Ordinal Approaches to Decomposing Between-group Test Score Disparities

David M. Quinn
University of Southern California

Andrew D. Ho
Harvard Graduate School of Education

Quinn, D.M., & Ho, A.D. (2020). Ordinal Approaches to Decomposing Between-group Test Score Disparities. *Journal of Educational and Behavioral Statistics.* Published online ahead of print.

The final published version of the article can be found here.

Acknowledgements: We are grateful to Ben Shear for assistance with programming and Sean Reardon for providing feedback on an early draft.

Abstract

The estimation of test score "gaps" and gap trends plays an important role in monitoring educational inequality. Researchers decompose gaps and gap changes into within- and between-school portions to generate evidence on the role schools play in shaping these inequalities. However, existing decomposition methods assume an equal-interval test scale and are a poor fit to coarsened data such as proficiency categories. This leaves many potential data sources ill-suited for decomposition applications. We develop two decomposition approaches that overcome these limitations: an extension of *V*, an ordinal gap statistic, and an extension of ordered probit models. Simulations show *V* decompositions have negligible bias with small within-school samples. Ordered probit decompositions have negligible bias with large within-school samples but more serious bias with small within-school samples. More broadly, our methods enable analysts to (1) decompose the difference between two groups on any ordinal outcome into portions within- and between some third categorical variable, and (2) estimate scale-invariant between-group differences that adjust for a categorical covariate.

Key words: *ordinal decomposition, achievement gap, test score gap, decomposition, ordinal methods, simulation study*

Ordinal Approaches to Decomposing Between-group Test Score Disparities

The estimation of test score "gaps"[1] and gap trends by race/ethnicity and income plays an important role in monitoring educational inequality. Researchers decompose gaps and gap changes into within- and between-school portions (or within- and between school districts, states, etc.) to generate evidence on the role that schools plays in shaping these inequalities (e.g., Fryer & Levitt, 2004; Hanushek & Rivkin, 2006; Quinn, 2015; Reardon, 2008). If school quality differs by race, we would expect to see racial gaps widen between schools over time, net of out-of-school factors. If students from different racial/ethnic backgrounds have access to differentially effective resources or instruction within the same schools, we would expect to see gaps widen within schools over time. By comparing the relative proportion of gap-change that occurs within versus between schools, we can generate hypotheses about the most effective way to allocate resources in pursuit of educational equity. Of course, patterns observed in gap-change decompositions may result from a variety of in- or out-of-school factors. Nonetheless, for descriptive research about large-scale trends, these decompositions help to narrow the range of plausible hypotheses.

Existing decomposition methods have two important limitations, however. First, as parametric methods, they assume that the test scale has equal-interval properties, an assumption that is often questionable and difficult to verify (Domingue, 2014; Lord, 1975; Zwick, 1992). Second, these models are a poor fit when only a coarsened version of the test score distribution is available. This leaves many potential data sources ineligible for decomposition applications. For example, most publicly-available data from state standardized testing – including data from the EDFacts Assessment Database (U.S. Department of Education, 2015) - come in the form of counts of students scoring within some number of ordered proficiency categories such as "needs

improvement," "proficient," and "advanced" (Ho & Reardon, 2012; Reardon, Shear, Castellano, & Ho, 2017). Similarly, coarsened data such as Advanced Placement scores (1-5) and English proficiency exam scores (e.g., performance levels on the English Language Proficiency Assessment for California) are not well-suited for existing decomposition methods.

In the present article, we develop and evaluate methods that overcome these limitations. Our methods are invariant to monotonic scale transformations and can be applied when only a coarsened version of the test scale is available. This expands the possibilities for decomposition applications and provides a means of testing the sensitivity of parametric decompositions to assumptions about the interval nature of the test scale.[2]

While our motivating examples focus on decomposing racial/ethnic test score gaps into within-and-between school portions, our methods can be applied to decompose the difference between two groups on any ordinal outcome into portions within and between any third categorical variable. In education contexts, this includes decomposing between-group gaps on outcomes such as GPA, socioemotional learning, or test scores into portions within- and between-district, within- and between-country, and so on. The method applies to decompositions of gaps in any psychological or behavioral outcome into between- and within- components, such as gender gaps in organizational structures. Finally, in large-sample analyses where continuous covariates are treated as categorical variables for explanatory simplicity – such as socioeconomic status cut into deciles, parental education into categories, or household income into income brackets – these approaches enable scale-invariant estimation of covariate-adjusted gaps.

We begin by describing parametric frameworks for decomposing test score gaps. We then describe the interval scale assumption and its relevance to parametric decompositions. We introduce our strategy for exploiting existing ordinal gap-estimation methods to develop ordinal

decomposition approaches. We then describe the simulations through which we evaluate the viability of these methods before presenting results from simulations and real data applications.

Parametric Gap Decomposition Frameworks

When decomposing a racial test score gap across schools, the goal is to estimate the proportion of the gap that lies within schools versus the proportion that lies between schools. The relative sizes of these proportions provide clues as to how resources might be productively allocated. That is, would the overall gap narrow more if we were able to equalize group performance within schools, or if we equalized group performance between schools? Yet defining and estimating the within versus between-school proportions is not straightforward. Here, we describe two contrasting approaches to parametric decomposition and a third approach that unifies the two. We begin the article discussing these parametric decompositions because our proposed ordinal decompositions are adaptations of these parametric methods.

Let δ represent a test score gap; for illustration, consider the Black-White test score gap and, for simplicity, consider a population of students who are either Black or White:

$$Y_i = \alpha + \delta Black_i + \epsilon_i \qquad (1)$$

where Y_i is a test score (assumed to represent the latent trait with equal accuracy across groups), $Black_i$ is an indicator for whether student i is Black, and ϵ_i is a random error term. Although in reality most of the test score variation lies within, rather than between, groups, we can learn about the success of equity-focused policies by tracking the between-group variation. A common approach for decomposing δ into within- and between-school portions is the school fixed effects decomposition, also known as the Oaxaca decomposition (Oaxaca, 1973). This decomposition can be achieved with the model:

$$Y_{is} = \beta_1 Black_{is} + \gamma_s + \epsilon_{is} \qquad (2)$$

where γ_s represents a set of school fixed effects (constrained to a zero sum). In this model, β_1 is interpreted as the average within-school gap, leaving $\delta - \beta_1$ as the between-school gap.

Hanushek and Rivkin (HR, 2006) argued that it is generally incorrect, or at least misleading, to interpret the ratio β_1/δ as the proportion of the total gap that lies within schools. Closing the gaps within all schools would sensibly mean that 100% of the remaining gap lies between schools. However, reducing β_1 to zero would not necessarily result in a gap equal to $\delta - \beta_1$.[3] HR therefore offered an alternative decomposition that weights the contribution of each school based on its racial make-up and size:

$$\delta = \left(\sum_s \frac{n_{bs}}{n_b}\bar{Y}_s - \sum_s \frac{n_{ws}}{n_w}\bar{Y}_s\right) + \left(\left(\frac{1}{n_w} + \frac{1}{n_b}\right)\sum_s (\bar{Y}_{bs} - \bar{Y}_{ws})\alpha_s(1-\alpha_s)n_s\right) \quad (3)$$

where $\frac{n_{bs}}{n_b}$ and $\frac{n_{ws}}{n_w}$ are, respectively, the share of Black or White students who are in school *s*, $\bar{Y}_s$ is the mean test score in school *s*, $\bar{Y}_{bs}$ and $\bar{Y}_{ws}$ are race-specific school means, and α_s is the proportion of students in school *s* who are Black. The first parenthetical on the right-hand side is HR's between-school gap, and the second parenthetical is HR's within-school gap. The choice of decomposition matters: In the ECLS-K:99, the Oaxaca decomposition led to the conclusion that 37% of the Black-White math gap lay between schools in the spring of fifth grade, while the HR decomposition suggested that 79% lay between schools (Reardon, 2008). Similar discrepancies arise when decomposing the gap-widening from kindergarten to fifth grade.

Reardon (2008) introduced a three-part decomposition showing the mathematical relationship between the Oaxaca and HR decompositions. Reardon's decomposition is accomplished by first fitting the model:

$$Y_{is} = \beta_0 + \beta_1 Black_{is} + \beta_2 \overline{Black}_s + \epsilon_{is} \quad (4)$$

where $\overline{Black_s}$ is the proportion of students in school s who are Black and other terms are as defined above. The adjusted gap in this model, β_1, is equivalent to the within-school gap estimated from the school fixed effects (Oaxaca) model in (2). Reardon showed that the overall Black-White gap can be expressed as:

$$\delta = \beta_1(1 - VR) + \beta_1 VR + \beta_2 VR \quad (5)$$

where β_1 and β_2 are from (4). VR is the variance ratio index of segregation, which can be expressed as the difference in average school proportion Black between Black and White students. Reardon called the first term on the RHS of (5) the "unambiguously within school" gap ($Unambig\ Within \equiv \beta_1(1 - VR)$), the center term the "ambiguous" gap (Ambig $\equiv \beta_1 VR$), and the last term the "unambiguously between school" gap ($Unambig\ Btwn \equiv \beta_2 VR$). When the ambiguous portion is added to the unambiguously within-school portion, the decomposition is equivalent to the school fixed effects decomposition. When the ambiguous gap is added to the unambiguously between-school gap, the decomposition is equivalent to the HR decomposition. Because Reardon's decomposition provides a useful framework for illustrating our ordinal methods, we describe it in more detail here.

Reardon explains the decomposition through a series of graphs similar to our stylized depictions in Figures 1 and 2. In Figure 1, observations are students from a population of approximately 500 schools, the y-axis is student test score, and the x-axis is the proportion of the student body at the student's school who are Black. The solid gray line is the fitted line for White students, the solid black line is the fitted line for Black students[4], and the dashed black line gives the predicted overall school mean test score. The x-coordinate for the diamond on the fitted line for Black students is the mean school proportion Black for Black students (making the y coordinate the overall test score mean for Black students) and the x-coordinate for the circle on

the fitted line for White students is the mean school proportion Black for White students. The horizontal distance between the two points is VR. The vertical distance between them is δ.

Figures 1 and 2 enable us to describe gap components graphically, in terms of two different approaches to closing them. In Reardon's (2008) three-part decomposition, the Oaxaca within-school gap – which we will call the "total within-school gap" ($Total\ Within = Unambig\ Within + Ambig = \beta_1$) – would be closed by anchoring the fitted lines for White and Black students in Figure 1 to the same y-intercept (without changing their slopes; i.e., equalizing Black-White mean performance within school without altering the relationship between school proportion Black and test scores). This scenario is depicted in the left panel of Figure 2. After overlaying the fitted lines in this way, the remaining vertical distance between the circle and the diamond is the unambiguously between school gap. In the right panel of Figure 2, we depict the closing of the HR within-school gap (or the unambiguously within school gap), which is accomplished by overlaying the White and Black fitted lines on the prediction line for school means (i.e., equalizing group mean performance within schools without changing schools' mean performance). The remaining vertical distance between the circle and the diamond is the HR between-school gap, which we will call the "total between school gap" ($Total\ Btwn = Unambig\ Btwn + Ambig = (\beta_1 + \beta_2)VR$). The Venn diagram in the upper right-hand corner of Figure 1 shows that $Total\ Btwn = Unambig\ Btwn + Ambig$ (where $Unambig\ Btwn \equiv \beta_2 VR$ and $Ambig \equiv \beta_1 VR$) and $Total\ Within = Unambig\ Within + Ambig$ (where $Unambig\ Within \equiv\ \beta_1(1 - VR)$).

Again, from the perspective of decomposition, the parameters of interest are these terms as proportions of δ (that is, $Prop.\ Total\ Btwn = \frac{(\beta_1+\beta_2)VR}{\delta}$, $Prop.\ Total\ Within = \frac{\beta_1}{\delta}$). In the K-5 rounds of the ECLS-K:99 data, Reardon (2008) found that, for Black-White gaps across

math and reading, $Prop.\ Total\ Btwn$ ranged from .78 to .91, and $Prop.\ Total\ Within$ from .27 to .67. These numbers suggest that eliminating the association between school proportion Black and test scores would narrow the Black-White gap by 33-73% (i.e., $1 - Prop.Total\ Within$), while closing gaps within schools without changing schools' mean test scores would narrow the overall gap by 9-22% (i.e., 1- $Prop.\ Total\ Btwn$).

In these decompositions, the estimated total gap – and the ratio of the decomposed elements to the total gap – will depend on the scale of *Y*. That is, estimates of β_1 and β_2 - and estimates of $Prop.Total\ Within$ and $Prop\ \ Total\ Btwn$ - will differ with non-linear transformations of *Y*. Additionally, if only coarsened data are available, $\delta,\ Prop\ \ Total\ Btwn,$ and $Prop.Total\ Within$ cannot be estimated properly from the models above. These facts motivate our development of ordinal decomposition methods. Before introducing these methods, we discuss the equal-interval scale assumption on which the parametric methods rely.

The Equal-Interval Scale Assumption

Most commonly-used test score gap statistics are based on a mean difference between groups, requiring the assumption that the test metric is interval-scaled (Spencer, 1983). However, given that achievement tests measure a latent construct, the question of whether it is possible to completely confirm the interval scale assumption is a controversial one (Domingue, 2014; Lord, 1975; Zwick, 1992). If a test scale is not interval, the interpretation of a gap expressed in terms of a mean difference is unclear because units do not correspond to the same "amount" of the construct at each point along the scale (Ballou, 2009). Furthermore, if we cannot know whether a test scale is interval, we have no basis on which to prefer one scale over another nonlinear transformation of the scale (Reardon, 2008). This is troubling because nonlinear transformations

of scale may change the magnitude, or even the sign, of a test score gap (Ho, 2009; Spencer, 1983).

Rescaling tests can have a particularly dramatic effect on estimates of gap trends (Bond & Lang, 2013; Nielsen, 2015). For example, Bond & Lang (2013) found that in the ECLS-K:99, Black-White gap-widening from K to grade 3 ranged from .05 SD to .64 SD under monotonic scale transformations (compared to .35 SD in the baseline metric). As we show in online Appendix A using ECLS-K:2011 data, transformations can change the within-school and between-school ratios to the total gap change just as dramatically. Among the transformations that yield the most extreme total gap changes, the total-within gap change can be 1.8 times the total gap-change or 0.13 times in the opposite direction. The total-between gap change can be 0.25 times the total gap change or 0.95 times. Converting the scale to percentile ranks can yield substantively meaningful (though less extreme) differences. In the original theta scale, 25% of math gap-widening over K occurs between schools; this climbs to 40% when using percentile ranks. A transformation-invariant approach to decomposing test score gaps avoids such confusion.

Two Ordinal Approaches to Gap Decomposition

We propose two primary approaches to scale-invariant gap decompositions that use only the ordinal information contained in a test scale. One approach involves assuming a latent normal distribution underlying each school-by-race distribution and fitting ordered probit models (Reardon et al., 2017). These models estimate school-by-race means on an effectively ordinal scale that can then be manipulated to achieve Reardon's (2008) decomposition. (Lockwood et al. [2018] develop a Bayesian extension of Reardon et al.'s [2017] approach that overcomes some challenges faced by the latter's direct MLE estimators; for simplicity, we build on Reardon et

al.'s [2017] ordered probit models and reserve investigation of Lockwood et al.'s [2018] extensions for future research). We show that this works well under ideal conditions when school-by-race sample sizes are large. We introduce a second approach motivated by small-sample scenarios. This approach involves decomposing *V*, a scale invariant gap statistic (Ho, 2009), by applying the ordinal analogues to Reardon's (2008) parametric decomposition.

The approaches are conceptually similar. They both identify scale-invariant within- and between-school gap components, and they close gaps by relying only on the ordinal information contained in test scores. A benefit of the ordered probit approach is that it produces estimates equivalent to those from the parametric decomposition (model 5) when school-by-race distributions are respectively normal (i.e., a common transformation exists that renders all distributions normal). That is, the interpretation of the ordered probit decomposition can be mapped onto the parametric decomposition without having to make the same scale assumptions required by the parametric decomposition. However, we find that the ordered probit decomposition requires large school-by-race sample sizes. This disqualifies many data sources, including NCES studies such as the ECLS-K and the ELS. While the *V* decomposition does not have stringent sample size requirements, its parameters do not equal the Model 5 parameters under respective normality. As discussed in more detail below, an implication is that the *V* decomposition parameters and ordered probit decomposition parameters will generally differ. However, we show that the differences are small in magnitude across a range of scenarios.

