COLLABORATIVE FOR
STUDENT GROWTH

WORKING PAPER

Using Assessment Metadata to Quantify the Impact of Test Disengagement on Estimates of Educational Effectiveness

By Megan Kuhfeld and Jim Soland

Kuhfeld, M. & Soland, J. (2019). *Using Assessment Metadata to Quantify the Impact of Test Disengagement on Estimates of Educational Effectiveness*. (The Collaborative for Student Growth at NWEA Working Paper).

nwea RESEARCH

Using Assessment Metadata to Quantify the Impact of Test Disengagement

on Estimates of Educational Effectiveness

Abstract

Educational stakeholders have long known that students might not be fully engaged when taking an achievement test, and that such disengagement could undermine the inferences drawn from observed scores. Thanks to the growing prevalence of computer-based tests and the new forms of metadata they produce, researchers have developed and validated procedures for using item response times—the seconds that elapse between when an item is presented and answered—to identify responses to items that are likely disengaged. In this study, we introduce those disengagement metrics for a policy and evaluation audience, including how disengagement might bias estimates of educational effectiveness. Analytically, we use data from a state administering a computer-based test to examine the effect of test disengagement on estimates of school contributions to student growth, achievement gaps, and summer learning loss. In so doing, we broaden the literature investigating how test disengagement can influence uses of aggregated test scores, provide guidance for policymakers and evaluators on how to account for disengagement in their own work, and consider the promise and limitations of using achievement test metadata for related purposes.

Educational stakeholders have long known that observed test scores can reflect more than achievement in subjects like math or reading alone. Observed scores can also be a function of factors irrelevant to the construct. For example, students must engage with the test sufficiently to demonstrate their content knowledge. In the presence of disengaged responses to items, observed achievement test scores may understate what students know and can do in that subject (Wise, 2015). While educators and policymakers may hypothesize that students are not fully engaged on tests, especially those with no stakes attached for the students, testing that hypothesis is not straightforward. If one assumes that such disengagement is widespread, then estimates of teacher, school, and program effectiveness may be influenced by the downward bias in observed achievement.

With the increasing prevalence of computer-based tests (CBTs), methods for detecting disengaged responding have expanded (Eklöf, Pavešič, & Grønmo, 2014; Guo et al., 2016; Rios, Liu, & Bridgeman, 2014; Wise, 2015). These methods often rely on student metadata captured when a CBT is administered. Specifically, most detection methods use response time, or the seconds that elapse between when an item is presented and answered. Research suggests that, in some cases, students respond to an item so quickly, its content could not have been understood (Eklöf et al., 2014; Guo et al., 2016; Rios et al., 2014; Wise, 2015). This behavior is often termed "rapid guessing" because responses are, when effectively detected, correct at a rate no better than chance (Wise & Kong, 2005a). Further, once a response is deemed a rapid guess, the student's test engagement can be summarized using the proportion of item responses on the test that were not rapid (Wise & Kong, 2005a). Wise (2015) catalogued nearly a decade's worth of research making a validity argument for the use of such metrics to identify disengaged item responses.

Thanks to these evolving metrics, the scope of the problem that disengaged responding poses has also begun to be quantified. At the student level, research suggests that students who rapidly guess on roughly 10% or more of the items on a test can have observed achievement scores that are biased downwards by anywhere from .2 standard deviations (Rios, Guo, Mao, & Liu, 2016) to .5 standard deviations (Wise, 2015), on average. Further, the downward bias in observed test scores can also bias correlations between those and other scores. For example, correlations between achievement test and Scholastic Aptitude Test (SAT) scores increased by anywhere from .08 to .12 units when rapid guessing was accounted for in the models (Kong, Wise, & Bhola, 2007; Rios et al., 2014; Swerdzewski, Harmes, & Finney, 2011; Wise & DeMars, 2006). These sources of bias are alarming given more recent research showing that rapid guessing occurs frequently. Soland, Jensen, Keys, Wolk, and Bi (2018) and Jensen, Rice, and Soland (2018) showed that the proportion of students rapidly guessing on 10% or more of the items on a test can reach more than .15 in middle school grades. In short, there is evidence that the bias introduced by test disengagement can be large and widespread.

Despite these findings, the potential impact of rapid guessing on estimates of educational effectiveness (whether at the teacher, school, program, or system level) is still largely unstudied. Emerging research suggests that test disengagement likely affects teacher value added (Jensen et al., 2018) and achievement gap estimates (Soland, 2018a, 2018b), though those effects are likely quite mild in most cases. That research builds on earlier studies showing that the rank orderings of schools can be influenced by rapid guessing (Setzer, Wise, van den Heuvel, & Ling, 2013), as well as those of countries that administer international achievement tests (Eklöf et al., 2014; Wise, Soland, and Bo, 2018; Zamarro, Hitt, & Mendez, 2016). However, there is otherwise little

research that provides detail on how rapid guessing might impact common policy and program evaluation metrics related to effectiveness.

There is also some debate in the measurement literature on how best to account for rapid guessing. The majority of studies, including several related to evaluation, remove students who rapidly guess on a significant portion of items on a given test, a process dubbed "motivation filtering" (Rios et al., 2016). This approach is justifiable under the assumption that there is no correlation between students' true achievement scores and their rapid guessing rates over the course of a test (i.e., that students are not rapidly guessing more often when they are less comfortable with the content). A range of studies chronicled by Wise (2015) provide evidence that this assumption might be justifiable. However, more recent work suggests there is likely a correlation between true achievement and rapid guessing, and that filtering out students can upwardly bias mean achievement estimates by as much as .4 standard deviations (Rios et al., 2016). Instead of filtering, Rios et al. (2016) suggest that less biased estimates of mean achievement can be produced by removing item responses deemed to be rapid guesses, re-estimating individual student achievement, and then taking the mean. To date, the implications of this debate have not been discussed in the context of program evaluation, nor has general consensus been reached.

In this study, we make several contributions to the existing literature. For one, we introduce the concept of rapid guessing for a policy and evaluation audience, including discussing different options for addressing it at the item and student levels. We also walk through empirical examples related to educational effectiveness and show how rapid guessing may influence the results for our sample. Specifically, we ask three research questions about whether rapid guessing may impact estimates of (1) school contributions to student growth (2)

achievement gaps, and (3) summer learning loss. For each case, we produce these estimates in ways that do, and do not, address rapid guessing. In our fourth research question, we briefly examine whether filtering may be introducing bias into our estimates of mean achievement and, thereby, provide insights into the tradeoffs associated with different approaches to addressing rapid guessing. Finally, we discuss how program evaluators and other stakeholders can use our results to examine the sensitivity of their findings to test disengagement.

Literature on Rapid Guessing

In this section, we review literature on (a) approaches to identifying rapid guessing, (b) approaches to accounting for rapid guessing in models, and (c) evidence on how rapid guessing may impact policy and evaluation metrics related to educational effectiveness.

