Addressing Practical and Technical Challenges when Measuring Achievement Gaps: Implications for Practice and Policy

Chair: Yeow Meng Thum, Northwest Evaluation Association (NWEA)
Discussant: Eric Hedberg, NORC at the University of Chicago
Section: Research Methods

Symposium Summary

Achievement gaps between student subgroups (including black-white and male-female gaps) have been documented for decades, in part because they can have real consequences for students, teachers, and school systems. For example, achievement gaps in mathematics between male and female students, as well as between black and white students, are posited as one source of differential participation in quantitative college majors and careers (Corbett & Hill, 2012). This differential participation likely contributes to wage disparities in the U.S. favoring white workers relative to black workers, and male workers relative to female workers (Corbett & Hill, 2012; Neal, 2006). As a result, achievement gaps are often a primary motivator for education policies that target educational equity in the United States.

A significant body of new research attempts to improve the usefulness of gaps research to practitioners by showing that gap estimates are scale-dependent, and proposing scale-invariant approaches to estimation (Ho, 2009; Ho & Reardon, 2012). Despite advances in measurement of achievement gaps, important practical and technical challenges remain, especially related to weighting, longitudinal modeling, and conditioning on covariates. Not addressing these challenges likely means that some gaps research will fail to represent differential student achievement accurately, estimate meaningful gaps for certain subgroups, and provide insight into how to close the gaps. These failings, in turn, make it much harder for teachers, administrators, policymakers, and other stakeholders to address differential achievement across subgroups.

Our symposium will lay out options for addressing these challenges associated with measuring and estimating achievement gaps. Two of these papers consider options for weighting incomplete student achievement data to more accurately estimate gaps that are reflective of state and national trends in gaps. The other two examine issues related to modeling gaps when the data are nested and longitudinal, including how to condition on covariates. The question of conditioning arises in part because of the new availability of metadata from computer adaptive tests on how engaged students are with the assessment. Previous research shows that conditioning achievement gaps on test engagement shrinks the black-white gap in reading by nearly 20 percent (Soland, 2016).

While the papers included address two broad themes, they are all related and build on each other. The first paper (Fahle & Gagne) shows that there are options for estimating nationally representative gaps using new methods and available federal data. Thum and Shear (paper two) extends the work of Fahle and Gagne to consider the implications of aggregating data when longitudinal data are available, including how best to model longitudinal gaps. Thum, Bo, and Soland presents an alternative strategy for weighting that uses some of the methods presented in the second paper. And, finally, Soland and Thum considers how to address low test engagement by estimating nested, longitudinal models that include engagement as a covariate.

Below, we provide detail on each of the four papers included in this panel. Together, they show that many common assumptions researchers tend to make when measuring and studying achievement
gaps may not lead to the most precise or accurate estimates, which has implications for practice, policy, and research.

**Study 1: Fahle and Gagne**

**Title:** Understanding Changes in Racial Achievement Gaps during Elementary and Middle School  
**Authors:** Erin Fahle & Josh Gagne (presenting author)

**Background / Context**

Racial/ethnic achievement gaps have been the focus of significant research to date; however, there is little available information on how these gaps change during elementary and middle school, and no information on whether those trends vary across local contexts. Prior studies using nationally representative data find that average white-black achievement gaps in math and reading are already large at the beginning of school, grow during early elementary school and stabilize by the end of middle school (Quinn, 2015; Fryer & Levitt, 2004; Reardon, 2008; Reardon & Galindo, 2006; Murnane et al., 2006). A limited amount of work has further investigated differential changes in white-black gaps across the achievement distribution (among high or low achieving students), but finds inconsistent evidence of change perhaps due to the scale-sensitivity of distributional gap measures (Sohn, 2012; Neal, 2006; Clotfelter et al., 2009). There has been comparatively little research on the patterns of change of average white-Hispanic achievement gaps in math and ELA. Research finds that gaps are also large upon entry into kindergarten but, in contrast to the white-black gaps, they shrink dramatically through first grade before stabilizing through eighth grade (Fryer & Levitt, 2004, 2006; Reardon & Galindo, 2006).