Ordered Probit Decompositions

Our first decomposition approach builds on work by Reardon and colleagues, who showed that ordered probit models can be used to estimate means and standard deviations of test scores for groups of students when only coarsened proficiency data are available (Reardon et al.,

2017). Imagine some test scale *y* that has been coarsened along some set of cut scores and assume that *y* is respectively normal across groups. Denote the scale in which *y* is normal for each group (with population SD=1) as *y**. Given the counts of students falling within each score bin, an ordered probit model can be fit to estimate each group's mean and SD in the *y** scale:

$$\pi_{gb} = \Phi\left(\frac{\mu_g^* - c_{b-1}^*}{\sigma_g^*}\right) - \Phi\left(\frac{\mu_g^* - c_b^*}{\sigma_g^*}\right) \quad (6)$$

where π_{gb} is the proportion of students from group g whose scores fall in achievement bin b, μ_g^* and σ_g^* are the means and standard deviations in the y^* metric for group g, c^* is a cut score defining a boundary of the bin, and Φ is the standard normal CDF. Because no monotonic scale transformation will alter a student's ranking, estimates of μ_g^* and σ_g^* will not differ across monotonic transformations of *y*. Model 6 can be fit using a homoskedastic ordered probit model (HOMOP) assuming a common variance across groups, a heteroskedastic model (HETOP) in which each group is allowed its own variance, or a partially heteroskedastic model (PHOP) in which some groups are assumed to have a common variance and others are not.

When working with student-level test scores, the first step is to discretize the distribution to match Model 6. We coarsen into 10 ordered bins by decile. Next, use these numbered bins to estimate school-by-race means with an ordered probit model and store the model-estimated school-by-race means in the y^* metric – i.e., $\widehat{\mu_{B}^*}_s$ and $\widehat{\mu_{W}^*}_s$(for Black and White sub-groups, respectively). The overall population gap is then estimated as: $\hat{\delta}^* = \sum_s(\widehat{\mu_{B}^*}_s \frac{n_s^B}{n^B}) - \sum_s(\widehat{\mu_{W}^*}_s \frac{n_s^W}{n^W})$. To estimate the total between-school gap, first estimate each school's overall mean $\widehat{\mu_s^*} = (\widehat{\mu_{B}^*}_s \frac{n_s^B}{n_s} + \widehat{\mu_{W}^*}_s \frac{n_s^W}{n_s})$, and then find: $\widehat{total\ btwn} = \sum_s(\widehat{\mu_s^*} \frac{n_s^B}{n^B}) - \sum_s(\widehat{\mu_s^*} \frac{n_s^W}{n^W})$. Use the equality from

(5) to solve for the total within-school gap: $\widehat{total\ within} = \frac{\hat{\delta}^* - \widehat{TB}}{1-VR}$. Finally, for interpretation, transform $\widehat{total\ btwn}$ and $\widehat{total\ within}$ into proportions of the total gap ($\frac{\widehat{total\ btwn}}{\hat{\delta}^*}$, $\frac{\widehat{total\ within}}{\hat{\delta}^*}$).

***V* Decomposition**

Our second approach decomposes *V*, an ordinal gap statistic defined as (Ho, 2009):

$$V = \sqrt{2}\Phi^{-1}\big(P(X_a > X_b)\big) = \sqrt{2}\Phi^{-1}\int_{-\infty}^{\infty} F_b(x) f_a(x) dx \qquad (7)$$

where $P(X_a > X_b)$ is the probability that a randomly chosen student from group *a* scored higher than a randomly chosen student from group *b*, $F_b(x)$ is the CDF for group *b*, and $f_a(x)$ is the PDF for group *a*. When the distributions for *a* and *b* are normal, *V* equals Cohen's *d*. Because *V* is estimated using only information about students' ordinal rankings, it is invariant to monotonic scale transformations. *V* can be estimated when students' individual scores are available, or when only coarsened test score data are available (Ho & Reardon, 2012).

As described next, our strategy for decomposing *V* is to apply the ordinal analogues to Reardon's (2008) graphical decomposition described earlier.

Estimating the total between-school *V* gap ($\boldsymbol{V^{(TB)}}$). Recall that in Reardon's parametric decomposition, the total between-school gap can be interpreted as the gap that would remain if gaps were closed within schools without changing overall school mean achievement. The ordinal analogue to this would be to equalize the empirical probability mass functions (PMFs) by race within each school without altering the school's marginal empirical PMF. Conceptually, this ordinal analogue can be established by mapping the separate Black and White PMFs within each school to the school's marginal PMF and then estimating *V*. This *V* represents the ordinal analogue to the total between-school gap, or $V^{(TB)}$. We map these PMFs through the following the procedure (assuming student-level test scores as the starting point):

1) For computational efficiency, we divide the sample into ten test score bins by decile.

2) Within each school *s*, we find $p_{sb}^{(all)}$, the proportion of all students (regardless of race) whose scores fall into each bin *b* (so that $\sum_b p_{sb}^{(all)} = 1$ for each school). These proportions define the school's marginal empirical PMF.

3) We give each school-by-race subgroup a new PMF matching the marginal PMF for their school. We do this by creating, separately for each school-by-race subgroup, a set of weights for the 10 test score bins. Each weight essentially answers the question, "In school *s*, how many students from subgroup *g* (Black or White) would have a test score falling in bin *b* (where *b* is a value from 1 to 10) if the PMF for subgroup *g* in school *s* matched the marginal PMF for school *s*?" For each school *s* and each bin *b*: $weight_{sb}^{B} = p_{sb}^{(all)} \times n_s^B$; $weight_{sb}^{W} = p_{sb}^{(all)} \times n_s^W$, where *B* and *W* superscripts represent Black or White students.[5] See online Appendix B for illustrative examples.

4) We apply the weights from step 3 while estimating V:

$$V^{(TB)} = \sqrt{2}\Phi^{-1}\left(P\left(X_{bs}^{(B)} > X_{bs}^{(W)}\right)\right)$$

where X is the integer value for bin *b* (and each school *s* has a complete set of bin numbers 1-10 for each racial group, indexed by the superscripts). This estimates the total between-school *V* gap ($V^{(TB)}$) because the application of the bin weights closes gaps within schools without altering schools' marginal PMFs. That is, these weights give Black and White students in the same school identical PMFs, which are also identical to the school's original marginal PMF. This is the ordinal analogue to Reardon's (2008) total-between gap, where Black/White mean test score differences are eliminated within schools without changing schools' overall mean scores. For interpretability, we focus on the proportion of the total *V* represented by $V^{(TB)}$ (i.e., $\frac{V^{(TB)}}{V^{(total)}}$).

Estimating the unambiguously between-school V gap ($V_{btwn}^{(B\ to\ W)}$, $V_{btwn}^{(W\ to\ B)}$). Recall that in Reardon's (2008) graphical decomposition, the unambiguously between-school gap is the gap that remains after closing the total-within gap. As shown in the left panel of Figure 2, the total-within gap is closed by raising the fitted line for Black students to share the same y-intercept as the fitted line for White students. The ordinal analogue to this would be to map the Black PMF in each school *s* to the White PMF in each school *s.* We do this by creating a new set of bin weights in each school for Black students. For bin b in school *s*, $weight_{sb}^{(B\ to\ W)} = p_{sb}^{(white)} \times n_s^B$. In other words, $weight_{sb}^{(B\ to\ W)}$answers the question, "If the Black empirical PMF in school *s* matched the White empirical PMF in school *s,* how many of school *s'*s Black students would have scores falling in bin b?" We then apply these weights (with bins for White students weighted proportionately to their empirical PMF) to estimate *V.* The resulting *V* estimate, or $V_{btwn}^{(B\ to\ W)}$, represents an ordinal analogue to the unambiguously between-school gap ($i.e., \beta_2 VR$).

In the parametric regression, the location of the y-intercept does not affect the unambiguously between-school gap, such that we arrive at the same unambiguously between-school gap whether we raise the Black fitted line to the White fitted line or lower the White fitted line to the Black fitted line. In the *V* decomposition, however, we will generally obtain different ordinal unambiguously between-school gaps depending on which PMF is mapped to which (for example, it is possible that mapping the Black to White PMF in School A advances a large number of School A's Black students ahead of a large number of School B's White students, while mapping the White to Black PMF in School A leaves a large number of School A's Black students scoring below the large number of School B's White students)[6]. Despite this lack of symmetry, the method for PMF-mapping within-school can be chosen for substantive reasons. It is preferable to narrow gaps by helping the lower-scoring group achieve the higher-scoring

group's PMF rather than by lowering the scores of the higher-scoring group, and this is the approach we recommend. For symmetry, an ad hoc solution is to average the two approaches.

For thoroughness, we also estimate the ordinal analogue to the unambiguously between-school gap by mapping the White PMF within school to the Black PMF ($V_{btwn}^{(W\ to\ B)}$). The procedure is the same as that for estimating $V_{btwn}^{(B\ to\ W)}$, except that we use the weights: $weight_{sb}^{(W\ to\ B)} = p_{sb}^{(Black)} \times n_s^W$. Again, for interpretability, we focus on these parameters expressed as proportions of the total V (i.e., $\frac{V_{btwn}^{(W\ to\ B)}}{V^{(total)}}, \frac{V_{btwn}^{(B\ to\ W)}}{V^{(total)}}$).

Comparing Parameters across Approaches

The parameters estimated from the ordered probit models will equal those estimated from the Reardon decomposition as long as the data are generated from Model 4 with a normally-distributed error term (with the trivial difference that ordered probit models linearly scale y to a population SD of 1). The ordered probit model is more robust than Model 4, in the sense that Model 6 will also fit any monotonically transformed data from Model 4.

The *V* decomposition parameters are more difficult to map to Model 4 due to the re-estimation of V using weights that effectively close gaps in different ways within each school. First, V is only equal to a parametric Black-White gap when the Black and White distributions are both normal. The overall V, the denominator of the proportions of interest, will equal δ when the mixture distribution of White students across every school is normal, and the mixture distribution of Black students across every school is normal. For the decomposed *V* elements to match their parametric analogues, the mixture distributions of Black and White students across schools must also be normal after mapping CDFs within schools as required for any given estimate. Strictly speaking, these assumptions cannot all hold at once. However, mixture normal distributions are often very close to normal in practice, as evidenced in Table 1 (discussed

below). Thus, conceptually, the *V* decomposition parameters can be thought of as "loose ordinal analogues" to Reardon's parametric decomposition in the sense that we arrive at decomposition proportions by mapping empirical PMFs in a manner analogous to Reardon's graphical exposition of the parametric decomposition.

Study Goals

We conduct simulations to evaluate the bias and RMSE of these two ordinal decomposition methods. We simulate from a range of plausible population parameters to evaluate each method's performance under realistic conditions. In all simulations, our target estimands are the overall population gap and the proportions of that overall gap represented by the relevant decomposition elements. Again, a primary concern with the ordered probit methods is their viability when school-by-race sample sizes are small. We therefore compare ordered probit decompositions to *V* decompositions in small-samples, and compare ordered probit models across large- and small-sample scenarios. Given that analysts must often choose their models without knowledge of the true data-generating process, we are interested in how well the ordered probit model performs when its variance assumptions are incorrect. We therefore fit PHOP, HOMOP, and HETOP models to data for which the data-generating process is PHOP. For all simulation scenarios, we compare the performance of the methods assuming student-level test scores as the starting point, versus proficiency count data as the starting point. Finally, we apply these methods in two real data sets: 1) the ECLS-K:99, which has student-level test scores and small within-school samples, and 2) population proficiency-count data from Georgia.

Simulation Plan

The top panel of Table 1 shows the parameters for our main simulations, followed by three additional scenarios designed to test our methods under more extreme conditions (high

heteroskedasticity, high segregation, and small overall standardized gap). In all cases, we simulate from an imagined population comprised of only Black and White students. Columns 2-5 of Table 1 show the data-generating parameters we used in the original simulated *y* scale. For simplicity, we make the data-generating model partially heteroskedastic (PHOP), where within-school SDs differ by race but do not differ across schools for students of the same race (columns 4 and 5). Subsequent columns show the parameters of interest that we estimated, organized by the ordered probit parameters and the *V* parameters. Recall that the ordered probit models do not estimate moments in the *y* scale, but rather moments in the y^* scale (in which the full sample SD equals 1). Because we do not constrain the SD in the simulated *y* scale to equal 1, δ^* does not equal δ in the populations from which we simulate. However, $Prop.\ Total\ Within$ and $Prop.\ Total\ Btwn$ do not vary with scale changes across *y* and *y**.

In columns 9-12, we give the relevant population parameters for *V* decompositions (see online Appendix C for derivations of population parameters). As noted, the parameters estimated in the *V* decomposition do not generally equal those estimated in the ordered probit approaches. Consequently, in the populations from which we simulate, *V* equals neither δ nor δ^*, and the decomposition proportions for $V^{(TB)}$, $V_{btwn}^{(B\ to\ W)}$, and $V_{btwn}^{(W\ to\ B)}$ do not equal analogous decomposition proportions in y or y^*. However, as seen in Table 1, they are similar (compare column 9 to 6, 10 to 7, 11 and 12 to $1 - Prop.\ Total\ Within$). In many simulation conditions, proportions $V_{btwn}^{(B\ to\ W)}$ and $V_{btwn}^{(W\ to\ B)}$ are similar, except in cases with high heteroskedasticity.

Main Simulations

As seen in the top panel of Table 1, our main simulations assess the decomposition methods under four combinations of parameter values that vary the within- and between-school contributions to δ (the overall unstandardized population gap). To achieve this, we fix the level

of segregation and the size of δ across all simulation scenarios before choosing four combinations of values for the β_1 and β_2 parameters. Specifically, we fix segregation by establishing (for simplicity) a population with equal shares of three types of schools, all of which have the same number of students: schools in which 10%, 50%, and 90% of students are Black (resulting in a population VR of $.42\bar{6}$). We sample a total of 150 schools from this population, with equal representation of each school type in the sample (such that $\widehat{VR}$ always equals *VR*). In the metric of our simulated test scores (y), we always make $\delta = -1$, to vary the proportions that each decomposition element represents of δ while holding δ constant. We give each school-by-race subgroup a normal distribution with school-by-race means distributed as:

$$\begin{aligned} \mathrm{M}_s^{(B)} &\sim N(\mu = \beta_1 + \beta_2 P_s(B=1),\ \sigma = .026) \quad (7) \\ \mathrm{M}_s^{(W)} &\sim N(\mu = \beta_2 P_s(B=1),\ \sigma = .011) \end{aligned}$$

where $\mathrm{M}_s^{(B)}$ is the test score mean for Black students in school *s*, $\mathrm{M}_s^{(W)}$ is the test score mean for White students in school *s*, and $P_s(B=1)$ is the school's proportion Black (.10, .50, or .90). Note that school-by-race means co-vary with school proportion Black, as in Figure 1. As seen in column 2 of Table 1, the four values of β_1 (i.e., total within) that we simulate across are -.4, -.6, -.8, and -1. Given that $\delta = -1$, these values for β_1 correspond to scenarios in which total-within represents 40%, 60%, 80%, and 100% of δ, a reasonable range given empirical findings. For example, across the first four rounds of math and reading tests in the ECLS-K:2011, observed percentages range from 70% to 106% (see end note 3 for explanation of how total-within can exceed 100% of δ). For each scenario with a given β_1, we solve for β_2 using equation (5) (see third column of Table 1). This results in scenarios in which total-between represents roughly 77.1%, 65.6%, 54.1%, or 42.7% of δ (column 7 of Table 1; compare to a range of 52% to 71% across the first four rounds of math and reading scores in the ECLS-K:2011).

After sampling school-by-race means, we draw student scores according to the model:

$$Y_{is} = \mathrm{M}_s^{(B)}(Black_{is}) + \mathrm{M}_s^{(W)}(1 - Black_{is}) + \epsilon_{is}^{B}(Black_{is}) + \epsilon_{is}^{W}(1 - Black_{is}) \quad (8)$$

where $Black_{is}$ is an indicator for whether student *i* is Black, ϵ_{is}^{B} is the error term for Black students, $\epsilon_{is}^{B} \sim N(0, .97)$, and ϵ_{is}^{W} is the error term for White students, $\epsilon_{is}^{W} \sim N(0, .89)$. These within-school SDs approximate the observed within-school-by-race SDs in the ECLS-K:2011 (which range from .79 to .98 across the first 4 rounds when scores are standardized each round).