Approaches for Identifying Rapid Guessing

Perhaps the most difficult technical challenge associated with identifying rapid guesses is setting a response time threshold separating engaged and disengaged responses (Guo et al., 2016; Wise, 2015; Wise & Kong, 2005a). In plain terms, how fast does a student need to respond in order to conclude that the content of the item was not fully understood? Initial approaches to setting thresholds involved visually inspecting response time distributions (Wise & Kong, 2005a; Wise & Ma, 2012). Many of these distributions were bimodal, with one of the modes occurring under 10 seconds (Wise & Kong, 2005a). Based on a variety of evidence (described shortly), researchers concluded that responses associated with these modes occurring under 10 seconds were rapid, or disengaged (Kong et al., 2007). Wise and Kong (2005b) used this finding to develop a threshold-setting process that examined the distribution of response times for an item, and set the threshold at 10% of the average time students took to answer the item, with a maximum threshold of 10 seconds.

While the visual inspection approach to setting thresholds is intuitive, there are disadvantages. For example, setting the exact cutoff can be somewhat arbitrary, and there are some items that are not bimodal at all (Guo et al., 2016). To help address this issue, newer approaches have been developed that set thresholds based on how often students with response times below a given cut point get the item correct. Guo et al. (2016) proposed setting the threshold at the response time below which students get the item right at a rate no better than chance. A similar and parallel approach was developed by Lee and Jia (2014). Analyses by other researchers (Goldhammer, Martens, Christoph, & Lüdtke, 2016) have found that these new threshold-setting processes are superior to visual inspection methods when examining criteria like those enumerated in Wise (2015).

Several pieces of evidence are typically used to support the use of rapid guessing as a measure of test disengagement on a given item. Research on related validity evidence was chronicled by Wise (2015) for those interested in greater detail. One validity criterion is that rates of rapid guessing across a test have been shown to correlate strongly with other measures of test engagement, including student self-reports conducted after the test's conclusion (Kong et al., 2007; Rios et al., 2014; Swerdzewski et al., 2011; Wise & Kong, 2005a). Rapid guessing behavior also tends to yield item responses that are correct at a rate consistent with random responding (Demars, 2007; Kong et al., 2007; Setzer et al., 2013; Wise, 2006), a validity criterion now used when setting thresholds (Guo et al., 2016; Goldhammer et al., 2016). For example, on an item with four response categories, students who rapidly guess tend to get those items correct about one quarter of the time.

Perhaps most controversially, there is evidence that rates of rapid guessing are not correlated with a student's true achievement (Demars, 2007; Kong et al., 2007; Rios et al., 2014;

Setzer et al., 2013; Wise & DeMars, 2010; Wise, Pastor, & Kong, 2009). This last piece of evidence is especially crucial to many inferences one might wish to make based on rapid guessing rates. If students are rapidly guessing not because they are disengaged, but because they surmise that they do not understand the content and move on, then such behavior may be a proxy for achievement rather than disengagement. In studies examining the connection between rapid guessing and true achievement, correlations between rapid guessing rates and scores on the SAT ranged from -.05 to .19, with a median correlation of .08 (Wise, 2015). These low to moderate correlations are generally used to suggest that rapid guessing is not primarily a proxy for low achievement (Wise, 2015).

However, more recent studies have questioned the finding that there is no correlation between rapid guessing and true achievement. Rios et al. (2016) showed that correlations between rapid guessing rates and achievement on other tests reported in earlier studies likely understated that relationship. Specifically, Rios et al. (2016) provided evidence that there is a correlation between rapid guessing and true achievement, and that motivation filtering can therefore bias mean test scores upwards by more than .4 standard deviations. Thus, when the assumption that true achievement and rapid guessing rates are uncorrelated is violated, mean achievement estimates can be biased.

Whatever the threshold setting process and evidence to support it, research has shown that rapidly guessing on 10% or more of the items on a test may be sufficient to call the validity of that observed score into question (Wise, 2015). In some cases, that rate has been shown to be as low as 6% of the items on a test (Rios et al., 2016). These cutoffs used to identify rates of rapid guessing that may introduce meaningful bias into observe test scores has implications for certain approaches to correcting for rapid guessing when using those scores, especially filtering.

Approaches to Accounting for Rapid Guessing in Models

In much of the rapid guessing research from the past twenty years reviewed by Wise (2015), aggregate test scores were corrected for rapid guessing by filtering out scores from examinees showing extreme rates of rapid guessing, an approach later dubbed "motivation filtering" (Rios et al., 2016). For example, gaps might be re-estimated after filtering out scores from examinees who rapidly guessed on 10% or more of the items on that test. Such an approach is justifiable only under the assumption that students being filtered out are not different in unobservable ways that might bias estimates. That assumption would be violated if rapid guessing is correlated with true achievement. Researchers felt such a strong assumption was justifiable given previously discussed studies finding that rapid guessing rates have low to modest correlations with observed achievement (DeMars, 2007; Kong et al., 2007; Setzer, Wise, van den Heuvel, & Ling, 2013; Wise 2006; Wise & Kong, 2005).

However, amidst a growing recognition that the assumption of no correlation between true achievement and rapid guessing may be too strong (Rios et al., 2016; Soland, 2018b), other methods to adjust for rapid guessing have been developed and supported with validity evidence. Wise and DeMars (2006) first developed an approach called "effort-moderated scoring" that operates at the item level by treating answers flagged for rapid guessing as uninformative in IRT models. In so doing, item responses deemed rapid guesses are treated as missing data in the models. This approach involves a classic bias-precision tradeoff: achievement estimates are noisier due to having fewer informative items, but do not rely on items that are likely biased due to rapid guessing. Therefore, effort-moderated scoring may be the method that comes closest to the ideal of knowing students' scores if they did not rapidly guess at all.

Ultimately, correcting for rapid guessing using the filtering approach versus the effort-moderated approach relies on very different assumptions. Effort-moderated scoring assumes that item responses associated with response times just above the rapid guessing threshold are yielding meaningful data. If a threshold for an item is set at, say, six seconds, but a student responds in seven seconds, the response is treated as engaged. Yet, it remains unclear whether responses barely above the threshold reflect engagement. If they are not, then effort-moderated scoring is likely to understate the impact of rapid guessing on observed achievement. However, this assumption is more mild than the one required by filtering, namely that there is no correlation between rapid guessing and true achievement (Wise, 2015; Wise & DeMars, 2006; Wise & Kingsbury, 2016). Therefore, filtering is much less conservative in its approach to rapid guessing, essentially throwing away any observed scores that may have been impacted by disengagement. While research increasingly suggests that effort-moderated scoring relies on more justifiable assumptions, both are still used regularly in practice (Wise, 2015).

Evidence on How Rapid Guessing May Impact Estimates of Educational Effectiveness

Vast bodies of literature consider the relevance of estimates of school contributions to student growth, achievement gaps, and summer learning loss to educational policy, evaluation, and practice (Braun, Chapman, & Vezzu, 2010; Briggs & Weeks, 2011; Gershenson & Hayes, 2018; McEachin & Atteberry, 2017; Reardon & Raudenbush, 2009; Reardon & Robinson, 2008). Given how extensive these respective literatures are, we will not review them here. Instead, we will only review literature relevant to how test engagement might impact these three types of estimates.

