	0.20	0.17
Model fit					
N	2618	2618	2618	2618	2618
Loglik. H0	.	.	.	.	.
LogLikelihood	-6259.313	-5517.668	-5513.931	-5516.206	-5480.239

df	4	13	16	19	25
AIC	12526.63	11061.34	11059.86	11070.41	11010.48
BIC	12550.11	11137.65	11153.78	11181.95	11157.23
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.335	0.336	0.336	0.345
level-2 fixed slopes	0.000	0.008	0.008	0.008	0.009
level-3 fixed slopes	0.000	0.026	0.027	0.027	0.036
all fixed slopes	0.000	0.369	0.371	0.371	0.391
level-2 random intercept	0.137	0.156	0.154	0.155	0.152
level-3 random intercept	0.060	0.046	0.047	0.047	0.035
all random intercepts	0.197	0.202	0.201	0.202	0.187
Proportion Reduction in...					
level-1 var. of previous model	.	0.802	0.001	0.004	0.020
level-2 var. of previous model	.	-0.157	0.011	0.002	0.032
level-3 var. of previous model	.	0.210	-0.002	-0.019	0.244

*$p < 0.01$. **$p < 0.001$.

Table A30

Model Results for District O2 Using Grand Mean Centering and 6-Category Race

District O2	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
InView					
Coefficient		0.01**	0.01**	0.01*	0.01**
Std. error		0.00	0.00	0.00	0.00
Math					
Coefficient		0.22	0.21	0.63	0.38*
Std. error		0.14	0.14	0.28	0.14
Reading					
Coefficient		1.15**	1.15**	1.06**	1.02**
Std. error		0.14	0.14	0.27	0.14
Black					
Coefficient			-0.16	-0.08	-0.08
Std. error			0.12	0.13	0.12
Latinx					
Coefficient			-0.20	-0.15	0.01
Std. error			0.12	0.12	0.12
Asian					
Coefficient			-0.10	0.03	0.22
Std. error			0.16	0.17	0.17
Nat. Amer.					
Coefficient			-0.36	-1.06	-0.45
Std. error			0.56	0.71	0.56
Other Race					
Coefficient			-0.10	-0.05	-0.07
Std. error			0.16	0.17	0.16
InView # Black					
Coefficient				0.00	
Std. error				0.00	
InView # Latinx					
Coefficient				0.00	
Std. error				0.00	
InView # Asian					
Coefficient				0.01	
Std. error				0.00	
InView # Nat. Amer.					
Coefficient				-0.01	
Std. error				0.01	
InView # Other Race					
Coefficient				0.01	
Std. error				0.01	
Math # Black					
Coefficient				-0.23	
Std. error				0.41	
Math # Latinx					

Coefficient				-0.51	
Std. error				0.36	
Math # Asian					
Coefficient				-0.97	
Std. error				0.57	
Math # Nat. Amer.					
Coefficient				-0.70	
Std. error				2.19	
Math # Other Race					
Coefficient				-0.99	
Std. error				0.60	
Reading # Black					
Coefficient				-0.08	
Std. error				0.39	
Reading # Latinx					
Coefficient				0.11	
Std. error				0.35	
Reading # Asian					
Coefficient				0.29	
Std. error				0.57	
Reading # Nat. Amer.					
Coefficient				-0.34	
Std. error				2.57	
Reading # Other Race					
Coefficient				0.28	
Std. error				0.57	
FRL Status					
Coefficient					-0.34**
Std. error					0.10
EL Status					
Coefficient					-0.48**
Std. error					0.11
Gender					
Coefficient					0.34**
Std. error					0.08
Intercept					
Coefficient	3.37**	3.34**	3.43**	3.34**	3.47**
Std. error	0.14	0.13	0.14	0.15	0.16
Level 3					
School Mean InView					
Coefficient		-0.03	-0.01	-0.01	-0.01
Std. error		0.02	0.02	0.02	0.02
School Mean Math					
Coefficient		-2.02	1.73	1.71	1.71
Std. error		2.14	2.40	2.41	2.43
School Mean Reading					
Coefficient		2.95	-0.48	-0.45	-0.25
Std. error		2.41	2.58	2.59	2.66
School Prop. Black					
Coefficient			4.79*	4.93*	2.42
Std. error			1.56	1.56	1.93