In both cases, most prior work on how racial/ethnic achievement gaps change over grades leverages national data (e.g., ECLS-K and NAEP-LTT) that limits their ability to examine variation in the grade trend within the U.S., as well as limits their ability to study grade-to-grade fluctuations in the gaps after the early elementary school years (see Fryer & Levitt, 2004, 2006; Reardon, 2008; Reardon & Galindo, 2006; Ferguson, 1998; Neal, 2006; Phillips, Crouse, & Ralph, 1998). Characterizing gap trends at a higher resolution and investigating the variation in those trends is a clear next step to understanding how local context affects racial/ethnic achievement gaps, and to identifying the types of districts where minority students are performing particularly well (or poorly) and achievement gaps within cohorts are narrowing (or widening).

**Purpose / Objective / Research Question**

In this paper, we characterize the annual changes in white-black and white-Hispanic achievement gaps as students progress from third through eighth grade, examining trends in gaps across the achievement distribution and variation in those trends due to local context. Measuring gap trends among high and low scoring students in addition to traditional mean gaps enables us to understand whether the limited changes reported in the literature are due to offsetting changes in the tails of the distribution or general stagnation across the distribution. Investigating the variation in grade trends across districts further provides insights into whether local variation in trends exists and whether it can be explained by contextual factors.
Population of Study

We study US school districts serving third through eighth grade students. The achievement data is drawn from the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) assessment. We use a subset of the data that includes math and ELA test scores for third through eighth grade over the 2008-09 through 2012-13 school years with approximately 33 million test records for 8 million students, 18,000 schools, and 4,000 districts from across the US. We also use data from the Stanford Education Data Archive (SEDA) to develop nonparametric covariate-balancing weights (Hainmueller, 2011) such that the NWEA sample can be weighted to be nationally representative.

Analysis

We estimate three district-level gap measures for both racial/ethnic comparisons in each subject, grade and year in our data. First, we estimate a measure of the mean achievement gap using the $V$-statistic, which is interpretable as the gap in common standard deviation units between two groups (Ho, 2009; Ho & Haertel, 2006, Ho & Reardon, 2012; Reardon & Ho, 2015). Second, we estimate the upper and lower tail “gaps” using the Proportion-Adjusted Relative Difference ($\lambda_{\theta}$), a percentile-specific proportion that provides information about the representation of whites compared with blacks/Hispanics at different percentiles, $\theta$, of the achievement distribution (Robinson & Lubienski, 2011). These “metric-free” statistics avoid assumptions about vertical-scaling and comparable test metrics inherent to comparisons across grades, years, and test scales.

For each of the three measures, we model changes in the white-black and white-Hispanic gaps over grades using precision-weighted Hierarchical Linear Modeling. This model enables us to not only investigate average grade trends, but also to test for variation in those trends across districts. Additionally, these models enable the inclusion of district covariates to explain both the variation in the magnitude of the gaps across districts, as well as to explain any variation in the gap trends across districts.

Preliminary Findings

Our preliminary findings show, first, that, on average, our estimated district level gaps align in magnitude with those from prior literature. The white-black gaps in our sample average 0.55 SD in reading and 0.67 SD in math. The white-Hispanic gaps are, on average, .57 SD in reading and .55 SD in math. Using a fully non-parametric grade specification, we further find that in our nationally weighted sample there is little absolute change in the gaps on average from third to eighth grade for both racial groups. However, there are some significant changes during elementary school, and the direction of these changes is different for each racial group, as was highlighted in prior research. Specifically, we find that the white-black gaps increase during this time, whereas the white-Hispanic gaps shrink slightly (see Figure 1). The next steps of our work will explore the variation in these trends across districts.

Conclusion

This paper adds to the current literature through systematically analyzing and characterizing changes in white-black and white-Hispanic achievement gaps as students move through elementary and middle school at a high resolution. Relative to prior literature, our study offers two key advantages:
(1) the use of district-level longitudinal data with multiple cohorts across six grades allows for investigation of variation in gap trends across districts while providing robustness to the observed results; and (2) the use of metric-free measures ensures comparability of the gap-sizes across grades.