Though emergent, research shows that wrongly assuming students are not rapidly guessing may impact fundamental inferences educational stakeholders wish to make about

evaluation and policy. For instance, studies have begun to investigate the effect of rapid guessing on rankings of schools and teachers. Setzer et al. (2013) filtered out college students who rapidly guessed on 10% or more of the items on a test, and examined how the rank ordering of 84 institutions changed. One institution with high overall rates of rapid guessing went from 58th to 17th overall. Four other institutions rose by at least five places after filtering, while several others shifted by one spot in the rankings. More recently, Jensen, Rice, and Soland (2018) produced two sets of teacher value added estimates, one using original test scores, the other using effort moderated scores. They found that, while rapid guessing was quite prevalent in middle school grades (in some cases, 15% of students rapidly guessed on 10% or more of the items on the test), the impact on teacher rank orderings was minimal, in part because there was a strong correlation between rapid guessing rates and lagged achievement.

Research has also considered the implications of test disengagement for evaluating the educational effectiveness of countries, such as when nations are rank ordered based on scores from international achievement tests (Eklöf et al., 2014; Wise, Soland, and Bo, 2018; Zamarro, Hitt, & Mendez, 2016). Two studies have shown that there are high correlations between disengagement rates on tests like the Programme for International Student Assessment (PISA) and observed achievement scores at the country level. These studies further hypothesize that such correlations could bias rank orderings of countries. However, Wise, Soland, and Bo (2018) showed that rapid guessing rates did not affect the rank orderings of schools on the OECD Test for Schools, a school-based version of the PISA. Wise et al. (2018) found that there was not much effect because students with exceptionally low achievement relative to the item difficulties on the test actually had their observed achievement scores inflated by rapid guessing in some cases rather than deflated. That is, some students actually performed better when rapidly

guessing rather than trying to answer the items in earnest. This finding likely occurred because the OECD Test for Schools is not computer adaptive and therefore presented students with items they found extremely difficult, unlike in much of the research on rapid guessing that uses computer-adaptive tests (Wise, 2015).

Research has also begun to show that there are differential rates of rapid guessing across student subgroups, which may affect achievement gap estimates for certain students. Studies consistently show that students who rapidly guess on 5-10% of items on a test (or more) can have their observed test scores biased downwards compared to their true scores, oftentimes by more than .2 standard deviations (Rios et al., 2016). If different groups of students rapidly guess at inconsistent rates, then their observed achievement scores are likely downwardly biased at differing rates, too. Soland (2018a) estimated achievement gaps conditional on rapid guessing rates over the course of the test and found that, while most gaps were not affected substantially, some changed by more than .1 standard deviation in later grades, especially male-female gaps. In a follow-up study, Soland (2018b) estimated achievement gaps using different approaches to correcting for rapid guessing, including filtering and estimating effort moderated scores. While results continued to suggest that gaps estimates could be sensitive to rapid guessing, and that the direction of male-female gaps could change from favoring girls to favoring boys, the magnitude of those changes was much more modest when relying on effort moderated scores compared to conditioning on rapid guessing rates (Soland, 2018a).

Currently, no studies examine whether rapid guessing affects estimates of summer learning loss. This omission is somewhat surprising given research showing that rates of rapid guessing often differ between fall and spring test administrations, perhaps because tests administered in spring are often associated with higher stakes under federal accountability

(Jensen et al., 2018; Wise, 2015). Further, if rates of rapid guessing do differ between fall and spring, then there could be implications for the growing body of research suggesting that estimates of school and teacher effectiveness differ substantively dependent on whether summer learning loss is accounted for in the statistical models (Gershenson & Hayes, 2018; McEachin & Atteberry, 2017). Estimates of summer learning loss are relevant to program evaluation in terms of how different educational programs might reduce summer loss, as well as how failing to account for what happens during the summer impacts evaluations (Gershenson & Hayes, 2018; McEachin & Atteberry, 2017). Our study is the first to consider whether rapid guessing biases estimates of summer learning loss and, potentially, estimates of teacher and school effectiveness.

Methods

In this section, we discuss our measures used, analytic sample, and modeling approach, including how each research question is addressed.

Measures

Test engagement. To identify rapid guesses, we followed the empirical approach developed by Wise and Kong (2005b) and Wise and Ma (2012) to set item thresholds. Under this approach, the threshold for rapid responding is 10% of the average time students take to answer an item with a maximum time of 10 seconds. These thresholds were previously set using a nationally representative sample of students, which helps ensure they are not sensitive to our sample. We compared a subset of items deemed rapid under the Wise and Ma (2012) method to the one developed by Guo et al. (2016), but found very high overlap, likely because the frequency with which students answer an item correctly was used as a criterion to validate the thresholds set using the Wise and Ma (2012) approach.

After the item thresholds were established, each item response was flagged as a rapid response if the response time was less than the item threshold. A student was flagged as disengaged if over 10% of his or her item responses were rapid, which in the context of the assessment used was typically four or more items.

Reading achievement. Student test scores from the NWEA MAP Growth reading assessment were used in this study. The MAP Growth assessments are CBTs typically administered three times a year in the fall, winter, and spring. Each test takes approximately 40 to 60 minutes depending on the grade and subject area. Students respond to assessment items in order (without the ability to return to previous items), and a test event is finished when a student completes all the test items (typically 40 items for reading). Test scores, called "RITs," are reported in an IRT-based metric. Further, the test is vertically scaled, allowing for arithmetic comparisons in evaluating growth across grades. In general, MAP Growth is not used for high-stakes purposes at the state level, though some districts and schools may use it to help screen students for special education and gifted programs (we discuss the low-stakes nature of the test in the limitations section).

Much of the literature on rapid guessing uses MAP Growth due to particular qualities of the test. For instance, MAP growth is not timed, so students are not rushed at the end in a way that might reduce their response times in ways unrelated to engagement (Wise & Kingsbury, 2016). Also, students must provide a response to every item, which means there are no missing responses to manage. Finally, the test is adaptive, reducing the likelihood that students are receiving extremely difficult items and responding by rapidly guessing (Wise & Kingsbury, 2016; Wise, 2006).

Sample

The data used in this study came from a longitudinal cohort of students in a Midwestern U.S. state who were in fourth grade in 2010–2011 and were tracked through their 8th-grade year in 2014–2015. For the current analysis, we restricted the sample to students who had at least two test scores in fourth or fifth grade. The main implication of this decision is that we excluded students who entered the sample during middle school. We did so because one of our research questions focuses on elementary schools' contributions to learning trajectories, and we were unable to identify the elementary school attended for students tested only in middle school grades. For the purposes of estimating these school contributions, we associated students with their modal elementary school. We further restricted our sample in those analyses to only include schools with at least 10 students, an approach with some precedent in the value-added literature (Loeb, Soland, & Fox, 2014; Soland, 2017). While decisions to limit our sample as we did may impact estimates of school contributions to student growth, our intention is not to generalize beyond our sample, nor to identify particular schools as effective or ineffective. Rather, while our results can provide insights into how rapid guessing might affect common policy and evaluation metrics related to efficacy, these analyses are primarily illustrative rather than broadly generalizable.