School Prop. Latinx					
Coefficient			3.95*	4.03*	2.90
Std. error			1.30	1.30	1.98
School Prop. Asian					
Coefficient			3.48	3.50	2.51
Std. error			1.68	1.68	2.44
School Prop. Nat. Amer.					
Coefficient			-8.26	-8.99	-10.98
Std. error			13.27	13.32	13.28
School Prop. Other Race					
Coefficient			4.53	4.77	1.57
Std. error			3.76	3.77	4.04
School Prop. FRL					
Coefficient					2.64
Std. error					1.37
School Prop. EL					
Coefficient					-1.16
Std. error					1.83
School Prop. Female					
Coefficient					0.97
Std. error					1.56
Random Effects					
Level 1					
Sigma-square					
Coefficient	6.36	3.46	3.46	3.45	3.39
Std. error	0.18	0.10	0.10	0.10	0.10
Level 2					
Tau					
Coefficient	1.09	1.36	1.34	1.33	1.32
Std. error	0.22	0.23	0.23	0.23	0.22
Level 3					
Tau					
Coefficient	0.47	0.34	0.28	0.29	0.26
Std. error	0.22	0.21	0.20	0.20	0.19
Model fit					
N	2618	2618	2618	2618	2618
Loglik. H0	.	.	.	.	.
LogLikelihood	-6259.313	-5520.805	-5509.67	-5522.457	-5480.278
df	4	10	20	35	26
AIC	12526.63	11061.61	11059.34	11114.91	11012.56
BIC	12550.11	11120.31	11176.74	11320.37	11165.18
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.324	0.326	0.329	0.334
level-2 fixed slopes	0.000	0.021	0.021	0.021	0.021
level-3 fixed slopes	0.000	0.027	0.032	0.032	0.037
all fixed slopes	0.000	0.372	0.379	0.382	0.393
level-2 random intercept	0.137	0.165	0.163	0.162	0.161
level-3 random intercept	0.060	0.042	0.034	0.035	0.031
all random intercepts	0.197	0.207	0.198	0.198	0.193
Proportion Reduction in...					
level-1 var. of previous model	.	0.786	-0.001	0.003	0.020

level-2 var. of previous model	.	-0.249	0.015	0.018	0.029
level-3 var. of previous model	.	0.278	0.179	0.160	0.247

*$p < 0.01$. **$p < 0.001$.

Table A31

Model Results for District O2 Using Grand Mean Centering and 2-Category Race

District O2	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
InView					
Coefficient		0.01**	0.01**	0.01**	0.01**
Std. error		0.00	0.00	0.00	0.00
Math					
Coefficient		0.22	0.21	0.22	0.39*
Std. error		0.14	0.14	0.22	0.14
Reading					
Coefficient		1.15**	1.15**	1.17**	1.03**
Std. error		0.14	0.14	0.21	0.14
FRL Status					
Coefficient					-0.33**
Std. error					0.09
EL Status					
Coefficient					-0.40**
Std. error					0.09
Gender					
Coefficient					0.35**
Std. error					0.08
Underserved					
Coefficient			-0.14	-0.13	-0.08
Std. error			0.09	0.09	0.09
InView # Underserved					
Coefficient				-0.00	
Std. error				0.00	
Math # Underserved					
Coefficient				-0.01	
Std. error				0.28	
Reading # Underserved					
Coefficient				-0.07	
Std. error				0.28	
Intercept					
Coefficient	3.37**	3.34**	3.42**	3.38**	3.48**
Std. error	0.14	0.13	0.14	0.14	0.15
Level 3					
School Mean InView					
Coefficient		-0.03	-0.02	-0.02	-0.01
Std. error		0.02	0.02	0.02	0.02
School Mean Math					
Coefficient		-2.02	-1.05	-1.20	0.56
Std. error		2.14	2.20	2.21	2.18
School Mean Reading					
Coefficient		2.95	1.98	2.10	1.13
Std. error		2.41	2.46	2.47	2.40