**Study 2: Thum and Shear**

**Title:** Sensitivity of achievement gap trend estimates to alternative treatments of nested data

**Authors:** Yeow Meng Thum & Benjamin R. Shear (presenting author)

**Background**

Gathering relevant data and analyzing them appropriately is key to understanding test score achievement gaps. The vast majority of achievement gap analyses in education compare student subgroups by their aggregate statistics, such as differences in subgroup means or subgroup percent proficient, collected from classrooms, schools, and districts (e.g., Bloom, Hill, Black, & Lipsey, 2008, Schulte & Stevens, 2014). This is driven in part by the wide availability of these aggregate data.

The multilevel modeling literature, e.g., Raudenbush and Bryk (2002), shows that analyzing aggregated data ignores clustering effects and can lead to a loss of statistical power, inflated Type I error rates, and biased parameter estimates. Explicitly modeling the multi-level data structure can mitigate such problems and support additional research questions about whether and how much key quantities such as the school achievement gap, varies between schools. As longitudinal student data become more accessible (e.g., the ECLS-K), it is important to evaluate how these issues impact conclusions about achievement gaps (Shear & Thum, 2016).

**Purpose**

This paper compares inferences about gender achievement gap trends based on an analysis of aggregate-level data to those based on multi-level growth models. The analyses are framed by expectations set by standard analytic results showing the differences between estimators of achievement gaps due to alternative levels of aggregation (e.g., Raudenbush & Bryk, 2001, pp. 114-115). Using longitudinal test scores for an age-cohort of students from a large number of schools, we focus on tests of the “equalization” hypothesis, which evaluates whether the mean difference in test scores between male and female students narrows from earlier to later grades.

**Design and Data**

We use longitudinal test score data from a computer-adaptive mathematics test administered in the fall and spring of 2nd through 5th grade for a single age-cohort (entering 2nd grade in 2007) attending public schools in a mid-Atlantic state. The sample includes 296,209 test scores for 66,585 students across 402 schools. Each school has observations for between 2 and 8 time points, while students have scores for between 1 and 8 time points. Table 1 shows the sample descriptive statistics, by gender. The third row of Table 1, labeled “Time,” shows how the time points were coded. A 1-unit change in time corresponds to a full calendar year.

**Models and Analyses**

Standard analytic results describing the impact of aggregation for nested longitudinal data are presented. We complement this with empirical results by fitting a sequence of three models that modify the level of aggregation and error structure of the data. First, we calculate the mean
difference in scores for male and female students at each time point, $D_t = m_{t,female} - m_{t,male}$, and model the trend in this aggregate gap over time using a linear regression model. There are eight outcomes (one per time point), and each is weighted by the sample size at time $t$:

$$D_t = \beta_0 + \beta_1 time + \epsilon_t.$$  

(1)

We then fit Model 2, which estimates gender-specific linear trends in the combined individual-level test scores, $Y_{it}$, for student $i$ and time $t$ as

$$Y_{it} = \beta_0 + \beta_1(time_{it}) + \beta_2(female_{i}) + \beta_3(time_{it} \times female_{i}) + \epsilon_{it}.$$  

(2)

Where $female_{i}$ is an indicator equal to 1 if a student is female and 0 otherwise. Finally, we fit a multi-level longitudinal growth curve model with random student effects and fixed gender effects, again using individual-level student scores as outcomes in the model:

$$Y_{it} = (\delta_{00} + \delta_{01} \times female_{i}) + (\delta_{10} + \delta_{11} \times female_{i}) + \epsilon_{it} \times time_{it} + \epsilon_{it} \times time_{i} + \epsilon_{it}$$

$$[r_{0i}, r_{1i}] \sim N([0,0], [\tau_{11}, \tau_{12}, \tau_{22}]).$$  

(3)

Where the $r_{0i}$ and $r_{1i}$ are student-level random effects. Precision weighting, by the standard error of measurement of scores, are applied for all multilevel models.

We also extend each model to include random school effects, across the $j = 1, ..., J$ schools. In Model 1, we estimate the mean difference separately in each school at each time point and fit a 2-level random effects model with the $D_{jt}$ school-specific gap estimates as outcomes. For Models 2 and 3, we add an additional level to each model. We refer to these models with school random effects as Models 1A, 2A and 3A, respectively. These models provide insights as to how school gender gaps and equalization vary among schools.