In total, the analytic sample included information for 22,055 students who were in 343 elementary schools in 2010-11 school year. The sample is 49% female, 64% White, 25% Black, seven percent Hispanic, and four percent Asian. Table 1 displays descriptive statistics by grade and term for the percent of students with disengaged test events (rapid guessing on 10% or more of the test items) and the average RIT score, including the unadjusted RIT, the RIT filtering out disengaged test events, and the effort-moderated RIT. The percentage of disengaged test events was consistently higher in the fall than spring, and went up as students moved from 4th to 8th

grade. The filtered RIT and effort-moderated RIT scores were consistently higher than the unadjusted RIT by approximately 1-2 RIT points, on average. Additionally, the sample variability was smaller within each time point after filtering out disengaged test takers.

Table 2 shows the proportion of students within the study sample who met the disengaged test event criterion by grade, term, and race/ethnicity. As seen in Soland (2018a), rates of disengaged test taking varied greatly by subgroup, with Black students showing the highest rates of rapid guessing followed by White and Hispanics students (their rapid guessing rates are fairly comparable), and Asian students consistently showing the lowest rates. As expected based on prior literature, rapid guessing rates generally increase as students get older across subgroups.

Table 2 also reveals that one term (7th grade fall in 2013-14) had unusually high rates of rapid responses across all groups. While we have not been able to identify with certainty why the rates are so much higher in this term, we believe it may have resulted from issues with how long it took items to render on screens for some test events. (As an aside, these types of patterns also reveal why tracking rapid responding can be useful for identifying abnormal patterns of test behavior that may indicate issues with test procedures.) To check whether our results are robust to this anomalous time point, we fit all of our models using a sample that excluded 7th grade. Across all estimated parameters, the statistical significance of differences between original and engagement-adjusted estimates did not change when the 7th grade assessment were excluded.

Modeling Approach

We used three-level random effects (hierarchical) linear models to investigate how rapid-guessing behavior affects estimates of schools' contributions to student growth, racial/ethnic achievement gaps in student learning, and summer learning loss. For each research question, the

dependent variable is the student's (unadjusted or adjusted) RIT score, y_{tij}, which is associated with year/term *t* for student *i* in school *j*. Each student contributes up to ten test scores (fall and spring for five school years), which are treated as repeated measures nested within students (level 2) and schools (level 3). The three-level model assumes students are nested within a single school over time, an assumption that is violated frequently given the grade span of focus for these analyses. To deal with this issue, we assigned each student to his or her modal elementary school for the entire test score history. One should also note that statistical models have been formulated to address complications that arise when matching students to teachers or schools when there is student mobility complicating the match, and that such models could be integrated with our growth model design (Bates, 2010; Daniel, 2012; Lockwood, McCaffrey, Mariano, & Setodji, 2007).

The growth structure is specified using a Compound Polynomial (CP) set-up, which is described in greater detail in Thum (2018) and Thum and Matta (2015; 2016). The CP modeling approach was chosen over more typical longitudinal designs such as the polynomial growth model because the CP model has been found to fit longitudinal data with clear seasonality (e.g., patterns of within-school year gains followed by summer losses) better than more conventional growth model designs (Thum, 2018). Additionally, the CP growth model design provides a useful parameterization for studying educational processes, allowing for the simultaneous estimation of a student's overall learning trajectory from fall of fourth grade to the fall of eighth grade and the average rate of within-school (fall-to-spring) growth across grades. Nonetheless, our findings on the effect of rapid guessing on the parameters we estimate are not generally sensitive to use of the CP model versus a more standard polynomial growth model. Further detail on the CP model is provided in the appendix.

The specification of our three-level hierarchical linear model is given in Equation 1:

Equation 1. Level-1 Model (Repeated observations of MAP scores (t) within student (i) and school (j)):

$$y_{tij} = \sum_{k=0}^{4} \pi_{kij} X_{ktij} + e_{tij}$$

Level-2 Model (student (i) within school (j)):

$$\pi_{0ij} = \beta_{00j} + r_{0ij}$$
$$\pi_{1ij} = \beta_{10j} + r_{1ij}$$
$$\pi_{2ij} = \beta_{20j} + r_{2ij}$$
$$\pi_{3ij} = \beta_{30j} + r_{3ij}$$
$$\pi_{4ij} = \beta_{40j} + r_{4ij}$$

Level-3 Model (school (j)):

$\beta_{00j} = \gamma_{000} + u_{00j}$, where γ_{000} is the predicted Fall score at 4th grade
$\beta_{10j} = \gamma_{100} + u_{10j}$, where γ_{100} is the linear growth rate of change of Fall scores
$\beta_{20j} = \gamma_{200} + u_{20j}$, where γ_{200} is the quadratic growth rate of change of Fall scores
$\beta_{30j} = \gamma_{300} + u_{30j}$, where γ_{300} is the predicted Fall $-$ Spring growth in 4th grade
$\beta_{40j} = \gamma_{400} + u_{40j}$, where γ_{400} is the linear growth rate of change for the Fall $-$ Spring change

Variance component specification:

$$e_{tij} \sim \mathrm{N}(0, \sigma^2_{tij}), \quad \boldsymbol{r_{ij}} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{T_{\beta}}), \quad \boldsymbol{u_j} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{T_{\gamma}}),$$

The first three terms ($X_0 - X_2$) represent a standard quadratic growth model estimating change in fall status from 4th to 8th grade, with X_0 representing the fall status (the start of 4th grade fall), X_1 representing fall-to-fall linear growth from 4th to 8th grade, and X_2 representing fall-to-fall quadratic growth. The second set of terms ($X_3 - X_4$) represent the part of the model where fall-to-spring (within school year) change is estimated. Since time is centered at fall of 4th grade, X_3 represents the predicted change in scores between the fall and spring of 4th grade, and X_4 represents the linear trend in the fall-to-spring change across grade-levels. The specification

of the full design matrix ($\mathbf{X}_{ij}$) is provided in the appendix. Random effects were included at the student and school-level for the all of the level-1 parameters.

We used this general modeling framework to address our first three research questions by adapting the model slightly for each. Further, within each question, we dealt with disengaged responses in three different ways. Thus, our combination of three models and three approaches to addressing rapid guessing led to a total of nine sets of parameter estimates. Below, we detail the models for each question followed by the three approaches to accounting for rapid guessing, including how we did analyses to see whether filtering might be introducing bias into mean achievement estimates in our sample.

Model 1. School contributions to student growth. For this question, we used the model shown in Equation 1. When examining school contributions to student growth, the primary school-level parameters of interest were (a) the average fall-to-fall linear growth, γ_{100}, (b) the fall-to-spring slope in fourth grade, γ_{300}, and (c) the rate of change in fall-to-spring gains from fourth to eighth grade, γ_{400}. In addition, we produced intraclass correlation coefficients (ICCs) associated with each of the parameters of interest. Finally, we compared school-level empirical Bayes estimates for each of the three parameters, a fairly common approach in the value-added literature (Jensen et al., 2018; Loeb et al., 2014; Soland, 2017).

Model 2. Racial/ethnic achievement gaps in student growth. To investigate gaps in student achievement and growth, we built on the basic CP model structure in Equation 1 by including time-invariant, student-level characteristics at level 2 to investigate whether variation in student learning trajectories is associated with race/ethnicity. Specifically, we included three indicators for race/ethnicity: Asian, Black, and Hispanic. By omitting an indicator for White students, we could estimate the development of White-Black, White-Hispanic, and White-Asian

achievement gaps in fourth grade and learning trajectories from 4th to 8th grade. For this question, we also limited our sample only to students identified as Asian, Black, Hispanic, or White.