School Prop. FRL					
Coefficient					3.43**
Std. error					1.04
School Prop. EL					
Coefficient					-0.30
Std. error					1.03
School Prop. Female					
Coefficient					0.82
Std. error					1.50
School Prop. Underserved					
Coefficient			1.50	1.51	1.20
Std. error			0.74	0.74	0.94
Random Effects					
Level 1					
Sigma-square					
Coefficient	6.36	3.46	3.46	3.45	3.39
Std. error	0.18	0.10	0.10	0.10	0.10
Level 2					
Tau					
Coefficient	1.09	1.36	1.33	1.33	1.29
Std. error	0.22	0.23	0.23	0.22	0.21
Level 3					
Tau					
Coefficient	0.47	0.34	0.35	0.36	0.26
Std. error	0.22	0.21	0.20	0.20	0.17
Model fit					
N	2618	2618	2618	2618	2618
Loglik. H0	.	.	.	.	.
LogLikelihood	-6259.313	-5520.805	-5518.693	-5522.031	-5488.778
df	4	10	12	15	18
AIC	12526.63	11061.61	11061.39	11074.06	11013.56
BIC	12550.11	11120.31	11131.83	11162.12	11119.22
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.324	0.325	0.326	0.335
level-2 fixed slopes	0.000	0.021	0.021	0.021	0.022
level-3 fixed slopes	0.000	0.027	0.028	0.028	0.037
all fixed slopes	0.000	0.372	0.374	0.375	0.393
level-2 random intercept	0.137	0.165	0.163	0.162	0.158
level-3 random intercept	0.060	0.042	0.042	0.043	0.032
all random intercepts	0.197	0.207	0.205	0.205	0.190
Proportion Reduction in...					
level-1 var. of previous model	.	0.786	0.001	0.003	0.020
level-2 var. of previous model	.	-0.249	0.018	0.021	0.050
level-3 var. of previous model	.	0.278	-0.014	-0.040	0.232

*$p < 0.01$. **$p < 0.001$.

Table A32

Model Results for District O3 Using Group Mean Centering and 6-Category Race

District O3	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
InView					
Coefficient		0.00	0.00*	0.00*	0.00
Std. error		0.00	0.00	0.00	0.00
Math					
Coefficient		0.47*	0.48*	0.50*	0.62**
Std. error		0.16	0.16	0.16	0.16
Reading					
Coefficient		1.07**	1.06**	1.05**	0.92**
Std. error		0.16	0.16	0.16	0.16
Black					
Coefficient			0.25	0.23	0.34
Std. error			0.15	0.15	0.15
Latinx					
Coefficient			0.23	0.23	0.42*
Std. error			0.15	0.15	0.15
Asian					
Coefficient			0.25	0.25	0.48
Std. error			0.20	0.20	0.20
Nat. Amer.					
Coefficient			0.80	1.27	0.92
Std. error			0.62	0.65	0.61
Other Race					
Coefficient			0.23	0.24	0.29
Std. error			0.20	0.20	0.20
InView # Black					
Coefficient				-0.00	
Std. error				0.01	
InView # Latinx					
Coefficient				-0.00	
Std. error				0.01	
InView # Asian					
Coefficient				0.00	
Std. error				0.01	
InView # Nat. Amer.					
Coefficient				0.00	
Std. error				0.03	
InView # Other Race					
Coefficient				-0.00	
Std. error				0.01	
Math # Black					
Coefficient				-0.08	
Std. error				0.56	
Math # Latinx					

Coefficient				1.04	
Std. error				0.54	
Math # Asian					
Coefficient				-0.04	
Std. error				0.77	
Math # Nat. Amer.					
Coefficient				3.44	
Std. error				2.34	
Math # Other Race					
Coefficient				1.24	
Std. error				0.75	
Reading # Black					
Coefficient				0.30	
Std. error				0.55	
Reading # Latinx					
Coefficient				-0.85	
Std. error				0.52	
Reading # Asian					
Coefficient				0.01	
Std. error				0.80	
Reading # Nat. Amer.					
Coefficient				-2.38	
Std. error				2.51	
Reading # Other Race					
Coefficient				-1.20	
Std. error				0.76	
FRL Status					
Coefficient					-0.40**
Std. error					0.12
EL Status					
Coefficient					-0.32
Std. error					0.13
Gender					
Coefficient					0.30*
Std. error					0.09
Intercept					
Coefficient	2.92**	2.93**	2.92**	2.92**	2.91**
Std. error	0.13	0.14	0.13	0.13	0.13
Level 2					
InView					
Coefficient		0.01	0.01	0.01	0.02
Std. error		0.01	0.01	0.01	0.01
Math					
Coefficient		0.88	1.84	1.86	1.60
Std. error		1.61	1.58	1.58	1.58
Reading					
Coefficient		-0.04	-0.80	-0.83	-1.04
Std. error		1.59	1.53	1.53	1.53
Black					
Coefficient			2.84	2.78	2.58
Std. error			1.75	1.75	1.83