Using the results from the six models, we estimate three quantities of interest: (1) the magnitude of the initial gap in fall of 2nd grade, (2) the final gap in the spring of 5th grade, and (3) the change in these gaps. We also test whether each of these quantities is significantly different from 0. “Equalization” has occurred if the change in the gap is statistically significant and the absolute value of the estimated gap in 5th grade is smaller than the absolute value of the gap in 2nd grade. We focus on whether conclusions about the magnitudes of these quantities and tests of the equalization hypotheses are consistent across the six models.

**Preliminary Results**

Table 2 presents the estimated initial and final gaps, as well as change in gaps and the p-values for the null hypothesis that the true value of each gap (or the change in gaps) is equal to 0. All models indicate a statistically significant ($p < 0.05$) reduction in the gap from fall of 2nd grade to spring of 5th grade, although the magnitude of the change varied across models. The initial gap was statistically significant for all models and levels of aggregation. The point estimates and statistical significance of the gaps varied substantially across the six models. Estimates for Model 3A, an initial gap of 0.941 (0.099), a final gap of 0.536 (0.136), and a difference of -0.405 (0.121), appear to reflect best the corresponding information in Table 1.

**Conclusion**

Employing student longitudinal scale scores from a large number of schools, our results show that inferences about achievement gaps and changes in achievement gaps can indeed differ when the nested nature of assessment data are more or less appropriately reflected in the analyses, as predicted
by commonly known analytic results. Thus, when longitudinal student data are employed for achievement gap estimation, it is prudent to also explore findings from multilevel models that take into consideration how repeated scores are nested within students and how students are nested within schools.

Study 3: Thum and Soland

Title: School Norms for Mathematics Achievement Status, Term-to-term Growth, and Gender Gap

Authors: Yeow Meng Thum (presenting author) and Jim Soland

Background and Purpose

Since the Coleman Report over 50 years ago, closing gaps in achievement, particularly those for ethnicity and gender, had remained an important goal of the No Child Left Behind Act (NCLB) of 2001 and the Every Student Succeeds Act (ESSA) of 2015. Turning to the problem of gender differences in this study, we first note that NAEP 4th grade mathematics achievement for females has persistently lagged behind their male classmates in US public schools, with 42% of males and 38% of females performing at or above proficient in 2015 (US DOE, 2015). Although this finding appears to be widespread (e.g., Robinson & Lubiensky, 2011; Penner & Paret, 2008), we consider new results from a very large assessment database.

In this study, we provide a profile of new estimates of the gender gap amongst 4th grade public school students in mathematics, based on student score trends from grades 3 through 5. The data and procedure provide nationally representative norms for achievement status and term-to-term growth for male and female students and their differences in both achievement status and growth. We also provide estimates of the school-level variation of each of these statistics and their relative importance in the form of school intra-class correlations to be compared, for example, to the efforts by Bloom, Hill, Black, and Lipsey (2008) and Hedges and Hedberg (2013). In combination, the results supply researchers and policy makers with the needed empirical benchmarks for assessing about how males and females in the nation’s 4th grade classrooms differ in mathematics.

Data and Design

This study employs student mathematics RIT scores for the NWEA Measures of Academic Progress® (MAP®) assessment, a computerized adaptive test that is used by over 6,000 districts across the US. All the available fall, winter, and spring MAP mathematics RIT scores (592,305 in total) for a random sample of 1443 country-wide partner schools serving 130,077 students who attended grades 3, 4, and 5. Also using the 4th graders’ 3rd and 5th grade MAP results in the analysis materially strengthens the description of how the 4th graders in who took MAP performed and grew.