Model 3. Summer learning loss. The primary goal of this analysis is to examine whether summer learning loss estimates are affected by the presence of rapid guessing. To estimate summer learning loss, we used the same CP modeling framework, but revised the coding of the design matrix to allow for the direct estimation of summer learning loss. In this model set-up, the first three terms ($X_0 - X_2$) represent a standard quadratic growth model for change in spring status from 4th to 8th grade (rather than change in fall status in prior models), X_3 represents spring-to-fall (summer) change between 4th and 5th grade, and X_4 represents the change in the summer learning slope across grades.

Methods for Accounting for Rapid Guessing

Three different approaches were used to estimate each model. The first approach used the unadjusted RIT score as the outcome variable in each model, providing a baseline for comparing the other two approaches to account for disengagement. The second approach is motivation filtering (Rios et al., 2016), where all scores from examinees who rapidly guessed on 10% or more of the items were removed before running the analyses. For the last approach, we used the effort-moderated scoring model described in the background section, and detailed in Wise and DeMars (2006), Wise and Kingsbury (2016), and Soland (2018b). This method re-scores achievement tests in a way that treats rapid guesses as uninformative. Therefore, in the presence of rapid guessing, a student's effort-moderated RIT score is calculated using the same Rasch model as a standard RIT score, but using only the items on which the student did not rapidly

guess. For any students who did not rapidly respond in a grade/term, the effort-moderated score estimate is simply his or her unadjusted RIT score.

We take two different approaches to comparing estimates that do not adjust for rapid guessing to those that do. First, we looked to see if the change in the coefficients between the models using the adjusted and unadjusted scores was statistically significant. We tested the significance by converting the coefficients to a Z-score using the below formula:

$$\frac{\beta_{1adjusted} - \beta_{1unadjusted}}{\sqrt{SE^2_{\beta 1adj} + SE^2_{\beta 1unadj}}}$$

where $SE^2_{\beta 1adj}$ is the squared standard error of the adjusted estimate and $SE^2_{\beta 1unadj}$ is the squared standard error of the unadjusted estimate. Second, to the extent possible, we also considered the practical significance of shifts in parameter estimates between adjusted and unadjusted models. For example, when estimating achievement gaps, we looked for cases in which the gap estimate switched from favoring one subgroup to favoring another.

Investigating Potential Bias Due to Filtering

In our final research question, we investigated whether there is evidence that filtering students may be introducing bias into our estimates, a finding made previously by Rios et al. (2016). We approached the question by examining mean effort-moderated scores and student demographics by proportion of items that are rapid guesses on a given test. Mean effort-moderated scores are meant to show whether achievement estimates that remove rapid guesses are still associated with rates of rapid guessing. If they are, then one might worry that students are more likely to rapidly guess if they have lower true achievement, and that removing them from the sample may upwardly bias mean achievement estimates in ways described by Rios et al. (2016).

Results

Question 1: Does Rapid Guessing Affect Estimates of School Contributions to Student Learning Trajectories?

Table 3 contains the fixed effects parameter estimates and ICCs from the unconditional three-level hierarchical linear model (HLM) estimated using the three different approaches for addressing disengaged test behavior. As shown in the table, differences in estimates of school contributions to fall-to-fall linear growth between the unadjusted and filtered models are significant at the .05 level. None of the other differences are significant. Beyond statistical significance, the magnitude of the changes in the estimates across models was modest at best. The linear growth rate in the fall of fourth grade ranged from 6.25 points with the unadjusted RIT scores to 6.67 points when filtering. The shifts in within-school year gain estimates in 4^{th} grade were similarly small, with estimates ranging from 7.58 to 7.63 RIT points.

Estimates of school contributions to student learning also do not appear to be sensitive to test engagement. Figure 1 shows scatterplots of empirical Bayes estimates of school level random effects for fall status, fall-to-fall growth, and fall-to-spring (within-year) growth for the first year of the sample. Each scatterplot compares estimates using original RIT scores to those using either filtering or effort-moderated scores for a total of six scatterplots. Correlations between original and adjusted estimates are high. When filtering is used, correlations range from .914 to .978. For effort-moderated scores, correlations are all roughly at or above .99.

One important question when studying school effectiveness is the percentage of variance that lies between schools in both initial status and student learning rates. The ICCs reported in Table 3 show sensitivity to the approach used to adjust for disengagement. The percent of variance that is between schools in students' longitudinal (fall-to-fall) growth shrank when test

engagement was addressed, going from 37% with the unadjusted RIT scores to 27% with the filtered RIT scores. The fall-to-fall ICC for the effort-moderated score is between the other two ICCs at 32%. Similar drops were seen in the ICC for fall-to-spring growth in fourth grade, with a drop of 15 percentage points between the unadjusted and filtered models.

Question 2: Does Rapid Guessing Affect Racial/ethnic Achievement Gaps in Student Learning Trajectories?

In Figure 2, we present estimated mean trajectories for different racial and ethnic groups using unadjusted scores (solid lines) and filtered RIT scores (dotted lines). In this plot, we can see the group trajectories of the filtered test takers are mostly parallel to those of all students, though with generally smaller summer learning drops for the latter. On average, filtered means are higher than unadjusted means (as one would expect), and these differences increase during middle school. Further, filtered scores increase most for Black students, which one would also expect given higher rates of rapid guessing among this subgroup. As a result, Black-White gaps appear to shrink when using filtered scores.

By incorporating student-invariant predictors at level 2 of the HLMs, we can more precisely quantify racial/ethnic achievement gaps with respect to fourth grade test scores, fall-to-fall growth, and fall-to-spring (within-year) growth. The results from these HLMs are presented in Table 4. While significant achievement gaps were observed between White and Black and White and Hispanic students in fourth grade, the fall status parameter estimates from the three models were not significantly different from each other, nor were any other parameter estimates.

However, for the fall-to-fall growth parameter, conclusions about how racial and ethnic achievement gaps grow or shrink across grades were sensitive to rapid guessing. Using the unadjusted RIT scores, Black students showed slightly smaller fall-to-fall learning gains than

White students (slope difference of -.06 points per year). By comparison, when disengaged scores were filtered out, Black students showed significantly greater fall-to-fall growth than White students (.43 RIT points). Results using effort-moderated scores differed from those using unadjusted scores in less dramatic ways, with the point estimates differing by .19 units. Similar shifts in gap estimates between White and Hispanic students in fall-to-fall growth were also observed. That is to say, our inferences about whether gaps are expanding as students move through school appear to be impacted by removing students who did not show engaged test behavior. However, while using effort-moderated scores also impacted the fall-to-fall linear slope estimates, the estimates were not significantly different than those based on the unadjusted scores.

Question 3: Does Rapid Guessing Affect Estimates of Summer Learning Loss?