To evaluate the performance of the ordered probit models under small sample scenarios, we use a within-school sample size of n=30 (approximately the within-school sample size for the ECLS-K studies). For large-sample scenarios, we use within-school sample sizes of n=300, approximating a situation in which complete school data are available.

For each simulated sample, we first apply Reardon's (2008) parametric decomposition to find δ, $Prop.Total\ Btwn$, and $Prop.\ Total\ Within$ in the *y* metric. We then coarsen the data to either 10 bins (as described above, representing a situation in which the analyst begins with student-level data) or 4 bins (representing a situation in which the analyst has only coarsened data). When coarsening to 4 bins, we use cut scores at the 20th, 50th, and 80th percentiles. These cut scores produce lower RMSEs compared to others (Ho & Reardon, 2012); testing regimes with different benchmarks may produce larger RMSEs than those reported here. After each coarsening, we estimate $\hat{V}_{total}$ (using Stata's *rocfit* routine), and follow the procedures described above to estimate the proportional decompositions for $\hat{V}^{(TB)}$, $\hat{V}_{btwn}^{(B\ to\ W)}$ and $\hat{V}_{btwn}^{(W\ to\ B)}$. Finally, for each coarsening, we convert the data to a matrix in which rows are school-by-race subgroups and columns are test score bins, with cell values indicating the number of students from a school-by-race subgroup whose score fell in that bin. Using this matrix, we estimate school-by-race subgroup means in the *y** metric (using the *hetop* command in Stata [Shear & Reardon, 2017]).

As noted, although the data-generating model in our simulations is partially heteroskedastic (PHOP), we estimate school-by-race means by fitting HOMOP, HETOP, and PHOP models to compare performance when the analyst's assumptions do and do not match reality.[7] After estimating these subgroup means, we follow the procedures described above to estimate the overall gap and each proportional decomposition element. We use Stata-MP 14.

More Extreme Conditions

We also evaluate these methods under more extreme population parameters and sampling scenarios. As seen in the second panel of Table 1, we simulate scenarios for high heteroskedasticity, high segregation, small standardized gap, and a scenario that allows for sampling error in school proportion Black, both at the school level and within school. See online Appendix D for additional description.

Estimates and Reporting

For each scenario, we run 1,000 simulations and find the estimated bias (mean difference across simulations between the estimates and the true value), and test whether the estimated bias is significantly different from zero ($\alpha = .05$). We also estimate the RMSE (square root of the mean squared difference between estimates and the parameter value) for each set of simulations. For all decomposition proportions, we report bias and RMSE on the proportion scale; for estimates of the overall population gap, we report bias and RMSE based on a metric representing the population SD for the given simulation scenario (given that β_1 and β_2 change across simulation scenarios, the overall population SD changes while the within-school SDs do not). For estimates of the overall *V* gap, we report bias and RMSE in pooled (across school-by-race subgroups) SD units (which are constant across simulation scenarios).

Simulation Results

Parametric Decompositions

For the parametric decompositions in our main simulations, no bias estimate for a decomposition proportion was more extreme than .0005, and no bias estimate for the overall gap was more extreme than .0004 SD (none was significantly different from zero at $\alpha = .05$; note that the ratio estimator is approximately unbiased in large samples). In the scenario allowing for sampling error in school proportion Black, estimates for $Prop.\ Total\ Btwn$ are biased due to error in the estimate of VR (with an estimated bias of .0068 in our simulations). For the sake of conserving space, we do not show results for the parametric decompositions.

V Decompositions

In Table 2, we present the bias and RMSE results from the *V* decompositions, and in Figure 3, we present the results for the bias in decomposition proportions graphically. Figure 4 shows the bias in estimates of the total population gap across all methods, represented in method-relevant SD units. Across all simulations for the *V* decompositions, no model failed to converge.

As seen in Table 2 and Figures 3 and 4, estimated bias for the *V* approach is generally small in magnitude, ranging from -.003 to .005 for proportion $\hat{V}^{(TB)}$, -.002 to <.001 for proportion $\hat{V}_{btwn}^{(B\ to\ W)}$, and <.001 to .002 for proportion $\hat{V}_{btwn}^{(W\ to\ B)}$. Bias estimates for proportions $\hat{V}_{btwn}^{(B\ to\ W)}$ and $\hat{V}_{btwn}^{(W\ to\ B)}$ are never significantly different from zero, but bias estimates for $\hat{V}^{(TB)}$ often are. As expected, estimates of the overall *V* gap show no evidence of bias, with bias magnitudes ranging from -.001 to .001 pooled SD across scenarios.

RMSE estimates are generally somewhat smaller when data are coarsened to 10 (versus 4) bins. For proportion $\hat{V}^{(TB)}$, RMSEs range from .013 to .02 for 10 bins and .015 to .02 for 4 bins, ranges similar to the precision of the parametric $Prop.\ Total\ Btwn$ (which ranged from .014 to .017 with 30 students per school). The RMSEs were larger for proportions $\hat{V}_{btwn}^{(B\ to\ W)}$ and

$\hat{V}_{btwn}^{(W\ to\ B)}$. For $\hat{V}_{btwn}^{(B\ to\ W)}$, RMSEs ranged from .046 to .047 and .044 to .045 for 4 and 10 bins, respectively, while RMSEs for $\hat{V}_{btwn}^{(W\ to\ B)}$ were .048 with 4 bins and .045 with 10. These RMSEs were also larger than those for their analogous parametric proportion unambiguously-between, which was .024 (precision differences for the decomposition proportions likely explain the differences in statistical significance of the bias estimates noted above). Because *V* decompositions showed negligible bias with 30 students per school, we did not run simulations with 300 students per school due to the computational intensity for large sample sizes (our real data applications show that a single large sample run is feasible).

Ordered Probit Decompositions

In Tables 3 and 4, we present the results for the HOMOP and PHOP models, respectively. In Figure 5, we present the bias results for the decomposition proportions across all models for simulations with 30 students per school, and in Figure 6, we present the results for simulations with 300 students per school. We relegate the HETOP model results to online Appendix E because the models predictably encountered convergence issues (and showed greater bias in the proportional decompositions compared to HOMOP and PHOP). For HOMOP and PHOP, estimated bias is often significantly different from zero, though the magnitudes of the bias vary substantially across methods and simulation scenarios. With 300 students per school, bias is small in magnitude across all models; with 30 students, bias can reach more substantial levels. While the PHOP model is the correct model given our data-generating procedure, the PHOP and HOMOP models generally perform quite similarly under our primary simulation conditions.

HOMOP Models. As seen in Table 3 and Figure 5, the estimated bias for the HOMOP models with 30 students per school is largest for the total-within proportions with 4 bins (-.033 to

.048). Bias decreases with 10 bins (ranging from -.008 to .013), and is generally smaller for $Prop.\ Total\ Btwn$ (-.028 to .019 with 4 bins, -.008 to .004 with 10). Estimates of the overall gap show estimated bias ranging from -.021 to .006 SD across bin numbers. With 300 students per school, all bias estimates are negligible, never becoming more extreme than -.003 SD for the overall gap, or +/- .001 for the decomposition proportions. The HOMOP models converged across all simulations.

PHOP Models. In Table 4, we present the results for the PHOP models. As is most evident in Figures 5 and 6, the magnitudes and patterns of bias for the decomposition proportions in the PHOP model are nearly identical to those found with the HOMOP model. For the overall gap estimates, however, the PHOP model shows larger bias estimates compared to the HOMOP model (see Figure 4). For PHOP, estimates of bias in the overall gap range from -.08 to -.039 SD with 30 students per school. Bias is substantially smaller with 300 students per school, never getting more extreme than -.008 SD. All of the PHOP models converged when simulations included 300 students per school. The vast majority of PHOP models converged with 30 students per school, but some scenarios did not have perfect convergence (with the lowest number of converged models being 993 out of 1000).

More Extreme Conditions

In Table 5, we present results from the ordered probit models under more extreme conditions. These simulations use 30 students per school, 10 bins, and $\beta_1 = -.4$. Online Appendix E includes results with 300 students per school that also show minimal bias.

Under high heteroskedasticity, the advantage of the PHOP model over the HOMOP model for estimating the decomposition proportions is greater than the advantage observed in the main simulations, with bias estimates here of .006 (PHOP) vs. .011 (HOMOP) for

$Prop.\ Total\ Btwn$, and -.011 vs. -.02 for $Prop.\ Total\ Within$. However, the HOMOP model still performs better than PHOP when estimating the overall gap (bias of -.012 vs. -.058 SD).

With high segregation, the PHOP and HOMOP models showed slightly less bias for decomposition proportions compared to the simulations with lower levels of segregation; however, bias was high for PHOP when estimating the overall gap.

With sampling error added for school proportion Black, PHOP and HOMOP again perform similarly for estimates of the decomposition proportions, with biases of .009 and -.011 for proportions total-between and total-within, respectively. The bias for the overall gap estimate is large for PHOP, at -.062 SD (vs. -.015 for HOMOP). In this scenario, the variance ratio index of segregation (VR) is estimated as well (in contrast to the other simulation scenarios, in which there is no sampling error in VR). The estimated bias in $\widehat{VR}$ was .016, with an RMSE of .03.

With a small standardized gap, the models performed relatively well for estimating the overall gap. For the gap proportions, however, bias and RMSE were larger compared to the large standardized gap (for both HOMOP and PHOP, bias of .01 for $Prop.\ Total\ Btwn$ and -.018 for $Prop.\ Total\ Within$). RMSEs reached .067 and .117 for $Prop.\ Total\ Btwn$ and $Prop.\ Total\ Within$, respectively.

In Table 6, we present the results for the *V* decompositions under the more extreme conditions. As before, the *V* decompositions often show smaller estimated bias compared to analogous decomposition proportions from the ordered probit models, and for overall gap estimates. The RMSEs for the decomposition elements were especially large for the small standardized gap, reaching .062, .169, and .166 for $\hat{V}^{(TB)}$, $\hat{V}_{btwn}^{(B\ to\ W)}$, and $\hat{V}_{btwn}^{(W\ to\ B)}$, respectively.

Real Data Applications

We further tested each decomposition approach using actual data from two sources. To compare performance of the methods in real data with small within-school sample sizes and student-level test scores, we apply the decompositions to the NCES's Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K:99). To evaluate the methods when population data are available as proficiency counts, we apply the decompositions to statewide data from Georgia's state testing program.

ECLS-K:99 Student-level Data

The ECLS-K:99 used a three-stage sampling design to obtain a nationally-representative sample of students attending kindergarten over the 1998-99 school year. We use math test scores from Black and White students collected in the fall of K, spring of K, fall of first grade, and spring of first grade. We drop school-switchers and students missing data on race, school, or test score. In the fall of first grade, data were collected from a random sub-sample (~30%) of students. Predictably given the sample sizes, ordered probit models did not converge.[8]

We present the *V* decompositions results in Table 7. For reference, we include parametric results for the total gap, $Prop.\ Total\ Btwn$, and proportion unambiguously-between school from Reardon's decomposition (using theta scores standardized to SD=1 at each wave). As seen, the total *V* gaps are similar to the total parametric gaps with wave-standardized theta scores, reflecting the near-normal distribution of theta. The $\hat{V}^{(TB)}$ proportions are also similar to the $Prop.\ Total\ Btwn$. The proportions for $\hat{V}_{btwn}^{(B\ to\ W)}$ and $\hat{V}_{btwn}^{(W\ to\ B)}$ differ from the proportions for the parametric unambiguously between-school gaps. Recall that the parametric unambiguously between-school gap is $\delta - \beta_1$ (i.e., the difference between the total gap and the within-school gap estimated from a school fixed effects model), and that only integrated schools contribute to the estimate of β_1. In contrast, $\hat{V}_{btwn}^{(B\ to\ W)}$ and $\hat{V}_{btwn}^{(W\ to\ B)}$ are estimated with information from all

schools, rendering these ordinal decompositions incomparable to their parametric analogues. We therefore include proportions for $\hat{V}_{btwn}^{(B\ to\ W)}$ and $\hat{V}_{btwn}^{(W\ to\ B)}$ estimated after dropping mono-racial schools. As seen in Table 7, these estimates are often closer to the parametric proportions for the unambiguously between-school gaps than are the original $\hat{V}_{btwn}^{(B\ to\ W)}$ and $\hat{V}_{btwn}^{(W\ to\ B)}$ proportions. Furthermore, in the ECLS-K:99, proportions $\hat{V}_{btwn}^{(B\ to\ W)}$ and $\hat{V}_{btwn}^{(W\ to\ B)}$ differed more from each other for a given decomposition than they did in the simulations. With and without dropping mono-racial schools, differences between the two estimates for a given round reached a maximum of .269.

Georgia Proficiency Count Data

We use statewide testing data from Georgia for grades 3-8 from 2011-2014, for which Georgia released school-level counts of students from each racial group scoring in each of three proficiency categories (we use only data from Black and White students). We describe the results in text, and include the tables in online Appendix G.

Across grades and years, total *V* gaps ranged from -.63 to -.80, proportion $\hat{V}^{(TB)}$ ranged from .57 to .69, proportion $\hat{V}_{btwn}^{(W\ to\ B)}$ ranged from .38 to .57, and proportion $\hat{V}_{btwn}^{(B\ to\ W)}$ ranged from .48 to .67 (Table G2, online Appendix G). The differences in proportion $\hat{V}_{btwn}^{(W\ to\ B)}$ and proportion $\hat{V}_{btwn}^{(B\ to\ W)}$ for a given decomposition were larger than those in the simulations, but were often smaller than in the ECLS-K. In the Georgia data, the maximum difference across decompositions was .15.

HOMOP and PHOP models converged across all grades and years, but HETOP converged for only 3 of 24 grade/year combinations. HETOP can fail when subgroups have zero counts in one or more score bin (online Appendix G includes results with and without sample

restrictions enabling convergence). The HOMOP and PHOP estimates were often similar (differing in absolute value by at most .02 *sd* for the total gap, .005 for $Prop.\ Total\ Btwn$, and .009 for $Prop.\ Total\ Within$). The differences in estimates from the ordered probit compared to the *V* decompositions were larger, but recall that *V* and ordered probit models estimate slightly different parameters. Overall gaps from the ordered probit models differed from the overall *V* by up to .06, and proportion $V^{(TB)}$ differed from ordered probit $Prop.\ Total\ Btwn$ by up to .03.

Discussion

These results demonstrate the viability of scale-invariant test score gap decompositions that can be applied when: 1) the interval nature of a test scale is questionable, and/or 2) only ordered proficiency data are available. The *V* decomposition showed negligible bias across a range of scenarios with small within-school sample sizes, though RMSEs were large when the overall standardized gap was small and within-school sample sizes were small. With large within-school sample sizes, bias for the ordered probit decompositions was also negligible. However, bias from the ordered probit approach was larger with small within-school sample sizes. In the ECLS-K:99 (small within-school samples), ordered probit models did not converge, and HETOP often required sample restrictions in Georgia data (large within-school samples).

Several factors are relevant when deciding between the *V* versus ordered probit decomposition. When comparing the relative bias and RMSE of each approach, one should recall that the population values being estimated are not generally the same for each (though will often be similar in practice). One practical consideration concerns within-school sample size. As demonstrated in the ECLS-K:99, ordered probit models may not be an option when within-school samples are small. Even in population data (Georgia data), HETOP model convergence often required dropping some school-by-race subgroups (though HOMOP and PHOP

converged). In such cases, *V* decomposition is a viable alternative. As discussed below, Bayesian HETOP models may offer an alternative in small-sample scenarios.

When ordered probit models are an option, there may be reasons to prefer them over *V* decomposition. As noted, one appeal of the ordered probit decomposition is that its components provide direct interpretational correspondence to the parametric decomposition. That is, unlike *V*, the ordered probit decomposition proportion values match the parametric decomposition proportion parameters when school-by-race distributions are normal. As such, ordered probit decomposition estimates can be interpreted as representing a particular underlying latent distribution, without having to assume the observed scale expresses that underlying latent distribution (as required by the parametric decomposition).