Latinx				
Coefficient		0.62	0.61	1.06
Std. error		1.38	1.38	1.45
Asian				
Coefficient		2.28	2.19	2.68
Std. error		1.98	1.99	2.03
Nat. Amer.				
Coefficient		-9.85	-9.60	-10.02
Std. error		6.82	6.84	6.87
Other Race				
Coefficient		-2.17	-2.18	-2.87
Std. error		2.02	2.02	2.10
FRL Status				
Coefficient				-0.21
Std. error				1.08
EL Status				
Coefficient				-1.98
Std. error				1.26
Gender				
Coefficient				-0.05
Std. error				1.31
Level 3				
InView				
Coefficient	0.00	0.03	0.03	0.03
Std. error	0.02	0.02	0.02	0.02
Math				
Coefficient	1.01	0.56	0.55	1.05
Std. error	1.97	1.88	1.89	2.00
Reading				
Coefficient	-0.44	0.20	0.18	-0.88
Std. error	2.15	2.06	2.07	2.29
Black				
Coefficient		4.63**	4.62**	4.52*
Std. error		1.28	1.29	1.55
Latinx				
Coefficient		4.06**	4.08**	5.11**
Std. error		1.06	1.07	1.55
Asian				
Coefficient		4.67*	4.74*	5.91*
Std. error		1.53	1.55	2.06
Nat. Amer.				
Coefficient		15.19	16.54	12.16
Std. error		11.54	11.67	12.40
Other Race				
Coefficient		3.83	3.75	3.94
Std. error		3.40	3.43	3.68
FRL Status				
Coefficient				-0.55
Std. error				1.30
EL Status				
Coefficient				-1.65

Std. error					1.63
Gender					
Coefficient					0.97
Std. error					1.74
Random Effects					
Level 1					
Sigma-square					
Coefficient	6.20	4.11	4.11	4.10	4.06
Std. error	0.20	0.13	0.13	0.13	0.13
Level 2					
Tau					
Coefficient	1.25	1.33	1.16	1.17	1.14
Std. error	0.25	0.25	0.23	0.23	0.22
Level 3					
Tau					
Coefficient	0.33	0.43	0.33	0.34	0.40
Std. error	0.25	0.28	0.23	0.23	0.24
Model fit					
N	2176	2176	2176	2176	2176
Loglik. H0	.	.	.	.	.
LogLikelihood	-5186.216	-4776.653	-4745.847	-4757.276	-4726.167
df	4	13	28	43	37
AIC	10380.43	9579.306	9547.695	9600.551	9526.335
BIC	10403.17	9653.214	9706.882	9845.017	9736.689
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.243	0.242	0.244	0.245
level-2 fixed slopes	0.000	0.010	0.022	0.022	0.025
level-3 fixed slopes	0.000	0.012	0.042	0.042	0.044
all fixed slopes	0.000	0.264	0.306	0.308	0.314
level-2 random intercept	0.160	0.167	0.144	0.144	0.140
level-3 random intercept	0.043	0.054	0.041	0.042	0.049
all random intercepts	0.203	0.221	0.185	0.186	0.189
Proportion Reduction in...					
level-1 var. of previous model	.	0.785	0.001	0.002	0.013
level-2 var. of previous model	.	-0.070	0.127	0.124	0.141
level-3 var. of previous model	.	-0.307	0.239	0.205	0.070

*$p < 0.01$. **$p < 0.001$.

Table A33

Model Results for District O3 Using Group Mean Centering and 2-Category Race

District O3	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
InView					
Coefficient		0.00	0.00*	0.00*	0.00
Std. error		0.00	0.00	0.00	0.00
Math					
Coefficient		0.47*	0.48*	0.49*	0.62**
Std. error		0.16	0.16	0.16	0.16
Reading					
Coefficient		1.07**	1.07**	1.06**	0.93**
Std. error		0.16	0.16	0.16	0.16
FRL Status					
Coefficient					-0.37*
Std. error					0.12
EL Status					
Coefficient					-0.23
Std. error					0.12
Gender					
Coefficient					0.29*
Std. error					0.09
Underserved					
Coefficient			0.14	0.15	0.21
Std. error			0.11	0.11	0.11
InView # Underserved					
Coefficient				-0.00	
Std. error				0.00	
Math # Underserved					
Coefficient				0.31	
Std. error				0.40	
Reading # Underserved					
Coefficient				-0.10	
Std. error				0.40	
Intercept					
Coefficient	2.92**	2.93**	2.92**	2.92**	2.90**
Std. error	0.13	0.14	0.14	0.14	0.14
Level 2					
InView					
Coefficient		0.01	0.02	0.02	0.02
Std. error		0.01	0.01	0.01	0.01
Math					
Coefficient		0.88	0.92	0.92	0.83
Std. error		1.61	1.59	1.59	1.60
Reading					
Coefficient		-0.04	-0.16	-0.15	-0.47
Std. error		1.59	1.56	1.56	1.58