Table 5 presents results from models designed to estimate summer learning loss. While we show all of the parameter estimates from the model, the primary estimate of interest is the coefficient representing change in achievement levels during the summer between 4^{th} and 5^{th} grade. As the table shows, the change in these estimates between unadjusted and adjusted models is significant when filtering, but not significant when using effort-moderated scores. Specifically, students in the unadjusted sample have an average drop of 1.68 RIT points over the summer compared to an average drop of 1.31 RIT points in the filtered sample and 1.47 points in the effort-moderated model.

Question 4: Is There Evidence That Motivation Filtering May Introduce Bias Into Mean Achievement Estimates?

Table 6 shows mean effort-moderate scores and proportions of students by race and sex for all students, students who were filtered out of the sample (more than 10% of item responses

were rapid), and those who were not (fewer than 10% of item responses were rapid). We provide these descriptive statistics for Spring of 8th grade in reading, a time-subject combination for which students were relatively more likely to rapidly guess. However, results are consistent in other time-subject combinations. Students who did not rapidly guess enough to be filtered out had much higher effort-moderated scores than students who did (roughly 14 RIT points). Thus, this table suggests that students who rapidly guess often have much lower mean achievement, even when estimated achievement removes items deemed rapid guesses. While effort moderated scores are not a perfect proxy for true achievement, these results should raise concerns that filtering students with high rates of rapid guessing from the sample may introduce bias by removing students who have low true achievement, are racial minorities, or both.

Discussion

Educators and researchers have long recognized test disengagement as a potential problem for the inferences one might wish to make based on observed test scores (Schnipke, 1996; Wise, 2015). Thanks to item metadata captured when tests are administered via computer, psychometricians have begun to quantify disengagement on items, which can in turn be used to estimate disengagement rates across a test. In particular, item response times have been used to identify rapid guessing behavior, which occurs when students respond to an item so quickly, its content could not have been understood (Rios et al., 2016; Wise & Kong, 2005b). While this so-called "rapid guessing behavior" has been shown to severely bias observed achievement test scores when it occurs often during a test event (Rios et al., 2016), relatively little is known about how this behavior might impact parameters relevant to evaluating the effectiveness of teachers, schools, programs, and school systems.

In this paper, we attempted to define rapid guessing for an evaluation and policy audience, discuss ways of correcting for rapid guessing, and illustrate the effects of rapid guessing on three metrics relevant to policy and evaluation: estimates of school effectiveness, achievement gaps, and summer learning loss. Through our examples, we show that the effect of rapid guessing on these metrics using our methods and sample is fairly minimal. On one hand, certain parameters related to school contributions to student growth and summer learning loss change significantly relative to unadjusted estimates when using motivation filtering. On the other, unadjusted and effort-moderated point estimates differ, but those differences are generally indistinguishable from zero. Even when using motivation filtering, the change in the point estimates are often not large with the exception of Black-White achievement gap estimates of fall-to-fall linear growth.

The discrepancy in findings between approaches to adjusting for rapid guessing might suggest uncertainty about the effect of rapid guessing were it not for research suggesting that motivation filtering can induce upward bias into mean achievement estimates. This bias, which likely occurs due to a correlation between true achievement and rapid guessing rates, can be as large as .4 standard deviations (Rios et al., 2016). That is, filtering can introduce bias in mean achievement estimates because students who are lower performing (even after accounting for the downward bias due to rapid guessing) are more likely to be removed. We presented filtered results despite the findings of Rios et al. (2016) because this approach to addressing rapid guessing is so predominant in the associated literature, chronicled by Wise (2015).

When we compared mean effort-moderated RIT scores in reading for students who were filtered out of estimates to those for students who were not filtered, the former had mean effort-moderated scores that were 14 RITs lower, on average, than students who were not filtered. This

finding, detailed in Table 6, suggests that there is a strong association between frequent rapid guessing and estimates of achievement that remove those rapid guesses. Thus, filtering removes students who are much lower achieving, even when accounting for rapid guessing when estimating achievement. While effort-moderated scores are imperfect proxies of true achievement (requiring one to assume that responses just above the rapid guessing threshold are providing valid data), this association does provide some evidence that the findings of Rios et al. (2016) may be applicable in our sample. Thus, the significant results we see using filtered results may be due to bias introduced by removing students with low true achievement, not actual differences in test engagement.

However, before concluding that test disengagement is unlikely to impact metrics relevant to educational effectiveness, there are a few additional factors to consider. Most importantly, rapid guessing metrics are conservative in two relevant ways. First, the thresholds used to identify rapid guesses are set very conservatively in order to avoid wrongly discarding item responses that are engaged (Demars, 2007; Rios et al., 2016; Wise & Kong, 2005b). For the purposes of program evaluation, where individual scores are somewhat less important than the aggregated metrics they produce, evaluators might be more willing to risk discarding engaged item responses to reduce bias in the aggregate. Second, rapid guessing rates are conservative as a measure of test disengagement given they only capture one particular behavior (Jensen et al., 2018; Soland, 2018a). That is, there are many other potential behaviors associated with disengagement not included in the measure. For example, a student could leave an item on the screen for a long period of time, pay it no attention, and arbitrarily select a response, which would not be captured by rapid guessing measures. Ultimately, if we could quantify all forms of disengagement, not just rapid guessing, the results might be more pronounced.

Limitations

There are still several ways that test disengagement could affect policy and evaluation metrics that have yet to be investigated. While the metrics we examined—school contributions to student growth, achievement gaps, and summer learning loss—are relevant to evaluation, the impact of rapid guessing on an actual evaluation of a specific program has never been explored. The literature could benefit from an examination of rapid guessing in the context of a particular program and an estimation of its impact. Hypothetically, one might imagine that students participating in a study with a pre- and post-test design could behave differently on the pre-test than the post-test if educators administering the assessments communicate the importance of the respective tests differently, whether intentionally or unintentionally.

Further, the results in our study may not generalize to other contexts. Our results pertain to the reading achievement for only a single cohort in a lone Midwestern state; one cannot be sure whether findings would generalize to other subjects, states, time periods, and student subgroups, though our findings generally match those from other studies (Jensen et al., 2018; Soland et al., 2018; Wise, 2015). Additional research should attempt to replicate our findings in different contexts. As discussed previously, our intent was not to produce generalizable results, but to use our sample as a way to illustrate how to account for rapid guessing when estimating parameters of interest and to explore the effect of rapid guessing on those parameters in our sample.

Our findings could also fail to generalize to different achievement tests. For instance, while studies suggest that test engagement, and rapid guessing in particular, occurs across most tests that have been studied, there are differences in rapid guessing rates dependent on whether the assessment is computer-adaptive versus fixed-form (Wise, 2015). Further, very little

research on rapid guessing has been conducted using tests with high stakes for students, which may reduce the likelihood that students rapidly guess (Swerdzewski et al., 2011). Results from our study could be replicated in a high-stakes testing context, and with tests that are not adaptive.

Challenges and Promise for Research on Educational Effectiveness

Program evaluators, policymakers, researchers, and other educational stakeholders may be wondering how best to apply this study to their own work. For simplicity, we will focus on the example of a program evaluation. While identifying rapid guesses and re-estimating achievement test scores to account for them is fairly straightforward, there are of course challenges. For instance, evaluators would need to be using a CBT that captures reliable item-level metadata on response times. The test would also need to be of sufficient length to remove rapidly guessed items without resulting in too much imprecision for the scores to be meaningful. While such conditions are not atypical, there are likely many evaluators who do not have a CBT, item-level data, associated metadata, or some combination of the three.