When opting for an ordered probit decomposition, the analyst must decide among HOMOP, PHOP, and HETOP. In practice, it is rarely possible to know which model's assumptions are more appropriate for a given application. Our simulations showed that when within-school SDs are constant within but not across racial groups and the differences in SDs by racial group are relatively small, the (incorrect) HOMOP model can outperform the (correct) PHOP model. When within-school SD differences by race were larger (high heteroskedasticity scenario), however, the (correct) PHOP model outperformed the (incorrect) HOMOP model for decomposition proportions (but not the overall gap). Prior work (Reardon et al., 2017) has shown that when HETOP is the correct model, SD estimates from HOMOP will be biased. However, the HOMOP model will generally produce estimates with smaller RMSE compared to the HETOP model when group sample sizes are less than 100. The exact sample size at which HETOP performs better than HOMOP will depend on factors such as the location of cut scores and extent to which group variances differ. In our application with the Georgia data, the

HOMOP model always converged without sample restrictions, but the HETOP model seldom converged without sample restrictions. Even when data are not homoskedastic, then, HOMOP may be best in terms of bias and convergence.

When *V* and ordered probit decomposition are both viable, one should consider the relative bias and RMSE of each approach. Across most of our scenarios, the *V* approach produced decomposition proportions with lower levels of bias compared to the ordered probit models. Additionally, estimates of the total *V* do not show evidence of bias, while estimates of the total gap showed bias across ordered probit simulations. Estimates of the overall *V* gap and proportion $\widehat{V}^{(TB)}$ showed smaller RMSEs than their ordered probit counterparts.

An advantage of the *V* decomposition is that it does not assume that the relationship between school proportion Black and test scores is linear, or that it is the same for Black and White students (as in Model 4). The performance of the *V* decomposition under such interactive or nonlinear data-generating models is an area for future research.

For decomposition applications with small within-school sample sizes, a Bayesian approach using the Fay-Herriot HETOP (FH-HETOP) models proposed by Lockwood et al. (2018) may offer a viable alternative to the "direct estimates" computed by MLE in Reardon et al.'s (2017) HETOP models. FH-HETOP models offer a solution to the convergence challenges faced by direct estimates, with the trade-off of introducing conditional bias into the estimation of group parameters through shrinkage (Lockwood et al., 2018). Through future research, it will be valuable to explore the conditions under which FH-HETOP may allow for viable ordinal decompositions when direct estimates are unobtainable.

We have focused on bias and RMSE and have not addressed uncertainty estimates. We recommend this for future research. The possibility of decomposing gaps across multiple levels

is also worth exploring; for example, can the between-school gap be partitioned into portions within- versus between school districts? We encourage exploration of these decomposition methods in other applications, including decompositions of other educational outcomes (e.g., GPA, measures of socio-emotional learning), or gap decompositions by an explanatory factor variable. Examples include situations in which income categories, parental education levels, or free-or-reduced-price lunch status are treated as unordered categories, as is common in large-scale analyses (e.g., Burkam et al., 2004; Condron, 2009). One can estimate an income-adjusted ordinal Black-White gap by following our decomposition methods, with the adaptation of using income bins, rather than schools, as the grouping variable. Our methods could also be applied to other scales whose equal-interval properties attract scrutiny, including psychological scales for constructs like "happiness" (Bond & Lang, 2019). For example, an ordinal "happiness gap" between people with and without disabilities can be estimated after adjusting for income. Finally, although our motivating example is cross-sectional, the decomposition of gap-changes is an important application to isolate the gap dynamics that occur as students advance through school, or to isolate dynamics that occur over the school year versus summer.

References

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4,* 351–383. doi:10.1162%2Fedfp.2009.4.4.351

Bond, T. N., & Lang, K. (2013). The evolution of the Black-White test score gap in Grades K–3: The fragility of results. *Review of Economics and Statistics*, *95*(5), 1468-1479.

Bond, T. N., & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy.*

Burkam, D. T., Ready, D. D., Lee, V. E., & LoGerfo, L. F. (2004). Social-class differences in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education*, *77*(1), 1-31.

Condron, D. J. (2009). Social class, school and non-school environments, and black/white inequalities in children's learning. *American Sociological Review*, *74*(5), 685-708.

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika, 79(1),* 1-19. DOI: 10.1007/S11336-013-9342-4

Fryer, R. G., & Levitt, S. D. (2004). Understanding the Black-White test score gap in the first two years of school. *Review of Economics and Statistics*, *86,* 447–464. doi:10.1162/003465304323031049

Hanushek, Eric A., and Steven G. Rivkin. 2006. "School Quality and the Black-White Achievement Gap." (No. w12651). National Bureau of Economic Research.

Ho, A.D. 2009. A Nonparametric Framework for Comparing Trends and Gaps Across Tests. *Journal of Educational and Behavioral Statistics,* 34: 201-228.

Ho, A.D., & Reardon, S.F (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics, 37(4),* 489-517.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher, 35(7),* 3-12.

Lockwood, J.R., Castellano, K.E., & Shear, B.R. (2018). Flexible Bayesian models for inference from coarsened, group-level achievement data. *Journal of Educational and Behavioral Statistics, 43,* 663-692.

Lord, F. M. (1975). The "ability" scale in item characteristic curve theory. *Psychometrika, 20,* 299-326.

Nielsen, E.R. (2015). Achievement estimates and deviations from cardinal comparability. Federal Reserve Working Paper. Retrieved from: https://www.federalreserve.gov/econresdata/feds/2015/files/2015040pap.pdf

Oaxaca, R. (1973). Male-female wage differentials in in urban labor markets. *International Economic Review, 14(3),* 693-709.

Quinn, D.M. (2015). Kindergarten black-white test score gaps: Re-examining the roles of socioeconomic status and school quality with new data. *Sociology of Education, 88,* 120-139.

Quinn, D. M., Desruisseaux, T. M., & Nkansah-Amankra, A. (2019). "Achievement Gap" Language Affects Teachers' Issue Prioritization. *Educational Researcher, 48*(7), 484-487.

Reardon, Sean F. 2008. "Thirteen Ways of Looking at the Black-White Test Score Gap." Working paper, Stanford University. Retrieved from: http://www.stanford.edu/group/irepp/cgi-bin/joomla/working-papers.html

Reardon, S.F., Shear, B.R., Castellano, K.E., & Ho, A.D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics, 42(1),* 3-45.

Shear, B.R., & Reardon, S.F. (2017). hetop: Stata module for estimating heteroskedastic ordered probit models with ordered frequency data.

Spencer, B. (1983). On interpreting test scores as social indicators: Statistical considerations. *Journal of Educational Measurement, 20,* 317–333. doi:10.1111/j.1745-3984.1983.tb00210.x

U.S. Department of Education. (2015). State assessments in reading/language arts and mathematics: School year 2012-13 EDFacts Data Documentation. Washington, DC. Retrieved from http://www.ed.gov/edfacts

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational

achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Measurement, 20*, 299-326.

Notes

[1]We call attention to recent evidence that the term "achievement gap" may lead people to place less priority on educational inequality due to the term's association with deficit framing (Quinn et al., 2019). In some instances in this article we use the term "gap" because of its familiarity, but we encourage a shift in framing to focus on the structural inequities that lead to between-group disparities in test scores (Ladson-Billings, 2006). We use the term "racial disparity" in test scores as synonymous with the more commonly used "racial test score gap."

[2]We have created, and made available online via Open Science, Stata .ado files that perform the parametric and ordinal decompositions (https://osf.io/urx6b/).

[3] Hanushek and Rivkin (2006) illustrate with an example. Imagine a sample of 1,000 schools in which only one is integrated. β_1 will be estimated using only information from students in that school. If the gap there is large relative to the overall gap, it will appear as though the within-school portion of the overall gap is large, even though closing the gap in that school will do little to close the overall gap. This also illustrates how the within-school gap from a school fixed effects model can represent over 100% of δ, if the gap in that single school is greater than δ.

[4]To render the model analogous to the school FE model, Black/White fitted lines are constrained to be parallel.

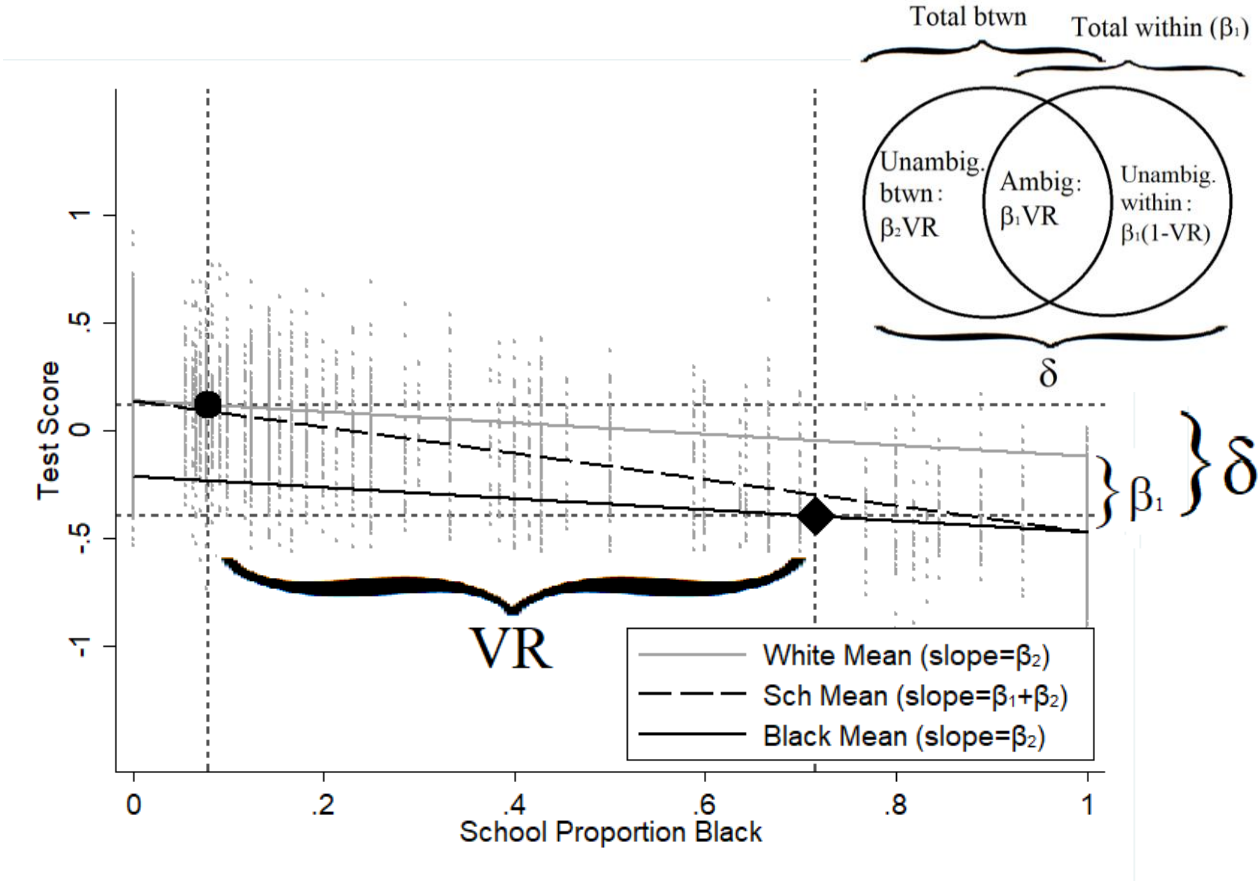
[5]The weights created for these methods are often not integer values. Because we estimate *V* using Stata's *rocfit* routine, which only allows integer weights, we multiply all weights by 100 and round to the nearest integer.

[6] Imagine School A has 4 Black students scoring at level 5 and 2 White students scoring at level 1. School B has 2 Black students scoring at level 4 and 4 White students scoring at level 2. Mapping the Black to White PMFs within these schools leaves 4 Black students in School A scoring above the 4 White students in School B, but mapping the White to Black PMFs within these schools leaves 2 Black students in School B scoring above the 4 White students in School A. These scenarios lead to different probabilities that a Black student will score above a White student.

[7] As written, the PHOP option for *hetop* allows the user to assume equal SD across groups with sample sizes below some minimum. We add constraints that allow us to assert common SD by race within schools.

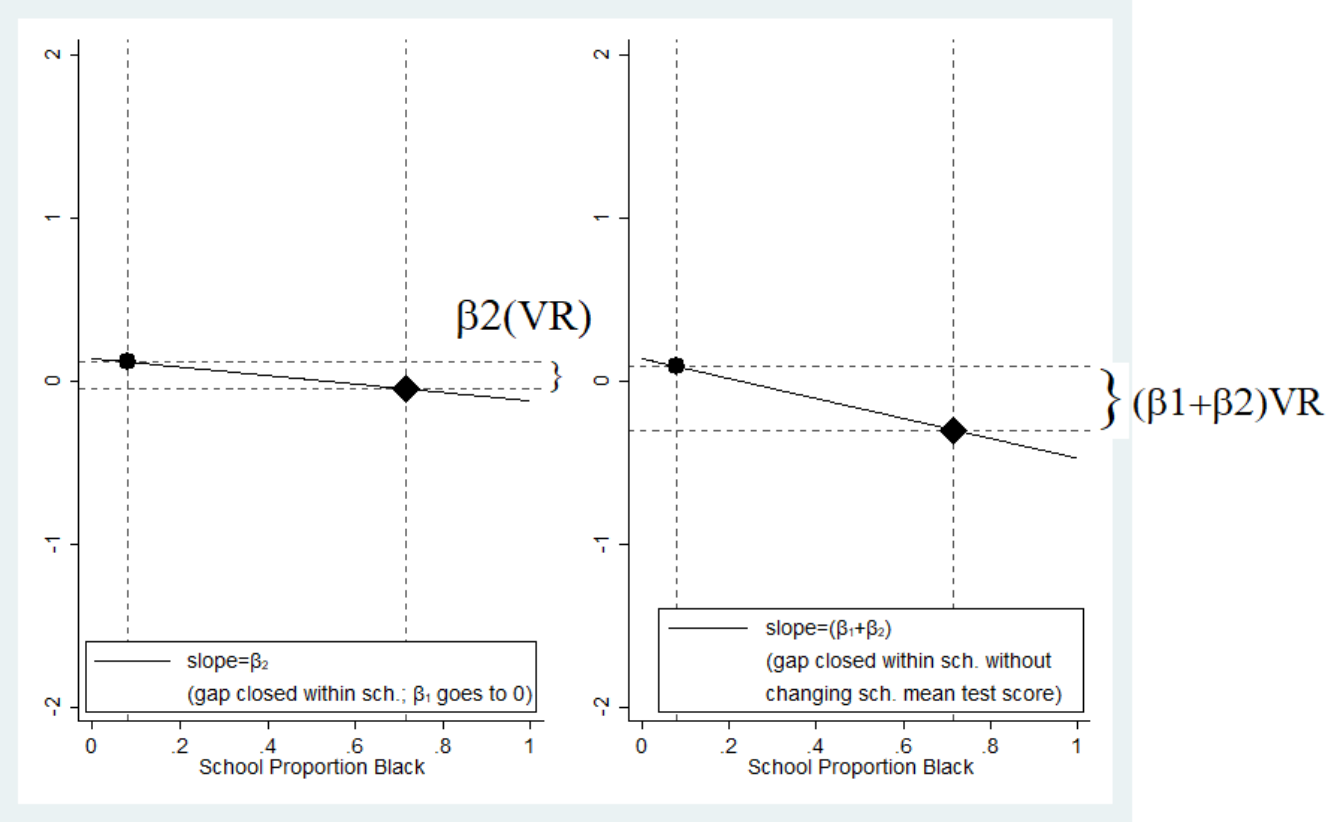
[8] The models converged when we restricted the sample to school-by-race cells with at least 20 students; however, this dropped all integrated schools.

Figure 1. Stylized Depiction of Reardon's (2008) Decomposition.