FRL Status					
Coefficient					-0.13
Std. error					1.10
EL Status					
Coefficient					-1.68
Std. error					1.21
Gender					
Coefficient					0.03
Std. error					1.34
Underserved					
Coefficient			1.12	1.13	1.34
Std. error			1.21	1.21	1.26
Level 3					
InView					
Coefficient		0.00	0.01	0.01	0.02
Std. error		0.02	0.02	0.02	0.02
Math					
Coefficient		1.01	0.32	0.31	0.42
Std. error		1.97	1.98	1.98	2.02
Reading					
Coefficient		-0.44	0.48	0.48	0.20
Std. error		2.15	2.17	2.17	2.22
FRL Status					
Coefficient					1.57
Std. error					1.14
EL Status					
Coefficient					-0.88
Std. error					0.95
Gender					
Coefficient					1.67
Std. error					1.76
Underserved					
Coefficient			1.87*	1.87*	1.92
Std. error			0.71	0.71	0.85
Random Effects					
Level 1					
Sigma-square					
Coefficient	6.20	4.11	4.10	4.11	4.06
Std. error	0.20	0.13	0.13	0.13	0.13
Level 2					
Tau					
Coefficient	1.25	1.33	1.28	1.28	1.30
Std. error	0.25	0.25	0.24	0.24	0.24
Level 3					
Tau					
Coefficient	0.33	0.43	0.44	0.45	0.42
Std. error	0.25	0.28	0.25	0.25	0.26
Model fit					
N	2176	2176	2176	2176	2176
Loglik. H0	.	.	.	.	.
LogLikelihood	-5186.216	-4776.653	-4771.539	-4776.627	-4753.936

df	4	13	16	19	25
AIC	10380.43	9579.306	9575.079	9591.254	9557.872
BIC	10403.17	9653.214	9666.043	9699.273	9700.004
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.243	0.242	0.242	0.246
level-2 fixed slopes	0.000	0.010	0.010	0.010	0.013
level-3 fixed slopes	0.000	0.012	0.023	0.023	0.027
all fixed slopes	0.000	0.264	0.275	0.275	0.286
level-2 random intercept	0.160	0.167	0.160	0.159	0.160
level-3 random intercept	0.043	0.054	0.055	0.055	0.052
all random intercepts	0.203	0.221	0.215	0.215	0.212
Proportion Reduction in...					
level-1 var. of previous model	.	0.785	0.002	0.001	0.012
level-2 var. of previous model	.	-0.070	0.036	0.037	0.028
level-3 var. of previous model	.	-0.307	-0.024	-0.031	0.028

*$p < 0.01$. **$p < 0.001$.

Table A34

Model Results for District O3 Using Grand Mean Centering and 6-Category Race

District O3	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
InView					
Coefficient		0.00*	0.00*	0.01	0.00
Std. error		0.00	0.00	0.00	0.00
Math					
Coefficient		0.47*	0.49*	-0.24	0.62**
Std. error		0.16	0.16	0.32	0.16
Reading					
Coefficient		1.06**	1.05**	1.48**	0.90**
Std. error		0.16	0.16	0.32	0.16
Black					
Coefficient			0.27	0.18	0.36
Std. error			0.15	0.15	0.15
Latinx					
Coefficient			0.22	0.16	0.42*
Std. error			0.14	0.15	0.15
Asian					
Coefficient			0.27	0.19	0.50
Std. error			0.19	0.20	0.20
Nat. Amer.					
Coefficient			0.71	0.98	0.83
Std. error			0.61	0.70	0.61
Other Race					
Coefficient			0.20	0.14	0.26
Std. error			0.20	0.20	0.20
InView # Black					
Coefficient				-0.01	
Std. error				0.00	
InView # Latinx					
Coefficient				-0.01	
Std. error				0.00	
InView # Asian					
Coefficient				0.01	
Std. error				0.01	
InView # Nat. Amer.					
Coefficient				-0.00	
Std. error				0.02	
InView # Other Race					
Coefficient				-0.00	
Std. error				0.01	
Math # Black					
Coefficient				0.43	
Std. error				0.45	
Math # Latinx					