For evaluators who do not have access to such data, research provides several indicators of when rapid guessing is more or less likely to be a concern. Specifically, studies including our own show that rapid guessing rates are higher in later grades and in reading (Jensen, Rice, & Soland, 2018). Research also suggests that there is more likely to be a correlation between true achievement and rapid guessing, which could disproportionately affect scores for low-achieving students, if the test is not adaptive (Rios et al., 2016). In terms of subgroup comparisons, the biggest disparities often occur between male and female students, which means inferences about their relative performance may be more questionable under the assumption of no disengagement (Soland, 2018a, 2018b). However, in our study, rapid guessing had a mild impact even in worst-

case scenarios like comparing male versus female reading achievement in later grades. How risk-averse the evaluator wishes to be is obviously an individual decision.

In the event evaluators do have appropriate data and wish to make sure their results are not sensitive to rapid guessing, there are several promising approaches. First, response time thresholds separating engaged and disengaged responses would need to be set. One straightforward option is to use the rule of thumb developed by Wise and Kong (2005) and Wise and Ma (2012), which sets the threshold at the 10^{th} percentile of the item's response time distribution with a maximum of 10 seconds. However, recent research suggests that thresholds can better avoid misclassification of responses by being set where students get the item correct at a rate no better than chance. For example, on a multiple-choice item with five response options, the threshold would be set at the point where responses are correct roughly 20% of the time with some sort of maximum time allowance. Fairly clear guidelines and additional details on these approaches are provided by Guo et al. (2016). Nuances aside, if evaluators want a rough check on rapid guessing rates, they could start by identifying what proportion of items had response times under ten seconds (the maximum allowed under many threshold-setting approaches), and how those rapid responses were clustered by student, treatment condition, and subgroup.

From there, achievement test scores would need to be corrected for rapid guessing. The most defensible and conservative approach is to re-estimate IRT-based scores treating rapid guessing as missing, use those scores in models, and compare results, just as we did in our own study. As previously mentioned, care must be taken to ensure that this approach does not introduce too much imprecision into achievement estimates. While motivation filtering is an option, it increases the likelihood of upwardly biasing mean estimates of achievement if there is any correlation between true achievement and rapid guessing rates, an assumption with some

evidence to support it (Rios et al., 2016), including in our sample. Nonetheless, filtering is much less labor intensive: evaluators could perform a basic sensitivity check by comparing filtered and unfiltered results, then move on to effort-moderated scoring if results are sensitive.

Finally, evaluators should be aware that there is considerable emerging research in the field of assessment metadata, including how it relates to test engagement, that is worth monitoring. For example, a large and growing body of research shows that disengagement is often an even bigger issues on surveys (Curran, 2015; Meade & Craig, 2012). Further, research has begun to develop ways to detect disengaged survey responses including using response time metadata, though these efforts are complicated because survey items do not have correct answers (Soland, Wise, & Gao, 2018).

There is also work afoot to identify new sources of metadata related to disengagement. For instance, Wise and Gao (2017) compared rapid guessing rates to how frequently students entered nonsensical text in response to open-ended items like essay questions, and found an association between the two behaviors. In the literature on disengaged responding among survey takers, initial research suggests that spending abnormally long amounts of time on an item may also signal disengagement (Curran, 2015). As the sources of metadata and their uses develop, evaluators and other stakeholders will likely be given new tools to safeguard their results against disengagement.

Conclusion

The use of response time metadata from CBTs to measure test disengagement provides new opportunities to account for construct irrelevant variance in studies of educational effectiveness. We described a method for identifying disengaged test takers based on flagging responses to multiple choice test items that were made in less time than would be needed to read

and understand the item (i.e., rapid guessing). This rapid guessing metric is useful for understanding overall student engagement on the assessment of interest, but also can highlight whether subgroups of interest were similarly engaged in testing across timepoints. We found both age and racial/ethnic differences in the degree to which students were disengaged test-takers. Similar analyses could be conducted in the context of evaluation studies to ensure that there are not significant differences in level of test engagement between treatment and control groups, as well as between pre- and post-test assessments.

We also presented two different approaches to account for test disengagement when studying educational processes. Our findings indicated that inferences about schools' contributions to growth, achievement gaps, and summer learning loss were mostly unaffected by rapid guessing behaviors. Nonetheless, these approaches are still worth investigating in other evaluation contexts as a sensitivity check to test whether observed improvements in student achievement are biased by differential test engagement.

References

Bates, J. K. R. D. (2010). Cross-classified models in the context of value-added modeling.

Braun, H., Chapman, L., & Vezzu, S. (2010). The black-white achievement gap revisited. *Education Policy Analysis Archives*, *18*, 21.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*, *36*(5), 616–637.

Curran, P. G. (2015). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

Daniel, L. H. (2012). *Comparing cross-classified growth models with and without the cumulative effect of teachers to a hierarchical growth model on cross-classified data* (PhD Thesis). University of Pittsburgh.

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*(1), 23–45.

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, *27*(1), 31–45.

Gershenson, S., & Hayes, M. S. (2018). The implications of summer learning loss for value-added estimates of teacher effectiveness. *Educational Policy*, *32*(1), 55–85.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-Taking Engagement in PIAAC. OECD Education Working Papers, No. 133. *OECD Publishing*.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A New Procedure for Detection of Students' Rapid Guessing Responses Using Response Time. *Applied Measurement in Education*, *29*(3), 173–183.

Jensen, N., Rice, A., & Soland, J. (2018). The Influence of Rapidly Guessed Item Responses on Teacher Value-Added Estimates: Implications for Policy and Practice. *Educational Evaluation and Policy Analysis*, 0162373718759600.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606–619.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(1), 8.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*(2), 125–150.

Loeb, S., Soland, J., & Fox, L. (2014). Is a Good Teacher a Good Teacher for All? Comparing Value-Added of Teachers With Their English Learners and Non-English Learners. *Educational Evaluation and Policy Analysis*.

McEachin, A., & Atteberry, A. (2017). The impact of summer learning loss on measures of school performance. *Education Finance and Policy*, *12*(4), 468–491.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519.

Reardon, S. F., & Robinson, J. P. (2008). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. *Handbook of Research in Education Finance and Policy*, 497–516.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2016). Evaluating the Impact of Careless Responding on Aggregated-Scores: To Filter Unmotivated Examinees or Not? *International Journal of Testing*, 1–31.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying Low-Effort Examinees on Student Learning Outcomes Assessment: A Comparison of Two Approaches. *New Directions for Institutional Research*, *2014*(161), 69–82.

Schnipke, D. L. (1996). *How Contaminated by Guessing Are Item-Parameter Estimates and What Can Be Done about It?*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34–49.

Soland, J. (2017). Is Teacher Value Added a Matter of Scale? The Practical Consequences of Treating an Ordinal Scale as Interval for Estimation of Teacher Effects. *Applied Measurement in Education*, *30*(1), 52–70.