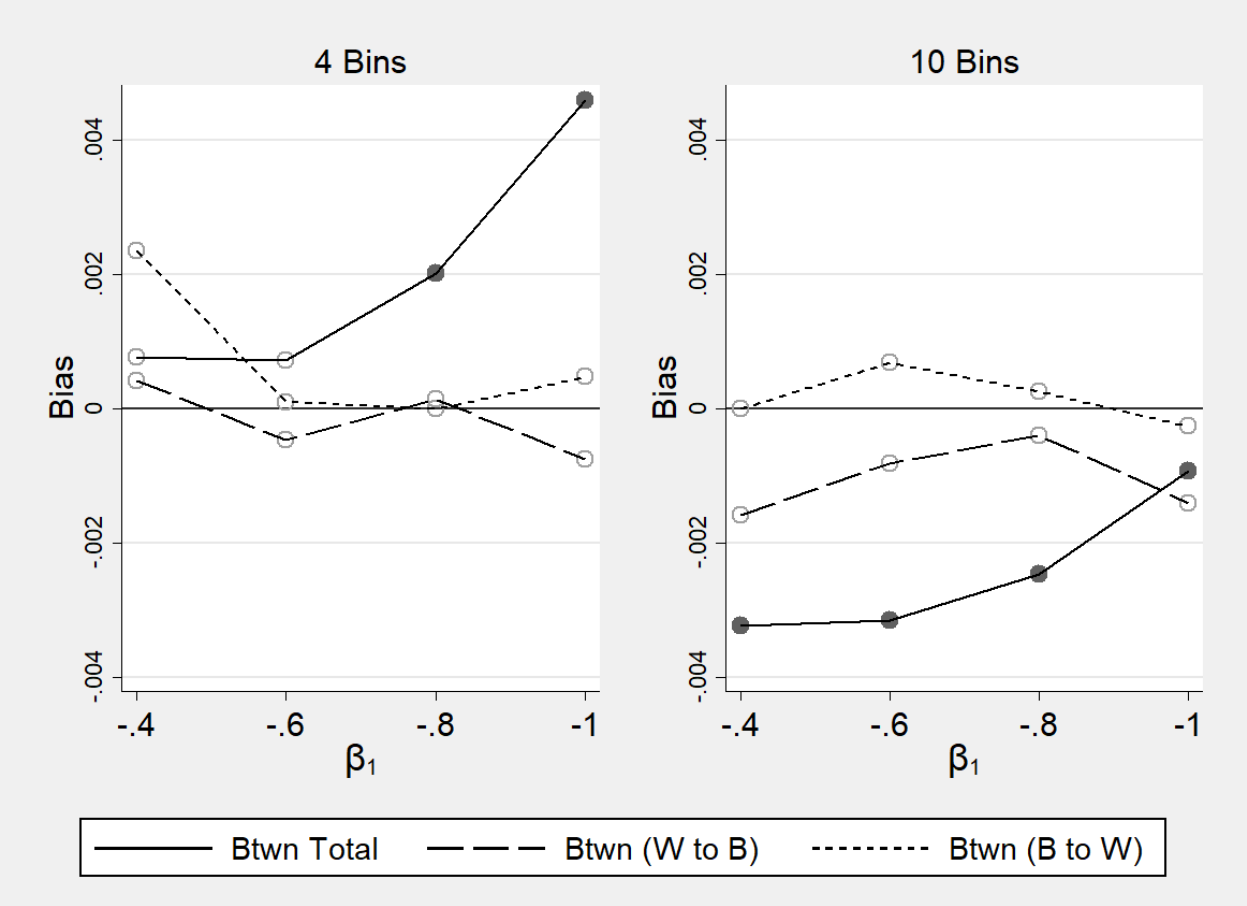
Note. Gray solid line is fitted line for White students, black solid line is fitted line for Black students, and black dashed line is predicted school means. Vertical distance between fitted lines for Black and White students is β_1, or the "total within-school" gap. Black circle on the fitted line for White students has as its x-coordinate the average school proportion Black for White students; its y-coordinate is overall mean test score for White students (with x- and y-coordinates identified by dotted gray lines). Black diamond on fitted line for Black students has as its x-coordinate the average school proportion black for Black students; its y-coordinate is overall mean test score for Black students (with these x- and y-coordinates identified by dotted gray lines). Vertical distance between these points equals overall Black-White test score difference, as indicated by δ. Horizontal distance between these two points equals the variance ratio index of segregation (VR).

Figure 2. Stylized Depiction of Reardon's (2008) Decomposition: Closing the "total within-school gap" (left) vs closing the "unambiguously within-school gap" (right).



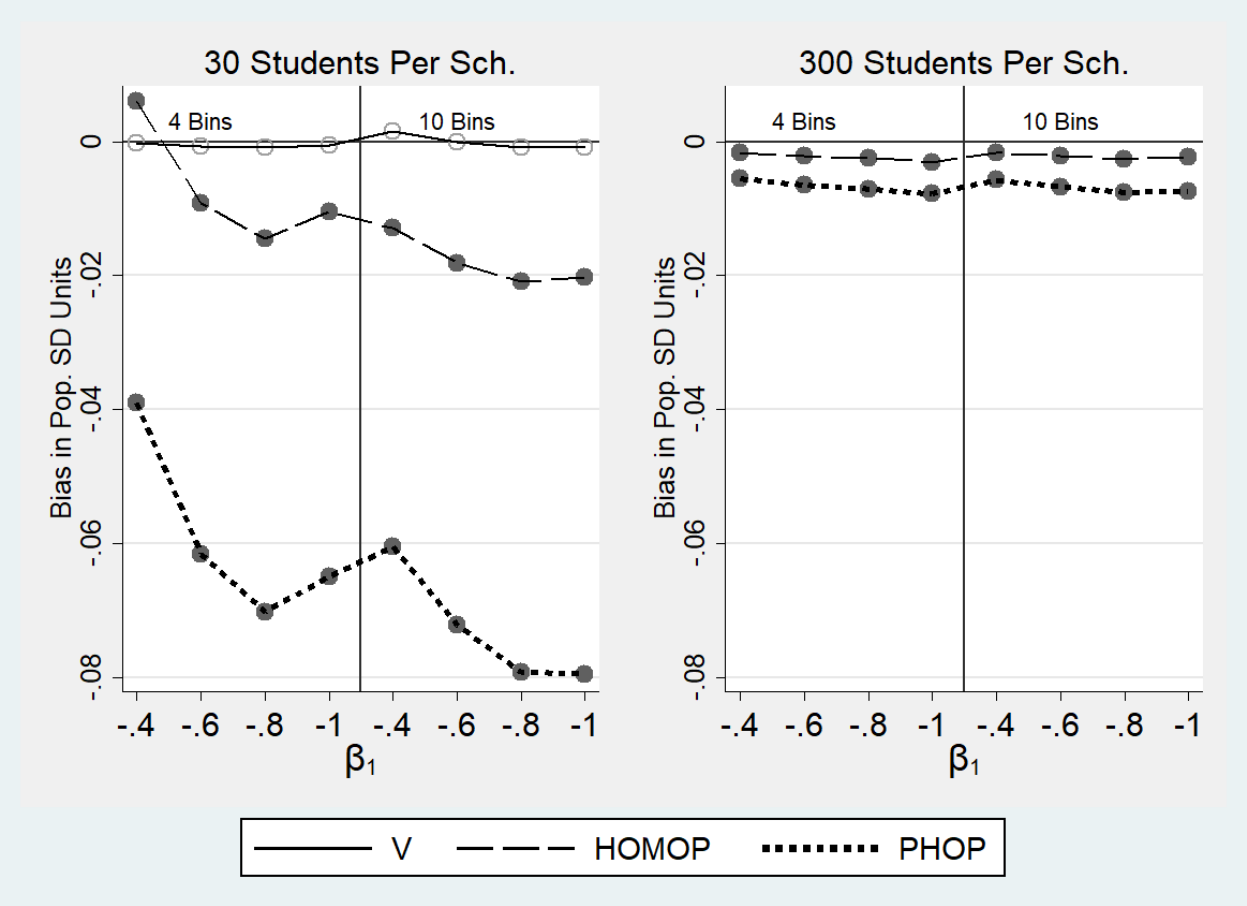
Note. VR=variance ratio index of segregation. In each panel, the x-coordinate for the black circle is the average school proportion Black for White students; the x-coordinate for the diamond is average school proportion Black for Black students. The y-coordinates for these points represent the average test score for White and Black students, respectively, after closing different portions of the overall gap. The left panel closes the total within-school gap, such that the remaining gap is the unambiguously between-school gap ($\beta_2 VR$). The right panel closes the unambiguously within-school gap, such that the remaining gap is the total between-school gap.

Figure 3. Estimated bias in *V* decomposition proportions by number of bins and β_1 values.



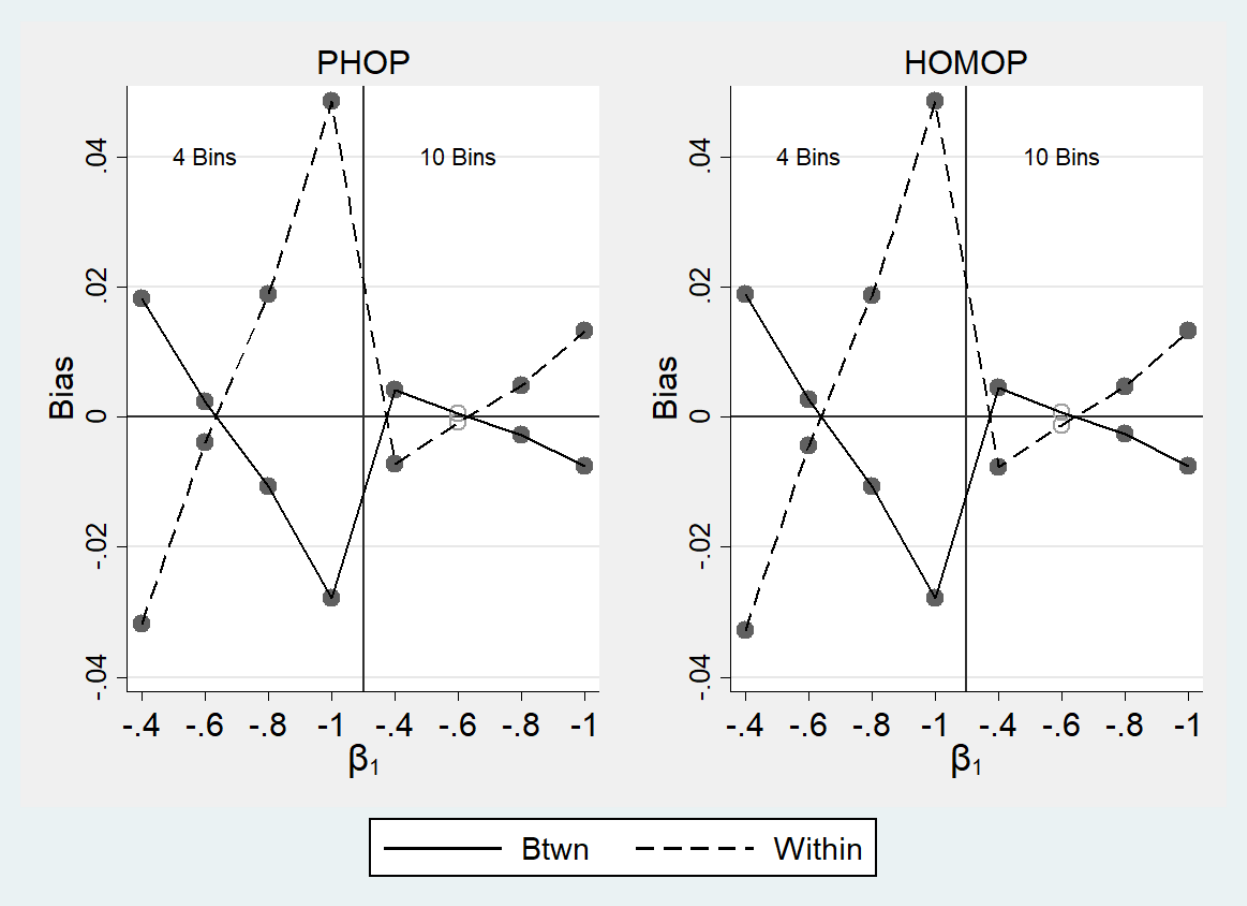
Note. Solid dots represent estimates that are significantly different from zero; hollow dots are not significantly different from zero. Estimates are based on 1,000 simulations per scenario, each with 150 schools with 30 students per school. With 10 bins, bins are equally-sized; 4 bins use cut scores at the 20th, 50th, and 80th percentiles. β_1= within-school gap in the simulated *y* metric from school fixed effects decomposition (total gap set to -1 in original test scale for all simulations).

Figure 4. Estimated bias in overall gap (*V* and δ^*) across bin numbers, values of β_1, and number of students per school.



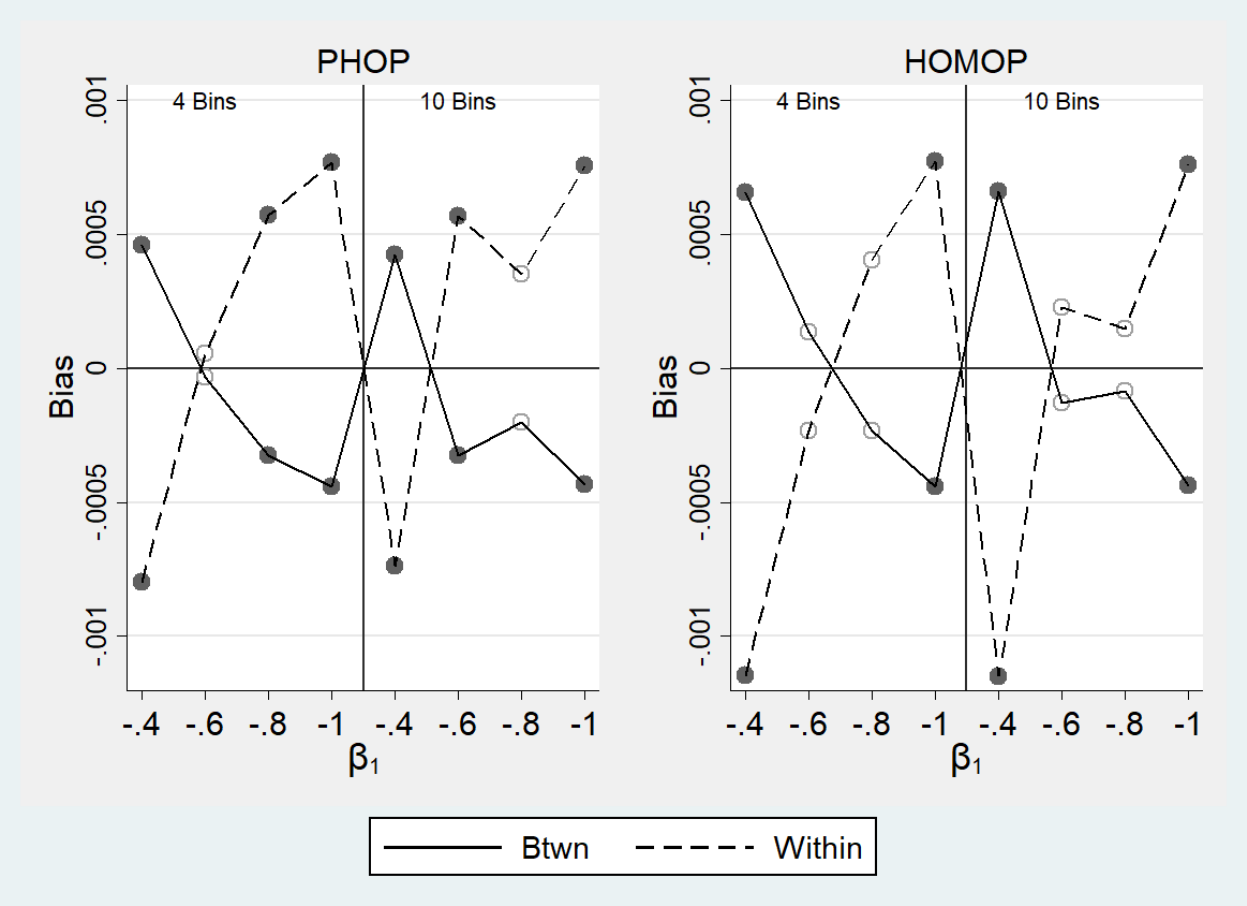
Note. Solid dots represent estimates that are significantly different from zero; hollow dots are not significantly different from zero. Estimates are based on 1,000 simulations per scenario, each with 150 schools. With 10 bins, bins are equally-sized; 4 bins use cut scores at the 20th, 50th, and 80th percentiles. β_1= within-school gap in the simulated *y* metric from school fixed effects decomposition (total gap set to -1 in original test scale for all simulations).