Coefficient				1.40**	
Std. error				0.41	
Math # Asian					
Coefficient				0.50	
Std. error				0.63	
Math # Nat. Amer.					
Coefficient				3.17	
Std. error				2.17	
Math # Other Race					
Coefficient				1.13	
Std. error				0.70	
Reading # Black					
Coefficient				-0.04	
Std. error				0.46	
Reading # Latinx					
Coefficient				-0.91	
Std. error				0.41	
Reading # Asian					
Coefficient				-0.44	
Std. error				0.66	
Reading # Nat. Amer.					
Coefficient				-1.81	
Std. error				2.33	
Reading # Other Race					
Coefficient				-0.88	
Std. error				0.72	
FRL Status					
Coefficient					-0.40**
Std. error					0.12
EL Status					
Coefficient					-0.33
Std. error					0.13
Gender					
Coefficient					0.29*
Std. error					0.09
Intercept					
Coefficient	2.92**	2.94**	2.72**	2.79**	2.84**
Std. error	0.13	0.14	0.16	0.17	0.18
Level 3					
School Mean InView					
Coefficient		-0.00	0.03	0.03	0.03
Std. error		0.02	0.02	0.02	0.02
School Mean Math					
Coefficient		0.51	0.07	-0.03	0.39
Std. error		1.98	1.89	1.89	2.02
School Mean Reading					
Coefficient		-1.48	-0.89	-0.87	-1.78
Std. error		2.15	2.07	2.07	2.31
School Prop. Black					
Coefficient			4.33**	4.29**	4.12*
Std. error			1.28	1.29	1.56

School Prop. Latinx					
Coefficient			3.79**	3.72**	4.64*
Std. error			1.06	1.07	1.56
School Prop. Asian					
Coefficient			4.44*	4.44*	5.40*
Std. error			1.54	1.55	2.08
School Prop. Nat. Amer.					
Coefficient			12.64	13.31	9.46
Std. error			11.52	11.55	12.42
School Prop. Other Race					
Coefficient			3.40	3.22	3.51
Std. error			3.39	3.40	3.70
School Prop. FRL					
Coefficient					-0.10
Std. error					1.32
School Prop. EL					
Coefficient					-1.28
Std. error					1.64
School Prop. Female					
Coefficient					0.66
Std. error					1.75
Random Effects					
Level 1					
Sigma-square					
Coefficient	6.20	4.11	4.10	4.08	4.05
Std. error	0.20	0.13	0.13	0.13	0.13
Level 2					
Tau					
Coefficient	1.25	1.31	1.27	1.28	1.25
Std. error	0.25	0.24	0.23	0.23	0.22
Level 3					
Tau					
Coefficient	0.33	0.44	0.29	0.29	0.37
Std. error	0.25	0.27	0.23	0.23	0.24
Model fit					
N	2176	2176	2176	2176	2176
Loglik. H0	.	.	.	.	.
LogLikelihood	-5186.216	-4775.517	-4758.866	-4767.305	-4743.716
df	4	10	20	35	26
AIC	10380.43	9571.034	9557.732	9604.61	9539.432
BIC	10403.17	9627.887	9671.437	9803.593	9687.248
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.242	0.242	0.247	0.246
level-2 fixed slopes	0.000	0.012	0.012	0.011	0.012
level-3 fixed slopes	0.000	0.012	0.042	0.042	0.044
all fixed slopes	0.000	0.265	0.295	0.300	0.301
level-2 random intercept	0.160	0.165	0.158	0.158	0.154
level-3 random intercept	0.043	0.055	0.036	0.036	0.045
all random intercepts	0.203	0.220	0.194	0.194	0.199
Proportion Reduction in...					
level-1 var. of previous model	.	0.788	0.001	0.007	0.014

level-2 var. of previous model	.	-0.055	0.032	0.028	0.051
level-3 var. of previous model	.	-0.332	0.354	0.350	0.165

*$p < 0.01$. **$p < 0.001$.