Because the NWEA MAP Growth Research Database, or GRD, is an archive for a very large number of partners, random sampling of schools reduces the anticipated computational burden of the analyses. To achieve meaningful generalizability, post-stratification weights, derived from comparing the NCES CCD-based distributional profile of partner schools with that of public school in the US, are introduced in the analysis to approximate nationally representative results for the 4th grade student population attending public schools in the US.
Model and Analysis

We denote the score received by student $i$ in school $j$ at test term $t=1,2,...,n_j$ by $Y_{ij}$. To estimate gender gap norms 4th graders, we describe the score trends of students and schools with the generic three-level hierarchical linear model

\[
\begin{align*}
\text{Within Student:} & \quad Y_{ij} = a'_{ij} \tau_j + e_{ij} \\
\text{Between Students:} & \quad \tau_j = X_{ij} \beta_j + r_j, \quad (1) \\
\text{Between Schools:} & \quad \beta_j = W_j \gamma + u_j
\end{align*}
\]

with regression coefficients $[\tau_j, \beta_j, \gamma]$ that correspond to polynomial growth design elements, $a'_{ij}$, student covariates $X_{ij}$, and school-level covariates $W_j$ employed. Of specific interest in this application is the dummy variable for males, such that $X_{ij} = \text{blockdiag}([1 \ Male_j])$, with an entry for each student growth component, $\tau_j$.

Using the selection matrices $H_\pi$ and $H_\beta$ to identify the coefficients among $\tau_j$ and $\beta_j$ (respectively) which are random, leads to the mixed-effects formulation of the model for student $i$ in school $j$

\[
Y_{ij} \sim MVN(A_{ij}x_{ij}W_i\gamma, \Sigma),
\]

where the components of variance for each level are

\[
e_{ij} \sim N(0, \sigma_{ij}^2), \quad r_{ij} \sim MVN(0, T_\pi), \quad u_{ij} \sim MVN(0, T_\beta),
\]

and

\[
\Sigma_{ij} = A_{ij}X_{ij}(H_\beta' T_\beta H_\beta)X_{ij}'A_{ij}' + A_{ij}(H_\pi' T_\pi H_\pi)A_{ij}' + \text{diag}(\sigma_{ij}^2).
\]

For any given testing term and subject characteristic, norms analysis merely performs conditional predictions based on the estimates of fixed-effects $[\hat{\gamma}, \text{Var}(\hat{\gamma})]$ and variance-covariance components $\hat{\sigma}^2, H_\pi' T_\pi H_\pi, H_\beta' T_\beta H_\beta$ for Model 2. For further details, see Thum and Hauser (2015).

Preliminary Results

Table 1 provides the predicted MAP mathematics mean and standard errors of male and female students who were 4th graders in the 2012-2013 school year. Mathematics performance for males and females in this population across the terms, though statistically significantly different, are practically indistinguishable. This is a very surprising finding, which is also at odds with much prior evidence. For one example, NAEP assessment of 2015 found that 4th grade males generally scored higher than females. Some tentative explanations for this finding, including the impact of smoothing by a growth curve analysis “smoothed” area being investigated.

Table 2 provides the predicted standard deviations by grade and term of school-level mean achievement for males and females in the population. ICCs for school-level means for males are reminiscent of the results of Hedges and Hedberg (2007). It appears however that the school means for males varied more than for females even as total score variation is quite comparable in magnitude. The result is higher school ICCs for males than for females. Table 2 also offers school ICCs for the gender gap. They are quite small, suggesting that gender gaps to be found in schools
may be quite small and they do not vary enough, assuming these result hold, to warrant research and policy attention.

Table 3 provides estimates to gender differences in term-to-term growth. The available results suggest that the spring grade 3 to fall grade 4 summer drop is larger for males than for females. This gender disparity seems to be greatly diminished when we examine summer drop from grade 4 to grade 5. The results also suggest that males and female 4th graders grew the same amounts from fall to spring during the school year, and also from the fall of grade 4 to the fall of grade 5.

Conclusions

This study joins recent efforts to lay an empirical foundation for interpreting and understanding the gender gap in achievement. We offered national MAP mathematics achievement status and growth norms by gender for both students and schools, extending the previous efforts by Thum and Hauser (2015). We have also added school norms for the gender gap, perhaps for the first time in the study of gender differences in mathematics. Much of the information developed, though new, are population marginal results. We are eager to further exploit NWEA GRD database and extend our procedures to furnish conditional norms, e.g. gender gap norms by school poverty deciles, to augment the tools we have developed thus far.