Figure 5. Estimated bias in ordered probit decomposition proportions across models, by number of bins and values of β_1; 30 students per school.



Note. Solid dots represent estimates that are significantly different from zero; hollow dots are not significantly different from zero. Estimates are based on 1,000 simulations per scenario, each with 150 schools. With 10 bins, bins are equally-sized; 4 bins use cut scores at the 20th, 50th, and 80th percentiles. Estimates for HETOP models with 4 bins are excluded from the graphs due to extreme levels of bias and low model convergence rates. β_1 = within-school gap in the simulated *y* metric from school fixed effects decomposition (total gap set to -1 in all simulations).

Figure 6. Estimated bias in ordered probit decomposition proportions across models, by number of bins and values of β_1; 300 students per school.



Note. Solid dots represent estimates that are significantly different from zero; hollow dots are not significantly different from zero. Estimates are based on 1,000 simulations per scenario, each with 150 schools. With 10 bins, bins are equally-sized; 4 bins use cut scores at the 20th, 50th, and 80th percentiles. β_1 = within-school gap in the simulated *y* metric from school fixed effects decomposition (total gap set to -1 in all simulations).

Table 1.
Data-Generating Parameters and Target Parameters for Recovery, across Simulation Scenarios and Decomposition Approaches

Data-generating Parameters (y scale)					Target Parameters for Recovery						
					Ordered Probit Parameters (y^* scale)			V Parameters			
	β_1	β_2	W/in sch SD, Black	W/in sch SD, White	δ^*	Prop. TB	Prop. TW	Overall V	Prop. $V^{(TB)}$	Prop. $V_{btwn}^{(B\ to\ W)}$	Prop. $V_{btwn}^{(W\ to\ B)}$
Main Simulations Ordered probit simulations: 30 or 300 students per school crossed with 4 or 10 coarsened bins V simulations: 30 students per school crossed with 4 or 10 coarsened bins											
Small β_1	-0.4	-1.406	.97	.89	-0.899	0.771	0.4	-1.009	0.742	0.628	0.582
Med. β_1	-0.6	-0.938	.97	.89	-0.924	0.656	0.6	-1.043	0.618	0.419	0.386
Large β_1	-0.8	-0.469	.97	.89	-0.941	0.541	0.8	-1.066	0.501	0.209	0.192
Fully β_1	-1	0.000	.97	.89	-0.946	0.427	1.0	-1.074	0.390	0.000	0.000
Extreme Conditions All simulations: 30 students per school and 10 coarsened bins											
High Het.	-0.4	-1.406	1.20	.80	-0.842	0.771	0.4	-0.931	0.748	0.747	0.519
High Seg.	-0.4	-0.920	.97	.89	-0.926	0.861	0.4	-1.051	0.840	0.758	0.731
Small Std. Gap	-0.4	-1.406	3.4	3.2	-0.298	0.771	0.4	-0.301	0.768	0.619	0.583

Note. Overall gap in the y metric (δ) is fixed to equal -1 for all scenarios. Prop.= proportion; TB=total between-school gap; TW=total within-school gap. Variance ratio index of segregation (VR) = .427 for all scenarios, except the high segregation scenario, in which VR=.652. High Het.= high heteroskedasticity condition; High seg.=high segregation condition; Small Std. Gap = small standardized gap condition.

Table 2.
Bias and RMSE from Simulated Estimates of *V* Decomposition Elements.

			Overall Gap V			Proportion Between Total $\hat{V}^{(TB)}$			Prop. Btwn (B to W) $\hat{V}_{btwn}^{(B\ to\ W)}$			Prop. Btwn (W to B) $\hat{V}_{btwn}^{(W\ to\ B)}$			
Num. Sts	Bins	β_1	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Converged
30	4	-0.4	0.000	0.033	0.859	0.001	0.020	0.233	0.000	0.047	0.783	0.002	0.048	0.121	1000
30	4	-0.6	-0.001	0.034	0.477	0.001	0.018	0.205	0.000	0.046	0.746	0.000	0.048	0.950	1000
30	4	-0.8	-0.001	0.035	0.439	0.002	0.015	<0.001	0.000	0.046	0.927	0.000	0.048	0.993	1000
30	4	-1	-0.001	0.035	0.593	0.005	0.015	<0.001	-0.001	0.047	0.608	0.000	0.048	0.757	1000
30	10	-0.4	0.001	0.032	0.141	-0.003	0.020	<0.001	-0.002	0.045	0.261	0.000	0.045	0.993	1000
30	10	-0.6	0.000	0.033	0.947	-0.003	0.017	<0.001	-0.001	0.044	0.551	0.001	0.045	0.631	1000
30	10	-0.8	-0.001	0.034	0.408	-0.002	0.015	<0.001	0.000	0.044	0.773	0.000	0.045	0.860	1000
30	10	-1	-0.001	0.034	0.428	-0.001	0.013	0.025	-0.001	0.044	0.314	0.000	0.045	0.849	1000

Note. Bias in overall gap is expressed in pooled SD units; bias in decomposition proportions are expressed in proportion units. Each scenario presents estimated bias and RMSE from 1,000 simulations. Converged = number of the 1,000 simulations for which models converged. Num. Sts. = number of students per school (in each of 150 schools, with equal representation from schools that are 10%, 50%, and 90% Black [vs. White]). Bins = number of bins that student-level data were coarsened to (bins of 10 are equally-sized; bins of 4 have cut scores at 20th, 50th, and 80th percentiles to represent a scenario in which only proficiency count data are available). β_1 = value of β_1 in data-generating model (see equations 7 and 8). p = p-value for test of null hypothesis that bias=0. For comparison, RMSEs in the parametric decompositions ranged from .014-.017 for proportion total between, .024 for proportion unambiguously between, and .027 for total gap.

Table 3.
Bias and RMSE from Simulated Estimates of HOMOP Decomposition Elements.

			Overall Gap			Proportion Total Between			Proportion Total Within			
Num. Sts	Bins	β_1	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Converged
30	4	-0.4	0.006	0.028	<0.001	0.019	0.031	<0.001	-0.033	0.053	<0.001	1000
30	4	-0.6	-0.009	0.026	<0.001	0.003	0.020	<0.001	-0.004	0.034	<0.001	1000
30	4	-0.8	-0.015	0.026	<0.001	-0.011	0.021	<0.001	0.019	0.037	<0.001	1000
30	4	-1	-0.010	0.024	<0.001	-0.028	0.035	<0.001	0.048	0.061	<0.001	1000
30	10	-0.4	-0.013	0.024	<0.001	0.004	0.019	<0.001	-0.008	0.033	<0.001	1000
30	10	-0.6	-0.018	0.027	<0.001	0.001	0.016	0.156	-0.001	0.028	0.156	1000
30	10	-0.8	-0.021	0.029	<0.001	-0.003	0.015	<0.001	0.005	0.026	<0.001	1000
30	10	-1	-0.020	0.029	<0.001	-0.008	0.017	<0.001	0.013	0.030	<0.001	1000
300	4	-0.4	-0.002	0.007	<0.001	0.001	0.006	<0.001	-0.001	0.010	<0.001	1000
300	4	-0.6	-0.002	0.007	<0.001	0.000	0.005	0.423	0.000	0.009	0.423	1000
300	4	-0.8	-0.002	0.008	<0.001	0.000	0.005	0.138	0.000	0.009	0.138	1000
300	4	-1	-0.003	0.008	<0.001	0.000	0.005	0.004	0.001	0.009	0.004	1000
300	10	-0.4	-0.002	0.007	<0.001	0.001	0.006	<0.001	-0.001	0.010	<0.001	1000
300	10	-0.6	-0.002	0.007	<0.001	0.000	0.005	0.410	0.000	0.009	0.410	1000
300	10	-0.8	-0.003	0.007	<0.001	0.000	0.005	0.566	0.000	0.008	0.566	1000
300	10	-1	-0.002	0.007	<0.001	0.000	0.005	0.004	0.001	0.008	0.004	1000

Note. Bias in overall gap is expressed in population SD units; bias in decomposition proportions are expressed in proportion units. Each scenario presents estimated bias and RMSE from 1,000 simulations. Converged = number of the 1,000 simulations for which models converged. Num. Sts. = number of students per school (in each of 150 schools, with equal representation from schools that are 10%, 50%, and 90% Black [vs. White]). Bins = number of bins that student-level data were coarsened to (bins of 10 are equally-sized; bins of 4 have cut scores at 20^{th}, 50^{th}, and 80^{th} percentiles to represent a scenario in which only proficiency count data are available). β_1= value of β_1 in data-generating model (see equations 7 and 8). p = p-value for test of null hypothesis that bias=0.

Table 4.
Bias and RMSE from Simulated Estimates of PHOP Decomposition Elements.

			Overall Gap			Proportion Total Between			Proportion Total Within			
Num. Sts	Bins	β_1	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Converged
30	4	-0.4	-0.039	0.048	<0.001	0.018	0.030	<0.001	-0.032	0.052	<0.001	995
30	4	-0.6	-0.062	0.066	<0.001	0.002	0.019	<0.001	-0.004	0.034	<0.001	1000
30	4	-0.8	-0.070	0.074	<0.001	-0.011	0.021	<0.001	0.019	0.037	<0.001	1000
30	4	-1	-0.065	0.069	<0.001	-0.028	0.035	<0.001	0.048	0.061	<0.001	1000
30	10	-0.4	-0.061	0.064	<0.001	0.004	0.019	<0.001	-0.007	0.033	<0.001	1000
30	10	-0.6	-0.072	0.075	<0.001	0.000	0.016	0.348	-0.001	0.028	0.348	999
30	10	-0.8	-0.079	0.082	<0.001	-0.003	0.015	<0.001	0.005	0.026	<0.001	999
30	10	-1	-0.080	0.082	<0.001	-0.008	0.017	<0.001	0.013	0.030	<0.001	993
300	4	-0.4	-0.006	0.009	<0.001	0.000	0.006	0.011	-0.001	0.010	0.011	1000
300	4	-0.6	-0.007	0.010	<0.001	0.000	0.005	0.852	0.000	0.009	0.852	1000
300	4	-0.8	-0.007	0.010	<0.001	0.000	0.005	0.036	0.001	0.009	0.036	1000
300	4	-1	-0.008	0.011	<0.001	0.000	0.005	0.005	0.001	0.009	0.005	1000
300	10	-0.4	-0.006	0.009	<0.001	0.000	0.006	0.018	-0.001	0.010	0.018	1000
300	10	-0.6	-0.007	0.009	<0.001	0.000	0.005	0.040	0.001	0.009	0.040	1000
300	10	-0.8	-0.008	0.010	<0.001	0.000	0.005	0.176	0.000	0.008	0.176	1000
300	10	-1	-0.007	0.010	<0.001	0.000	0.005	0.004	0.001	0.008	0.004	1000

Note. Bias in overall gap is expressed in population SD units; bias in decomposition proportions are expressed in proportion units. Each scenario presents estimated bias and RMSE from 1,000 simulations. Converged = number of the 1,000 simulations for which models converged. Num. Sts. = number of students per school (in each of 150 schools, with equal representation from schools that are 10%, 50%, and 90% Black [vs. White]). Bins = number of bins that student-level data were coarsened to (bins of 10 are equally-sized; bins of 4 have cut scores at 20^{th}, 50^{th}, and 80^{th} percentiles to represent a scenario in which only proficiency count data are available). β_1 = value of β_1 in data-generating model (see equations 7 and 8). p = p-value for test of null hypothesis that bias=0.

Table 5.
Bias and RMSE from Simulated Estimates of PHOP and HOMOP Decomposition Elements under Extreme Conditions.

	Overall Gap			Proportion Total Between			Proportion Total Within			
	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	P	Converged
	High Heteroskedasticity									
PHOP	-0.058	0.062	<0.001	0.006	0.022	<0.001	-0.011	0.038	<0.001	999
HOMOP	-0.012	0.025	<0.001	0.011	0.023	<0.001	-0.020	0.041	<0.001	1000
	High Segregation									
PHOP	-0.059	0.063	<0.001	0.003	0.015	<0.001	-0.003	0.043	0.018	1000
HOMOP	-0.016	0.025	<0.001	0.003	0.015	<0.001	-0.004	0.043	0.007	1000
	Sampling Error for Segregation and School Proportion Black									
PHOP	-0.062	0.066	<0.001	0.009	0.023	<0.001	-0.011	0.035	<0.001	1000
HOMOP	-0.015	0.027	<0.001	0.009	0.024	<0.001	-0.011	0.036	<0.001	1000
	Small Standardized Gap									
PHOP	-0.028	0.041	<0.001	0.010	0.067	<0.001	-0.018	0.117	<0.001	1000
HOMOP	-0.005	0.028	<0.001	0.010	0.067	<0.001	-0.018	0.117	<0.001	1000

Note. Bias in overall gap is expressed in population SD units; bias in decomposition proportions are expressed in proportion units. Each scenario presents estimated bias and RMSE from 1,000 simulations. Converged = number of the 1,000 simulations for which models converged. All simulation scenarios use 150 schools with 30 students per school, data coarsened to 10 equally-sized bins, and $\beta_1 = -.4$ in data-generating model (see equations 7 and 8). *p* = *p*-value for test of null hypothesis that bias=0. See Table 1 for parameters used in data-generating models.

Table 6.
Bias and RMSE from Simulated Estimates of V Decomposition Elements under Extreme Conditions.

	Overall Gap V			Proportion Between Total $\hat{V}^{(TB)}$			Prop. Btwn (B to W) $\hat{V}_{btwn}^{(B\ to\ W)}$			Prop. Btwn (W to B) $\hat{V}_{btwn}^{(W\ to\ B)}$			
	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Converged
High heteroskedasticity	0.002	0.032	0.117	0.000	0.021	0.645	-0.003	0.048	0.023	0.002	0.050	0.179	1000
High segregation	0.004	0.033	<0.001	-0.003	0.017	<0.001	0.000	0.040	0.938	-0.002	0.041	0.170	1000
Sampling error	0.004	0.035	0.001	0.004	0.024	<0.001	0.007	0.052	<0.001	0.011	0.054	<0.001	1000
Small std. gap	0.000	0.032	0.939	0.003	0.062	0.198	0.003	0.169	0.638	0.008	0.166	0.141	1000

Note. Prop.=proportion. Bias in overall gap is expressed in pooled population SD units; bias in decomposition proportions are expressed in proportion units. Each scenario presents estimated bias and RMSE from 1,000 simulations. Converged = number of the 1,000 simulations for which models converged. All simulation scenarios use 150 schools with 30 students per school, data coarsened to 10 equally-sized bins, and $\beta_1 = -.4$ in data-generating model (see equations 7 and 8). p = p-value for test of null hypothesis that bias=0. See Table 1 for parameters used in data-generating model.

Table 7.
Parametric and *V* Decompositions for Black-White Math Test Score Gaps across Waves in the ECLS-K:99.

	Fall K		Spring K		Fall Grade 1		Spring Grade 1	
	Param	V	Param	V	Param	V	Param	V
Total Gap	-0.693	-0.753	-0.734	-0.771	-0.663	-0.691	-0.746	-0.811
Decomp proportions								
TB, $\hat{V}^{(TB)}$	0.773	0.756	0.764	0.741	0.753	0.732	0.770	0.730
Unambig. Btwn	0.326		0.316		0.334		0.331	
$\hat{V}_{btwn}^{(W\ to\ B)}$		0.741		0.763		0.842		0.756
$\hat{V}_{btwn}^{(B\ to\ W)}$		1.024		0.991		0.948		0.955
$\hat{V}_{btwn}^{(W\ to\ B)}$ (no seg.)		0.352		0.272		0.628		0.249
$\hat{V}_{btwn}^{(B\ to\ W)}$ (no seg.)		0.400		0.254		0.458		0.143
Mean $\overline{Black_s}$ (Black, White)	0.728	0.066	0.722	0.067	0.705	0.075	0.723	0.066
N (Black White)	1838	7601	2015	8395	651	2547	2008	8388
Sch-level N	838		903		282		900	

Note. Param=parametric estimates. Parametric gaps are in wave-standardized theta scores. Data are coarsened to 10 bins for *V* decompositions. Ordered probit models failed to converge. $\overline{Black_s}$ = school proportion Black. No seg. = schools with mono-racial samples dropped. K=kindergarten. Note that Prop. $\hat{V}_{btwn}^{(W\ to\ B)}$ and Prop. $\hat{V}_{btwn}^{(B\ to\ W)}$ can be greater than Prop. $\hat{V}^{(TB)}$ if there are enough majority-Black schools whose small number of White students are low-scoring.