Table A35

Model Results for District O3 Using Grand Mean Centering and 2-Category Race

District O3	Model 1	Model 2	Model 3a	Model 3b	Model 3c
Fixed Effects					
Level 1					
InView					
Coefficient		0.00*	0.00*	0.01*	0.00
Std. error		0.00	0.00	0.00	0.00
Math					
Coefficient		0.47*	0.48*	0.01	0.62**
Std. error		0.16	0.16	0.25	0.16
Reading					
Coefficient		1.06**	1.05**	1.29**	0.92**
Std. error		0.16	0.16	0.26	0.16
FRL Status					
Coefficient					-0.36*
Std. error					0.12
EL Status					
Coefficient					-0.24
Std. error					0.12
Gender					
Coefficient					0.29*
Std. error					0.09
Underserved					
Coefficient			0.15	0.15	0.21
Std. error			0.11	0.11	0.11
InView # Underserved					
Coefficient				-0.01	
Std. error				0.00	
Math # Underserved					
Coefficient				0.78	
Std. error				0.32	
Reading # Underserved					
Coefficient				-0.39	
Std. error				0.32	
Intercept					
Coefficient	2.92**	2.94**	2.83**	2.84**	2.97**
Std. error	0.13	0.14	0.15	0.16	0.18
Level 3					
School Mean InView					
Coefficient		-0.00	0.01	0.01	0.02
Std. error		0.02	0.02	0.02	0.02
School Mean Math					
Coefficient		0.51	-0.22	-0.23	-0.26
Std. error		1.98	1.99	1.99	2.03
School Mean Reading					
Coefficient		-1.48	-0.55	-0.62	-0.69
Std. error		2.15	2.18	2.17	2.23

School Prop. FRL					
Coefficient					1.94
Std. error					1.15
School Prop. EL					
Coefficient					-0.65
Std. error					0.96
School Prop. Female					
Coefficient					1.40
Std. error					1.76
School Prop. Underserved					
Coefficient			1.70	1.67	1.70
Std. error			0.72	0.72	0.86
Random Effects					
Level 1					
Sigma-square					
Coefficient	6.20	4.11	4.10	4.09	4.06
Std. error	0.20	0.13	0.13	0.13	0.13
Level 2					
Tau					
Coefficient	1.25	1.31	1.26	1.26	1.26
Std. error	0.25	0.24	0.23	0.23	0.23
Level 3					
Tau					
Coefficient	0.33	0.44	0.46	0.46	0.44
Std. error	0.25	0.27	0.25	0.25	0.26
Model fit					
N	2176	2176	2176	2176	2176
Loglik. H0	.	.	.	.	.
LogLikelihood	-5186.216	-4775.517	-4771.925	-4773.985	-4758.482
df	4	10	12	15	18
AIC	10380.43	9571.034	9567.85	9577.971	9552.964
BIC	10403.17	9627.887	9636.073	9663.249	9655.298
R^2: Prop. of Variance Explained by...					
level-1 fixed slopes	0.000	0.242	0.241	0.243	0.246
level-2 fixed slopes	0.000	0.012	0.012	0.011	0.012
level-3 fixed slopes	0.000	0.012	0.022	0.022	0.027
all fixed slopes	0.000	0.265	0.275	0.277	0.285
level-2 random intercept	0.160	0.165	0.157	0.157	0.157
level-3 random intercept	0.043	0.055	0.057	0.057	0.055
all random intercepts	0.203	0.220	0.214	0.214	0.212
Proportion Reduction in...					
level-1 var. of previous model	.	0.788	0.002	0.004	0.012
level-2 var. of previous model	.	-0.055	0.041	0.040	0.038
level-3 var. of previous model	.	-0.332	-0.041	-0.033	-0.002

*$p < 0.01$. **$p < 0.001$.

Table A36

Proportion of Teacher Rating Scale Variance that was Unexplained Between Teacher Variance across Models Using Group Mean Centering Strategy and 2-Category Race

District	Model 1	Model 2	Model 3a	Model 3b	Model 3c
C	0.104	0.106	0.105	0.105	0.105
H	0.246	0.242	0.241	0.241	0.240
M1	0.222	0.196	0.189	0.190	0.170
M2	0.249	0.231	0.233	0.233	0.223
M3	0.119	0.133	0.134	0.133	0.130
O2	0.137	0.156	0.154	0.155	0.152
O3	0.160	0.167	0.160	0.159	0.160