Tables

Table 1. MAP Mathematics: Predicted RIT mean and standard error estimates for US public school 4th graders in 2011-2012 school year by gender

<table>
<thead>
<tr>
<th>Grade</th>
<th>Term</th>
<th>Male</th>
<th>Est.</th>
<th>s.e.</th>
<th>Female</th>
<th>Est.</th>
<th>s.e.</th>
<th>Male-Female</th>
<th>Est.</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Winter</td>
<td>198.42</td>
<td>0.18</td>
<td>197.90</td>
<td>0.16</td>
<td>0.52</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Spring</td>
<td>203.82</td>
<td>0.19</td>
<td>203.25</td>
<td>0.17</td>
<td>0.58</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Fall</td>
<td>202.01</td>
<td>0.18</td>
<td>201.88</td>
<td>0.16</td>
<td>0.13</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Winter</td>
<td>208.92</td>
<td>0.19</td>
<td>208.50</td>
<td>0.17</td>
<td>0.42</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Spring</td>
<td>213.65</td>
<td>0.21</td>
<td>213.32</td>
<td>0.19</td>
<td>0.33</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Fall</td>
<td>211.94</td>
<td>0.21</td>
<td>211.74</td>
<td>0.18</td>
<td>0.20</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. MAP Mathematics: Predicted RIT standard deviation estimates and school intra-class correlations (ICC) for US public school 4th graders in 2011-2012 school year by gender

<table>
<thead>
<tr>
<th>Grade</th>
<th>Term</th>
<th>School</th>
<th>Total</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Winter</td>
<td>6.43</td>
<td>6.00</td>
<td>1.30</td>
</tr>
<tr>
<td>3</td>
<td>Spring</td>
<td>6.90</td>
<td>6.07</td>
<td>1.30</td>
</tr>
<tr>
<td>4</td>
<td>Fall</td>
<td>6.55</td>
<td>5.68</td>
<td>1.31</td>
</tr>
<tr>
<td>4</td>
<td>Winter</td>
<td>6.99</td>
<td>6.13</td>
<td>1.31</td>
</tr>
<tr>
<td>4</td>
<td>Spring</td>
<td>7.45</td>
<td>6.62</td>
<td>1.31</td>
</tr>
<tr>
<td>5</td>
<td>Fall</td>
<td>7.34</td>
<td>6.46</td>
<td>1.44</td>
</tr>
</tbody>
</table>
Table 3. MAP Mathematics: Predicted term-to-term RIT growth mean and standard error estimates for US public school 4th graders in 2011-2012 school year by gender

<table>
<thead>
<tr>
<th>Terms</th>
<th>Male</th>
<th>Female</th>
<th>Male-Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring Grade 3 to Fall Grade 4</td>
<td>-1.82</td>
<td>-1.37</td>
<td>-0.45</td>
</tr>
<tr>
<td>Spring Grade 4 to Fall Grade 5</td>
<td>-1.71</td>
<td>-1.59</td>
<td>-0.13</td>
</tr>
<tr>
<td>Fall Grade 4 to Spring Grade 4</td>
<td>11.65</td>
<td>11.44</td>
<td>0.20</td>
</tr>
<tr>
<td>Fall Grade 4 to Fall Grade 5</td>
<td>9.93</td>
<td>9.86</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Study 4: Soland and Thum

**Title**: Estimating conditional nonparametric gaps: an example involving student test-taking engagement

**Author**: James Soland (presenter), Yeow Meng Thum

**Background and Purpose**

Oftentimes, researchers wish to estimate achievement gaps that are conditional on a variable other than group membership. For example, studies have presented achievement gaps between male and female students by percentile of the achievement distribution. While conditioning on additional covariates is straightforward in a traditional regression framework, the problem becomes more complex when using nonparametric or so-called “metric-free” gap estimates (Ho, 2009). These estimators, which rely only on the ordinal information from the test score distribution, are motivated by considerable research showing that achievement gaps are often dependent on the scale used (Ho, 2009; Ho & Reardon, 2012).