Appendix A. Sensitivity of Gap Change Decompositions to Scale Transformations.

In Table A1, we show the sensitivity to scale transformations of decompositions of the standardized Black-White math gap change over kindergarten in the ECLS-K:2011. Our sample includes only Black and White students with fall and spring K test scores who did not change schools over the K school year. Across various monotonic transformations, we show the minimum and maximum total standardized gap change along with the ratio of total-within gap change to total gap change and the ratio of total-between gap change to total gap change that were obtained under the respective transformation.

As a reference, the top row shows these decompositions for the original theta scores, which have been standardized to a mean of 0 and SD of 1 at each wave. The next two rows show decomposition results using the theta scores after applying skewness-inducing exponential transformations (see Reardon & Ho, 2015) and then re-standardizing the scores. Skewness ranges from -2 to 2 were motivated by observable skewness ranges in state test score data (Ho & Yu, 2015). Across transformations, the total gap can vary from shrinking by .11 SDs to increasing by .15 SDs. Under the transformation that yields a shrinking of .11 SD, the total-within:total change ratio is larger than the total-between:total change ratio, at 1.01 and .57, respectively. In contrast, under the transformation that yields a gap-narrowing of .15, these decomposition ratios flip. Here, the total-within:total change ratio is smaller than the total-between: total change ratio, at .13 and .95, respectively.

Inspired by Bond and Lang (2013), we apply exponential transformations to the theta scores (though unlike Bond and Lang, we apply monomial transformations for simplicity). We first add a constant to the theta scale to ensure positive values for all scores, then transform scores as: $\theta' = \theta + \theta^n$, with n ranging from 2 to 8, and finally re-standardize scores. Across

these transformations, the total gap change ranges from -.28 to -.06 SD. When total change is -.28 SD, the total-within:total gap change ratio is .81; this ratio increases to 1.5 when total gap change is -.06 SD. The total-between:total gap change ratio moves in the other direction. When total change is -.28 SD, this ratio is .66; when total gap change is -.06 SD, the ratio is .36. Our ordinal decomposition methods produce gap statistics that are constant across all of these scale transformations.

Table A1.
Total Black-White Gap Changes and Change Decompositions under Various Scale Transformations, Fall - Spr. Kindergarten (ECLS-K:2011).

	Total Gap Change (fall to spring)	Ratio of "Total Within" Change to Total Gap Change	Ratio of "Total Between" Change to Total Gap Change
Standardized Theta Scores	-0.050 (-.589 to -.639)	1.791	0.245
Percentile Rank of Theta (Std.)	-.071 (-.577 to -.648)	1.415	0.404
Skew Transformations to Theta (Std.)			
Min (c=0.5)	-0.114 (-.481 to -.595)	1.016	0.572
Max (c=-0.5)	0.150 (-.516 to -.366)	0.126	0.947
Exponential Transformations to Theta (Std.; up to 8th power)			
Min (8^{th} power)	-0.278 (-.27 to -.548)	0.807	0.660
Max (2^{nd} power)	-0.059 (-.589 to -.639)	1.523	0.358

Note. For skew and exponential transformations, top row shows total gap change and decomposition ratios in original scale for reference. Subsequent rows show the minimum or maximum total gap change estimated across a range of different scale transformations, along with the decomposition ratios obtained with that transformation. Ranges given in parentheticals in "total gap change" column show gaps in fall and spring of K. All scores are wave-standardized to mean=0, SD=1 after transformations. Exponential transformations are performed after adding a constant to all theta scores, ensuring positive values and monotonicity; transformations take the form $\theta' = \theta + \theta^n$, with n ranging from 2 to 8 (powers for min and max gap change given in table). Values for *c* parameter in skew transformations yielding min and max total gap change listed in table (see Reardon & Ho [2015]). Black n =1894; White n=7193; school n=800. Sampling weight W1C0 applied.

Appendix B. Illustration of Weighting Procedure for V Decomposition.

Table B1. Bin weights for Black and White students for sample school with 50% Black and 50% White students when n=30 per school.

Bin	Number White Sts.	Number Black Sts.	Total number sts	School's Marginal PMF	Bin Weight Black	Bin Weight White
1	2	1	3	0.10	1.5	1.5
2	0	3	3	0.10	1.5	1.5
3	0	2	2	0.07	1	1
4	3	4	7	0.23	3.5	3.5
5	3	2	5	0.17	2.5	2.5
6	0	1	1	0.03	0.5	0.5
7	0	0	0	0.00	0	0
8	3	2	5	0.17	2.5	2.5
9	1	0	1	0.03	0.5	0.5
10	3	0	3	0.10	1.5	1.5
Sum	15	15	30	1	15	15

Table B2. Bin weights for Black and White students for sample school with 90% Black and 10% White students when n=30 per school.

Bin	Number White Sts.	Number Black Sts.	Total number sts	School's Marginal PMF	Bin Weight Black	Bin Weight White
1	0	2	2	0.07	1.8	0.2
2	1	4	5	0.17	4.5	0.5
3	0	3	3	0.10	2.7	0.3
4	0	1	1	0.03	0.9	0.1
5	0	2	2	0.07	1.8	0.2
6	0	0	0	0.00	0	0
7	1	6	7	0.23	6.3	0.7
8	1	3	4	0.13	3.6	0.4
9	0	3	3	0.10	2.7	0.3
10	0	3	3	0.10	2.7	0.3
Sum	3	27	30	1	27	3

Appendix C. Calculating Population Values for *V* Decomposition

For each value of school probability Black in our population, school-by-race means are normally distributed in the original *y* metric. Therefore, assuming equally-sized schools for simplicity, the overall (i.e., across schools) White mean at a given school probability Black is $\beta_2\left(E\left(\overline{Black_s}^{(White)}\right)\right)$ and the overall Black mean is $\beta_2\left(E\left(\overline{Black_s}^{(Black)}\right)\right)+\beta_1$. By the law of total variance, the overall race-specific variances (conditional on school probability Black) are sums of the within-school variances and the variance of the means. After calculating the overall race-specific means and variances at each school probability Black (recall that school-by-race means are normally distributed, conditional on school proportion Black), we find the overall population CDF for White students by applying the formula for the CDF of the mixture of normal distributions (we use the R package *nor1mix* to find all mixture distributions):

$$F_{white}(x)=\sum_p w_p^{(w)}\Phi\left(\frac{x-\mu_p^{(w)}}{\sigma_p^{2(w)}}\right) \quad \text{(B1)}$$

where *p* indexes a particular school probability black, Φ is the normal CDF, $w_p^{(w)}$ is a weight giving the proportion of the total White population that attends schools with school proportion Black *p* (or $w_p^{(w)}=\frac{1-P(b)_p}{\sum_p(1-P(b)_p)}$), $\mu_p^{(w)}$ is the true White mean across schools with probability Black *p*, *a*nd $\sigma_p^{2(w)}$ is the true White variance across schools with probability Black *p*.

Similarly, we find the overall population PDF for Black students using the formula for the PDF of a mixture of normals:

$$f_{black}(x)=\sum_p w_s^{(b)}\phi\left(x,\mu_s^{(b)},\sigma_s^{2(b)}\right) \quad \text{(B2)}$$

where ϕ is the normal PDF and the weight $w_s^{(b)}=\frac{P(b)_s}{\sum_s P(b)_s}$. This gives a total *V* of:

$$V_{total} = \sqrt{2}\Phi^{-1}\int_{-\infty}^{\infty}\left(\sum_p w_s^{(w)}\Phi\left(\frac{x-\mu_s^{(w)}}{\sigma_s^{2(w)}}\right)\right)\left(\sum_p w_s^{(b)}\phi\left(x,\mu_s^{(b)},\sigma_s^{2(b)}\right)\right)dx \quad \text{(B3)}$$

To solve for the true value of the total between-school V (i.e., $\hat{V}^{(TB)}$), we find the overall Black PDF and the overall White CDF when the Black and White CDFs at each school probability Black are equal to the school marginal CDF at that school probability Black. This yields, for each racial group, the mixture distribution for the overall population that would result if the within-school distributions for each racial group matched the actual marginal (Black/White) distribution within school. We then find V as in B3, using these newly weighted overall distributions by race.

When mapping the Black CDF within school to the White CDF, we keep all White parameters unchanged. For Black students in schools where $p(b)_s \neq 1$, we assign them the parameter values of White students in the same school and find the marginal Black CDF that would result if Black students' within-school CDFs matched those of White students in the same school. We then use the new Black PDF to apply B3 and find $V_{btwn}^{(B\ to\ W)}$. Following a similar but reversed procedure to map the White distribution to the Black distribution within school, we find $V_{btwn}^{(W\ to\ B)}$.

We find the overall population gaps and gap proportions in the y^* metric by first finding SD_y, the overall population SD in the y metric, using the law of total variance (for each simulation scenario). Given that the gap in the y metric is -1, the overall gap in the y^* metric is $\frac{-1}{SD_y}$, and the decomposition proportions are equal in both y and y^* (for the population, we use VR=.42666667).

Appendix D. Parameters for Extreme Simulation Scenarios.

We also evaluate these methods under more extreme population parameters and sampling scenarios. As seen in the second panel of Table 1 in the main text, we simulate a "high heteroskedasticity" scenario to compare the relative performance of HOMOP, PHOP, and HETOP models under different levels of true heteroskedasticity. For this scenario, we assign the White distribution within schools a SD of .80 (4th column of Table 1 in main text) and the Black distribution within schools a SD of 1.20 (3rd column of Table 1 in main text). This represents a relatively large coefficient of variation (approximately 0.54, higher than the maximum of .3 used in simulations by Reardon et al. [2017] when evaluating HETOP models). Second, to examine how the methods perform when mono-racial schools are present, we include a "high segregation" scenario, in which schools are either 0%, 10%, 50%, 90%, or 100% Black (sampling 30 schools of each type), yielding *VR*=.652 (compare to a *VR* range of approximately .5 to .63 across different relevant analytic samples in various waves of the ECLS-K:2011). Third, we simulate a "small standardized gap" scenario in which we fix the overall population gap to approximately -.30 SD (similar to the smallest observed standardized Black-White gap across rounds and subjects in the ECLS-K:2011). We achieve the small standardized gap by increasing the within-school SDs while holding other parameters constant (see columns 4 and 5 of the bottom panel of Table 1 in main text). Finally, we examine a scenario that allows for sampling error in school proportion Black, both at the school level and within school. For these simulations, we randomly draw the true school proportion Black for each school, where each proportion (.10, .50, .90) has a 1/3 chance of being pulled for each sampled school. We then randomly sample students from each school, where each student's probability of being Black is given by the school's true

proportion Black. For all supplementary simulations, we use $\beta_1 = -.4$, 150 schools, 30 students per school, and 10 achievement bins to keep the number of simulation scenarios manageable.

Appendix E. Simulation Results for HETOP Models.

Main Simulations

In Table E1, we present results from the HETOP model. With 30 students per school and 4 bins, the HETOP models had convergence issues (with 12% to 26% of models converging). The models that did converge yielded high estimates of bias and large RMSEs across all parameters, often driven by a relatively small number of simulations that produced extreme estimates. With 10 bins and 30 students, a greater number of HETOP models converged (56%-78%), and the bias estimates were substantially smaller. In fact, bias estimates for the overall gap were smaller here than in the PHOP models; however, bias in the proportional decompositions were larger than the PHOP and HOMOP models.

With 300 students per school, the HETOP convergence issues were all but eliminated (with only one model across all simulations failing to converge), and bias was dramatically reduced. Across bin sizes, bias in the overall gap estimates was smaller for HETOP than PHOP (never more extreme than -.008 SD). Bias in the proportional decompositions were generally larger than for PHOP and HOMOP, but still small, with approximately +/- .002 as the most extreme estimates.

More Extreme Simulations

In Table E2, we present the simulation results for the HETOP models under the more extreme conditions. With high heteroscedasticity, HETOP again exhibited convergence problems (597/1000 converging) and performed worse than HOMOP and PHOP for the proportional decompositions. Bias for estimating the overall gap was small, (-.008 SD), but RMSE was large (.135). With high segregation, bias remained high when estimating decomposition proportions (849/1000 HETOP models converged). With sampling error added

for school proportion Black, the estimated bias for the decomposition proportions was small (.007 and -.007 for TB and TW, respectively), but only 381 out of 1,000 simulations converged. The proportional decompositions showed the largest bias under the small standardized gap condition.

Table E1.
Bias and RMSE from Simulated Estimates of HETOP Decomposition Elements.

			Overall Gap			Proportion Total Between			Proportion Total Within			
Num. Sts	Bins	β_1	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Converged
30	4	-0.4	0.530	0.316	<0.001	-2.667	9.559	0.049	4.651	16.673	0.049	223
30	4	-0.6	0.670	0.388	<0.001	-1.241	2.150	<0.001	2.165	3.750	<0.001	260
30	4	-0.8	0.715	0.340	<0.001	-0.893	0.584	<0.001	1.558	1.018	<0.001	183
30	4	-1	0.624	0.256	<0.001	-0.732	0.403	<0.001	1.277	0.703	<0.001	119
30	10	-0.4	-0.030	0.113	<0.001	-0.043	0.278	<0.001	0.075	0.486	<0.001	750
30	10	-0.6	-0.036	0.124	<0.001	-0.040	0.244	<0.001	0.070	0.425	<0.001	782
30	10	-0.8	-0.037	0.136	<0.001	-0.045	0.225	<0.001	0.078	0.392	<0.001	732
30	10	-1	-0.046	0.107	<0.001	-0.034	0.158	<0.001	0.060	0.275	<0.001	561
300	4	-0.4	-0.003	0.007	<0.001	0.001	0.006	<0.001	-0.002	0.010	<0.001	1000
300	4	-0.6	-0.004	0.008	<0.001	0.000	0.005	0.942	0.000	0.009	0.942	1000
300	4	-0.8	-0.004	0.009	<0.001	-0.001	0.005	<0.001	0.001	0.009	<0.001	999
300	4	-1	-0.005	0.009	<0.001	-0.001	0.005	<0.001	0.002	0.009	<0.001	1000
300	10	-0.4	-0.004	0.008	<0.001	0.001	0.006	<0.001	-0.001	0.010	<0.001	1000
300	10	-0.6	-0.005	0.008	<0.001	0.000	0.005	0.066	0.001	0.009	0.066	1000
300	10	-0.8	-0.006	0.009	<0.001	0.000	0.005	0.010	0.001	0.008	0.010	1000
300	10	-1	-0.006	0.009	<0.001	-0.001	0.005	<0.001	0.001	0.008	<0.001	1000

Note. Bias in overall gap is expressed in population SD units; bias in decomposition proportions are expressed in proportion units. Each scenario presents estimated bias and RMSE from 1,000 simulations. Converged = number of the 1,000 simulations for which models converged. Num. Sts. = number of students per school (in each of 150 schools, with equal representation from schools that are 10%, 50%, and 90% Black [vs. White]). Bins = number of bins that student-level data were coarsened to (bins of 10 are equally-sized; bins of 4 have cut scores at 20th, 50th, and 80th percentiles to represent a scenario in which only proficiency count data are available). β_1 = value of β_1 in data-generating model (see equations 7 and 8). p = p-value for test of null hypothesis that bias=0.

Table E2.
Bias and RMSE from Simulated Estimates of HETOP Decomposition Elements under Alternative Populations and Sampling Scenarios.

	Overall Gap			Between			Within			
	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Converged
	High Heteroscedasticity									
HETOP	-0.008	0.135	0.243	-0.098	0.361	<0.001	0.171	0.630	<0.001	597
	High Segregation									
HETOP	-0.025	0.131	<0.001	-0.053	0.330	<0.001	0.158	0.961	<0.001	849
	Sampling Error for Segregation and School Proportion Black									
HETOP	-0.044	0.062	<0.001	0.007	0.015	<0.001	<0.001	0.023	<0.001	381
	Small Standardized Gap									
HETOP	0.002	0.062	0.643	-0.130	0.340	<0.001	0.227	0.593	<0.001	524

Note. Bias in overall gap is expressed in population SD units; bias in decomposition proportions are expressed in proportion units. Each scenario presents estimated bias and RMSE from 1,000 simulations. Converged = number of the 1,000 simulations for which models converged. All simulation scenarios use 150 schools with 30 students per school, data coarsened to 10 equally-sized bins, and $\beta_1 = -.4$ in data-generating model (see equation 7 and 8). p = p-value for test of null hypothesis that bias=0.