Table A37

Proportion of Teacher Rating Scale Variance that was Unexplained Between Teacher Variance across Models Using Grand Mean Centering Strategy and 6-Category Race

District	Model 1	Model 2	Model 3a	Model 3b	Model 3c
C	0.104	0.108	0.107	0.108	0.107
H	0.246	0.244	0.241	0.241	0.241
M1	0.222	0.189	0.173	0.189	0.190
M2	0.249	0.241	0.235	0.233	0.232
M3	0.119	0.130	0.130	0.132	0.130
O2	0.137	0.165	0.163	0.162	0.161
O3	0.160	0.165	0.158	0.158	0.154

Table A38

Proportion of Teacher Rating Scale Variance that was Unexplained Between Teacher Variance across Models Using Group Mean Centering Strategy and 2-Category Race

District	Model 1	Model 2	Model 3a	Model 3b	Model 3c
C	0.104	0.108	0.108	0.108	0.108
H	0.246	0.244	0.242	0.243	0.241
M1	0.222	0.189	0.191	0.191	0.192
M2	0.249	0.241	0.240	0.239	0.237
M3	0.119	0.130	0.129	0.129	0.126
O2	0.137	0.165	0.163	0.162	0.158
O3	0.160	0.165	0.157	0.157	0.157

Table A39

Plausible Values for Teacher Rating Scale from Final Model using Group Mean Centering and 6-Category Race by District

District	TRS Mean	TRS SD	90% Plausible Values
C	64.65	20.76	[53.53, 75.82]
H	57.52	25.09	[36.87, 77.64]
M1	35.33	10.94	[27.45, 43.47]
M2	35.41	11.41	[26.56, 44.80]
M3	25.30	11.00	[18.63, 31.97]
O2	3.28	2.80	[1.50, 5.04]
O3	2.87	2.76	[1.15, 4.67]

Note. TRS = Teacher Rating Scale.

Table A40

Plausible Values for Teacher Rating Scale from Final Model using Group Mean Centering and 2-Category Race by District

District	TRS Mean	TRS SD	90% Plausible Values
C	64.65	20.76	[53.61, 75.81]
H	57.52	25.09	[36.85, 77.67]
M1	35.33	10.94	[28.08, 42.82]
M2	35.41	11.41	[26.72, 44.67]
M3	25.30	11.00	[18.61, 31.89]
O2	3.28	2.80	[1.45, 5.08]
O3	2.87	2.76	[1.03, 4.77]

Note. TRS = Teacher Rating Scale.

Table A41

Plausible Values for Teacher Rating Scale from Final Model using Grand Mean Centering and 6-Category Race by District

District	TRS Mean	TRS SD	90% Plausible Values
C	64.65	20.76	[54.25, 76.93]
H	57.52	25.09	[39.66, 81.04]
M1	35.33	10.94	[28.04, 43.92]
M2	35.41	11.41	[27.03, 45.26]
M3	25.30	11.00	[19.39, 32.59]
O2	3.28	2.80	[1.58, 5.36]
O3	2.87	2.76	[1.01, 4.68]

Note. TRS = Teacher Rating Scale.

Table A42

Plausible Values for Teacher Rating Scale from Final Model using Grand Mean Centering and 2-Category Race by District

District	TRS Mean	TRS SD	90% Plausible Values
C	64.65	20.76	[54.42, 77.12]
H	57.52	25.09	[39.97, 81.31]
M1	35.33	10.94	[27.83, 43.71]
M2	35.41	11.41	[26.92, 45.23]
M3	25.30	11.00	[19.30, 32.40]
O2	3.28	2.80	[1.61, 5.35]
O3	2.87	2.76	[1.12, 4.82]

Note. TRS = Teacher Rating Scale.

Table A43

Variation in the Number of Students Identified as a Result of 0.33 SD Unit Differences in TRS

District	Always	Current	Additional	Total
C	708	163	182	1,053
H	950	272	329	1,551
M1	164	43	46	253
M2	157	31	38	226
M3	145	43	45	233
O2	209	53	58	320
O3	178	40	42	260
Total	2,511	645	740	3,896

Note. The Always column contains the number of students who would always be in the top 10% (using the original cut-off) regardless of whether their TRS were increased or decreased by 0.33 SD units. The Current column contains the number of students who would currently be identified, but would not be identified if their TRS scores were decreased by 0.33 SD units. The Additional column is the number of students who would not currently be identified, but would be identified if their TRS were increased by 0.33 SD units.