This paper builds on work by Reardon, Shear, Castellano, and Ho (in press) to estimate nonparametric gaps using heteroskedastic ordered probit (HETOP) models. By coarsening available test score data and fitting a HETOP model, achievement gaps can be conditioned on additional covariates of interest. Specific approaches to implementing this method are explored by estimating achievement gaps that are conditional on student test-taking engagement. Research shows that students oftentimes respond to a test question before fully understanding its content, and that this “rapid-guessing behavior” is correlated with subgroup status (Wise, 2015).

**Data**

This study uses reading and mathematics scores from the Measures of Academic Progress (MAP), a vertically-scaled assessment given in most states across the country. Rather than use all available MAP data in an attempt to estimate nationally representative gaps, estimates are generated by state to explore whether there is heterogeneity in both gaps and engagement. Heterogeneity might occur for a variety of reasons. First, there could be regional differences, which is why we look at five states from four regions: Colorado (West), Michigan (Midwest), Minnesota (Midwest), New York (Northeast), and South Carolina (Southeast). These specific states were chosen because they have the highest proportion of students using MAP within that region. To help determine whether results generalize to the states in the sample, gaps were re-estimated using a raking procedure that uses weights to match sample marginal distributions to known population marginals. Specific to this
study, we use known, state-specific marginal distributions of race and MAP scores to adjust the estimates so that the sample marginals match those known, state-specific marginals with a certain degree of statistical certainty. Gaps were also re-estimated using school-level information on poverty and other factors from the Common Core of Data to create within-state school-level weights documented in Thum (2015). However, neither of the methods changed results substantively, therefore only the unweighted, sample-based results are reported.

Methods

To estimate a HETOP model, RIT scores are first coarsened into a set of \( K \) ordered categories defined by \( K - 1 \) thresholds. This variable, called \( RIT' \), takes on values \( k = \{1,2,...,K\} \). Then an ordered probit model is used

\[
pr(RIT'_{its} = k \mid b_i) = \Phi(\mu_k - X_{its}b) - \Phi(\mu_{k-1} - X_{its}b) \quad (1)
\]

where \( \Phi \) is the cumulative density function of the standard normal distribution. Given this model, the gap can be estimated using a latent variable interpretation

\[
RIT^*_{its} = \beta_0 + \beta_1 b_i + \epsilon_{itsc} (\exp(b_i \gamma_0)) \quad (2)
\]

where \( RIT^*_{its} \) is a z-score expressed in standard deviation units. Under this model, the variances for black and white students are allowed to differ. To identify the model, we set \( \beta_0 = 0 \) and \( \sigma^2(e_{itsc}) = 1 \) for the reference group (white students). For black students, \( b_i = 1 \) and \( \gamma_0 \) is therefore a term that allows the error variance to differ relative to white students. From there, the \( V \) gap can be estimated as

\[
\frac{\beta_1}{\sqrt{\exp(2\gamma_0)+1}} \quad (3).
\]

Equation 7 is equivalent to the more generic formula for \( V \) in Ho and Haertel (2006) and Ho (2009)

\[
V = \frac{\bar{x}_b - \bar{x}_w}{\sigma_{bw}} \quad (4)
\]

where \( \bar{x}_b \) is the mean for black students, \( \bar{x}_w \) is the mean for white students, and \( \sigma_{bw} \) equals the unweighted pooled standard deviation or

\[
\sigma_{bw} = \sqrt{\frac{\sigma_{b}^2 + \sigma_{w}^2}{2}} \quad (5).
\]

To produce conditional estimates of \( V \), Equation 6 can be re-estimated with additional covariates for our RTE thresholds:

\[
RIT^*_{its} = \beta_0 + \beta_1 b_i + \beta_2 RTE2 + \beta_3 RTE3 + \epsilon_{itsc} (\exp(b_i \gamma_0)) \quad (6)
\]

and a \( V \) statistic conditional on RTE bin can be estimated using the formula in equation 3.

Preliminary Results

Preliminary results using the HETOP approach show that achievement gaps shift considerably when conditioned on test engagement. As shown in the below figure, black-white gaps in reading can shrink by 20 percent when conditioned on test engagement.
Figure 1. Metric-free gaps estimates in reading by time period and test engagement (RTE) bin

Note: Student is in RTE1 if RTE = 1, RTE2 if RTE<1 and RTE ≥ .9, RTE3 if RTE<.
Citations