Appendix F. Ordered Probit Simulations for Extreme Conditions with 300 Students per School.

Table F1.
Bias and RMSE from Simulated Estimates of HETOP, PHOP and HOMOP Decomposition Elements under Alternative Populations and Sampling Scenarios.

	Overall			Between			Within			
	Bias	RMSE	p	Bias	RMSE	p	Bias	RMSE	p	Converged
	High Heteroskedasticity									
HETOP	-0.004	0.008	<0.001	0.001	0.006	0.001	-0.001	0.011	0.001	1000
PHOP	-0.006	0.009	<0.001	0.000	0.006	0.022	-0.001	0.011	0.022	1000
HOMOP	-0.003	0.007	<0.001	0.006	0.009	0.000	-0.010	0.015	0.000	1000
	High Segregation									
HETOP	-0.005	0.008	<0.001	0.002	0.005	0.000	-0.001	0.015	0.013	1000
PHOP	-0.006	0.009	<0.001	0.002	0.005	0.000	-0.001	0.015	0.029	1000
HOMOP	-0.002	0.007	<0.001	0.002	0.005	0.000	-0.001	0.015	0.001	1000
	Sampling Error for Segregation and School Proportion Black									
HETOP	-0.004	0.009	<0.001	0.000	0.011	0.234	0.000	0.010	0.135	1000
PHOP	-0.005	0.010	<0.001	-0.001	0.011	0.069	0.000	0.010	0.798	1000
HOMOP	-0.001	0.008	<0.001	0.000	0.011	0.249	0.000	0.010	0.117	1000
	Small Standardized Gap									
HETOP	-0.002	0.009	<0.001	0.002	0.020	0.005	-0.003	0.035	0.005	1000
PHOP	-0.003	0.009	<0.001	0.001	0.020	0.018	-0.003	0.035	0.018	1000
HOMOP	-0.001	0.009	0.004	0.002	0.020	0.012	-0.003	0.035	0.012	1000

Note. Bias in overall gap is expressed in population SD units; bias in decomposition proportions are expressed in proportion units. Each scenario presents estimated bias and RMSE from 1,000 simulations. Converged = number of the 1,000 simulations for which models converged. All simulation scenarios use 150 schools with 300 students per school, data coarsened to 10 equally-sized bins, and $\beta_1 = -.4$ in data-generating model (see equation 7 and 8). p = p-value for test of null hypothesis that bias=0.

Appendix G. Georgia State Proficiency Data Applications.

In Table G1, we present school-level sample sizes for the Georgia proficiency count data. These data are reported as school-by-race-by-score-bin counts for each grade level and year. In other words, the state reports the number of Black and White students in each grade at each school who scored in each of the three possible proficiency categories (we use data from 2011-2014). As discussed in the main text, HETOP can encounter convergence issues when cell counts are sparse. As seen in Table G1, it is relatively rare for a school-by-race-by-grade level group to have non-zero counts in only one of the three score bins. However, it is more common for a school-by-race-by-grade level group to have non-zero counts in only two of the three score bins.

In Table G2, we present the results from the *V* decompositions applied to the Georgia state proficiency data (described in main text).

In Table G3, we present the results from the ordered probit decomposition models. As described in the main text, few of the HETOP models converged, due to subgroups with empty proficiency count cells. We therefore conducted supplementary analyses dropping offending school-by-race subgroups in order to achieve model convergence, and present these results in Table G4. We did this in two steps. First, we identified school-by-race subgroups with at least one empty proficiency count cell and pooled them together into race-by-school-percent-Black-tercile groups. If a given race-by-school-percent-Black bin still had at least one empty proficiency count cell, we dropped that set of observations from the data set. We then fit the ordered probit models to obtain the school-by-race estimates as described in the main text.

Table G1. Counts for schools and school-by-grade-race-by-bin cells from Georgia proficiency count data.

			Black students						White students					
Year	Grade	Total sch N	N Schs with at least 1 Black st.	N schs with 1 empty cell	N schs with 2 empty cells	N schs no Black students in lower third	N schs no Black students in middle third	N schs no Black students in top third	N Schs with at least 1 White st.	N schs with 1 empty cell	N schs with 2 empty cells	N schs no White students in lower third	N schs no White students in middle third	N schs no White students in top third
2011	3	1247	970	20	0	17	0	3	911	53	0	52	0	1
	4	1238	976	41	0	21	0	20	892	70	0	67	0	3
	5	1231	979	169	3	159	0	13	887	234	1	231	1	3
	6	550	479	42	1	4	1	38	443	16	0	6	0	10
	7	520	464	30	1	23	1	7	418	21	0	20	0	1
	8	524	464	43	1	19	1	24	424	24	1	19	0	6
2012	3	1244	962	31	0	24	1	6	906	61	0	59	1	1
	4	1235	957	42	0	22	0	20	887	51	0	50	0	1
	5	1225	972	133	1	113	0	21	873	202	0	197	0	5
	6	559	488	51	1	11	0	41	444	10	0	7	0	3
	7	532	475	40	1	32	1	8	426	35	0	34	0	1
	8	523	464	48	0	21	0	27	425	31	0	21	0	10
2013	3	1231	942	21	0	15	0	6	898	38	1	37	2	0
	4	1224	952	56	0	40	0	16	875	94	0	93	0	1
	5	1216	945	191	1	189	1	2	865	303	2	303	2	0
	6	553	483	30	0	9	0	21	445	16	0	14	0	2
	7	530	476	34	1	28	1	6	425	36	1	35	1	1
	8	526	466	43	0	28	0	15	426	32	0	28	0	4
2014	3	1224	953	36	2	33	2	3	894	45	0	44	1	0
	4	1217	954	53	2	32	0	23	874	56	0	55	0	1
	5	1211	948	183	1	175	0	9	858	284	2	284	2	0
	6	551	479	40	0	16	0	24	448	18	0	15	0	3
	7	532	474	33	1	22	1	11	425	30	0	30	0	0

8	530	471	34	0	24	0	10	428	29	0	25	1	3

Note. Data are reported by state as number of students by race in each school who scored within each of three proficiency score bins.

Table G2

V Decompositions for Black-White Math Gaps, Georgia State Testing Data.

Year	Grade	Total *V*	Prop. $\hat{V}^{(TB)}$	Prop. $\hat{V}_{btwn}^{(W\ to\ B)}$	Prop. $\hat{V}_{btwn}^{(B\ to\ W)}$	Black N	White N	Sch N	Mean $\overline{Black}_s^{(black)}$	Mean $\overline{Black}_s^{(white)}$
2011	3	-0.75	0.65	0.51	0.61	44224	53749	1247	0.753	0.204
	4	-0.73	0.64	0.48	0.61	45659	54310	1238	0.752	0.208
	5	-0.67	0.66	0.51	0.63	45604	54574	1231	0.751	0.208
	6	-0.73	0.57	0.38	0.48	45530	54891	550	0.709	0.241
	7	-0.65	0.60	0.39	0.54	45246	54304	520	0.705	0.246
	8	-0.63	0.60	0.44	0.53	44524	53266	524	0.703	0.248
2012	3	-0.79	0.65	0.51	0.61	43063	53182	1244	0.751	0.202
	4	-0.75	0.65	0.51	0.61	42984	52846	1235	0.753	0.201
	5	-0.65	0.67	0.52	0.65	44990	53405	1225	0.753	0.208
	6	-0.70	0.58	0.38	0.49	46386	54205	559	0.712	0.246
	7	-0.69	0.58	0.39	0.50	44916	54222	532	0.706	0.244
	8	-0.64	0.61	0.42	0.55	45451	54112	523	0.706	0.247
2013	3	-0.74	0.67	0.54	0.64	42939	53016	1231	0.753	0.200
	4	-0.80	0.65	0.52	0.61	42566	52237	1224	0.753	0.202
	5	-0.68	0.69	0.55	0.67	42810	52239	1216	0.754	0.202
	6	-0.76	0.58	0.39	0.49	45913	53188	553	0.713	0.248
	7	-0.75	0.58	0.38	0.49	46231	53977	530	0.710	0.248
	8	-0.69	0.61	0.44	0.55	45256	54145	526	0.708	0.244
2014	3	-0.69	0.69	0.57	0.65	44295	52382	1224	0.754	0.208
	4	-0.78	0.67	0.53	0.63	42837	52268	1217	0.752	0.203
	5	-0.73	0.69	0.56	0.67	42574	51728	1211	0.752	0.204
	6	-0.77	0.59	0.38	0.52	44765	52547	551	0.708	0.249
	7	-0.80	0.58	0.37	0.50	46684	53397	532	0.712	0.251
	8	-0.74	0.61	0.43	0.54	46715	53956	530	0.711	0.251

Note. Data were reported by the state as school-by-race counts in three proficiency categories. $\overline{Black}_s$ = school proportion Black.

Table G3.
Ordered Probit Decompositions for Black-White Math Gaps, Georgia State Testing Data.

		Overall			Proportion TB			Proportion TW			N			Mean Sch. Proportion Black	
Year	Grade	HOMOP	HETOP	PHOP	HOMOP	HETOP	PHOP	HOMOP	HETOP	PHOP	Black	White	Schools	Black Sts	White Sts
2011	3	-0.699		-0.714	0.673		0.670	0.726		0.732	44224	53749	1247	0.753	0.204
	4	-0.692		-0.701	0.657		0.656	0.753		0.754	45659	54310	1238	0.752	0.208
	5	-0.630		-0.644	0.682		0.679	0.695		0.701	45604	54574	1231	0.751	0.208
	6	-0.692		-0.696	0.586		0.586	0.777		0.778	45530	54891	550	0.709	0.241
	7	-0.617		-0.624	0.618		0.616	0.707		0.710	45246	54304	520	0.705	0.246
	8	-0.601		-0.607	0.617		0.616	0.703		0.704	44524	53266	524	0.703	0.248
2012	3	-0.733		-0.750	0.676		0.672	0.718		0.726	43063	53182	1244	0.751	0.202
	4	-0.703		-0.713	0.670		0.668	0.738		0.741	42984	52846	1235	0.753	0.201
	5	-0.622		-0.632	0.686		0.685	0.690		0.693	44990	53405	1225	0.753	0.208
	6	-0.660	-0.664	-0.664	0.599	0.593	0.598	0.752	0.762	0.753	46386	54205	559	0.712	0.246
	7	-0.637		-0.653	0.606		0.601	0.733		0.742	44916	54222	532	0.706	0.244
	8	-0.606		-0.611	0.625		0.624	0.693		0.694	45451	54112	523	0.706	0.247
2013	3	-0.685		-0.701	0.692		0.688	0.689		0.697	42939	53016	1231	0.753	0.200
	4	-0.739		-0.753	0.674		0.672	0.726		0.731	42566	52237	1224	0.753	0.202
	5	-0.636		-0.651	0.714		0.711	0.639		0.646	42810	52239	1216	0.754	0.202
	6	-0.709	-0.714	-0.715	0.603	0.599	0.602	0.742	0.750	0.745	45913	53188	553	0.713	0.248
	7	-0.697		-0.707	0.602		0.600	0.739		0.743	46231	53977	530	0.710	0.248
	8	-0.651		-0.657	0.637		0.637	0.676		0.677	45256	54145	526	0.708	0.244
2014	3	-0.650		-0.665	0.718		0.714	0.620		0.629	44295	52382	1224	0.754	0.208
	4	-0.720		-0.735	0.690		0.687	0.687		0.695	42837	52268	1217	0.752	0.203
	5	-0.679		-0.693	0.713		0.710	0.636		0.643	42574	51728	1211	0.752	0.204
	6	-0.714	-0.718	-0.722	0.611	0.603	0.609	0.719	0.734	0.723	44765	52547	551	0.708	0.249
	7	-0.738		-0.752	0.599		0.595	0.744		0.751	46684	53397	532	0.712	0.251
	8	-0.693		-0.701	0.632		0.630	0.681		0.686	46715	53956	530	0.711	0.251

Note. Data were reported by the state as school-by-race counts in three proficiency categories. Blank HETOP cells indicate that the HETOP model would not converge.

Table G4.
Ordered Probit Decompositions for Black-White Math Gaps, Georgia State Testing Data (altered sample to achieve HETOP convergence).

		Overall			Proportion TB			Proportion TW			N		
Year	grade	HOMOP	HETOP	PHOP	HOMOP	HETOP	PHOP	HOMOP	HETOP	PHOP	Black	White	Schools
2011	3	-0.688	-0.682	-0.703	0.676	0.668	0.673	0.705	0.725	0.710	43798	51905	1234
	4	-0.666	-0.664	-0.675	0.652	0.646	0.652	0.748	0.763	0.749	44867	51178	1212
	5	-0.567	-0.585	-0.581	0.660	0.657	0.658	0.664	0.675	0.667	42040	42839	1095
	6	-0.683	-0.684	-0.687	0.581	0.578	0.581	0.788	0.796	0.789	44172	54482	534
	7	-0.607	-0.610	-0.613	0.609	0.600	0.607	0.711	0.726	0.713	44420	52613	506
	8	-0.580	-0.582	-0.585	0.598	0.593	0.597	0.718	0.728	0.719	43267	51533	502
2012	3	-0.711	-0.709	-0.727	0.675	0.662	0.672	0.709	0.732	0.715	42500	50540	1224
	4	-0.685	-0.675	-0.694	0.679	0.666	0.678	0.721	0.752	0.722	42202	50855	1220
	5	-0.557	-0.574	-0.565	0.675	0.667	0.674	0.666	0.680	0.667	41919	43544	1115
	6	-0.652	-0.652	-0.656	0.604	0.600	0.604	0.743	0.753	0.745	44650	53812	543
	7	-0.625	-0.637	-0.639	0.597	0.584	0.593	0.747	0.774	0.752	43731	51357	515
	8	-0.594	-0.593	-0.598	0.614	0.609	0.613	0.698	0.709	0.699	43949	52308	496
2013	3	-0.679	-0.671	-0.695	0.696	0.685	0.692	0.671	0.692	0.679	42572	51789	1217
	4	-0.718	-0.716	-0.732	0.678	0.670	0.675	0.702	0.725	0.706	41367	48495	1186
	5	-0.578	-0.600	-0.601	0.672	0.664	0.668	0.569	0.587	0.574	38112	36268	1017
	6	-0.702	-0.707	-0.709	0.610	0.605	0.609	0.725	0.735	0.728	45088	52543	540
	7	-0.698	-0.705	-0.707	0.605	0.598	0.603	0.722	0.733	0.725	44875	51782	514
	8	-0.645	-0.647	-0.650	0.625	0.619	0.624	0.667	0.680	0.668	43873	51675	506
2014	3	-0.642	-0.633	-0.657	0.703	0.699	0.700	0.623	0.627	0.630	43647	50442	1204
	4	-0.705	-0.702	-0.719	0.689	0.681	0.686	0.676	0.689	0.682	41791	50097	1190
	5	-0.603	-0.605	-0.622	0.661	0.641	0.657	0.572	0.624	0.578	38438	35865	1033
	6	-0.697	-0.699	-0.704	0.611	0.603	0.608	0.719	0.734	0.722	43394	51142	535
	7	-0.736	-0.746	-0.749	0.603	0.597	0.600	0.735	0.746	0.741	45732	52125	515
	8	-0.689	-0.689	-0.697	0.629	0.620	0.627	0.672	0.689	0.676	45474	52387	508

Note. Data were reported by the state as school-by-race counts in three proficiency categories. To achieve HETOP convergence, school-by-race subgroups with at least one empty proficiency count cell were combined into race-by-terciles of school percent black. Race-by-terciles of school percent black with at least one empty proficiency count cell were then dropped (same samples across all estimation approaches).

References (Appendices)

Bond, T. N., & Lang, K. (2013). The evolution of the Black-White test score gap in Grades K–3: The fragility of results. *Review of Economics and Statistics*, *95*(5), 1468-1479.

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