Soland, J. (2018). Are Achievement Gap Estimates Biased by Differential Student Test Effort? Putting an Important Policy Metric to the Test. *Teachers College Record*.

Soland, J. (2018). The Achievement Gap or the Engagement Gap? Investigating the Sensitivity of Gaps Estimates to Test Motivation. *Applied Measurement in Education*.

Soland, J., Jensen, N., Keys, T., Wolk, E., & Bi, S. (2018). *Is Low Test Motivation a Sign of Disengagement from School? Examining Indicators of Dropout Conditional on Rapid Guessing Scores.* Manuscript revised and resubmitted.

Soland, J., Wise, S., & Gao, L. (2018). *Avoiding False Positives When Identifying Disengaged Survey Responses: New Evidence Using Response Time Metadata*. Manuscript submitted for publication.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114.

Wise, S. L. (2015). Effort Analysis: Individual Score Validation of Achievement Test Data. *Applied Measurement in Education*, *28*(3), 237–252.

Wise, S. L., & DeMars, C. E. (2006). An Application of Item Response Time: The Effort-Moderated IRT Model. *Journal of Educational Measurement*, *43*(1), 19–38.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling Student Test-Taking Motivation in the Context of an Adaptive Achievement Test. *Journal of Educational Measurement*, *53*(1), 86–105.

Wise, S. L., & Kong, X. (2005a). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wise, S. L., & Kong, X. (2005b). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. In *annual meeting of the National Council on Measurement in Education, Vancouver, Canada*.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205.

Zamarro, G., Hitt, C., & Mendez, I. (2016). When Students Don't Care: Reexamining International Differences in Achievement and Non-Cognitive Skills. EDRE Working Paper No. 2016-18.

Table 1

Statistics on Analytical Sample

		Fall					Spring				
Year	Grade	N	Dis-engaged test events	RIT (un-adjusted)	RIT (Engaged test events)	RIT (Effort-moderated)	N	Dis-engaged test events	RIT (un-adjusted)	RIT (Engaged test events)	RIT (Effort-moderated)
2010-11	4	12,668	0.07	197.1 (15.6)	198.0 (15.3)	197.3 (15.4)	14,949	0.05	205.7 (14.6)	206.5 (14.1)	205.9 (14.5)
2011-12	5	16,945	0.07	205.0 (15.4)	206.0 (14.9)	205.2 (15.2)	18,883	0.05	211.7 (14.8)	212.6 (14.2)	211.9 (14.6)
2012-13	6	15,251	0.12	209.1 (15.6)	210.8 (14.8)	209.6 (15.3)	15,862	0.10	214.8 (14.9)	216.5 (13.8)	215.2 (14.5)
2013-14	7	14,262	0.20	213.9 (15.6)	216.2 (14.3)	214.5 (15.1)	14,729	0.10	218.8 (15.0)	220.5 (13.9)	219.2 (14.6)
2014-15	8	12,238	0.14	217.6 (15.7)	219.9 (14.5)	218.3 (15.2)	12,113	0.11	221.9 (15.5)	223.7 (14.3)	222.3 (15.0)

Note. Standards deviations are in parentheses. Disengaged test events occurred when a student rapidly guessed on 10% or more of the items.

Table 2

Proportion of Students with Disengaged Test Events (Rapid Guessed on 10% or More Items) by Gender and Race/ethnicity

Grade	Term	Year	White	Black	Asian	Hispanic
4	Fall	2010-11	0.06	0.09	0.03	0.05
5	Fall	2011-12	0.06	0.11	0.02	0.06
6	Fall	2012-13	0.11	0.16	0.03	0.12
7	Fall	2013-14	0.19	0.24	0.25	0.20
8	Fall	2014-15	0.13	0.17	0.03	0.15
4	Spring	2010-11	0.05	0.08	0.02	0.04
5	Spring	2011-12	0.04	0.08	0.02	0.04
6	Spring	2012-13	0.09	0.15	0.04	0.12
7	Spring	2013-14	0.10	0.14	0.01	0.10
8	Spring	2014-15	0.10	0.14	0.02	0.09

Table 3

Estimates of School's Contributions to Student Growth Accounting for Rapid Guessing

	Fixed effects			Intraclass Correlation		
Parameter	No adjustment	Filtering	Effort-moderated score	No adjustment	Filtering	Effort-moderated score
Intercept	196.55 (0.43)	196.84 (0.42)	196.63 (0.43)	0.24	0.23	0.24
Fall-to-fall linear slope	6.25 (0.15)	6.67* (0.12)	6.54 (0.14)	0.37	0.27	0.32
Fall-to-fall quadratic slope	-0.35 (0.03)	-0.38 (0.02)	-0.37 (0.03)	0.35	0.31	0.32
Fall-spring change (4th grade)	7.58 (0.17)	7.60 (0.15)	7.63 (0.16)	0.60	0.45	0.54
Fall-spring change linear slope	-0.98 (0.06)	-1.02 (0.05)	-1.05 (0.06)	0.64	0.56	0.57

Note. Standards errors are in parentheses. * signifies difference between unadjusted and adjusted coefficients is significant at the .05 level.

Table 4

Estimates of Racial/ethnic Gaps in Student Growth Accounting for Rapid Guessing

		Fixed effects		
Parameter	Group comparison	No adjustment	Filtering	Effort-moderated score
Intercept (4th grade fall)	White students	198.96 (0.39)	199.39 (0.37)	199.04 (0.39)
	Asian -White Difference	0.40 (0.89)	0.19 (0.91)	0.33 (0.90)
	Black -White Difference	-7.40 (0.41)	-7.70 (0.40)	-7.37 (0.40)
	Hispanic -White Difference	-5.47 (0.52)	-5.95 (0.50)	-5.56 (0.51)
Fall-to-fall linear slope	White students	6.24 (0.16)	6.47 (0.13)	6.47 (0.15)
	Asian -White Difference	0.95 (0.36)	1.09 (0.37)	0.95 (0.35)
	Black -White Difference	-0.06 (0.20)	0.43 (0.18)	0.13 (0.19)
	Hispanic -White Difference	-0.15 (0.26)	0.46 (0.22)	-0.03 (0.24)
Fall-to-fall quadratic slope	White students	-0.35 (0.03)	-0.35 (0.03)	-0.37 (0.03)
	Asian -White Difference	-0.02 (0.07)	-0.05 (0.07)	-0.03 (0.07)
	Black -White Difference	0.00 (0.04)	-0.08 (0.04)	-0.03 (0.04)
	Hispanic -White Difference	0.11 (0.05)	-0.02 (0.05)	0.09 (0.05)
Fall-spring change (4th grade)	White students	7.51 (0.18)	7.46 (0.16)	7.56 (0.17)
	Asian -White Difference	0.70 (0.30)	0.74 (0.30)	0.69 (0.30)
	Black -White Difference	-0.01 (0.19)	0.21 (0.19)	-0.01 (0.18)
	Hispanic -White Difference	0.28 (0.28)	0.39 (0.27)	0.32 (0.27)
Fall-spring change linear slope	White students	-0.95 (0.07)	-0.95 (0.06)	-1.01 (0.06)
	Asian -White Difference	-0.14 (0.13)	-0.20 (0.11)	-0.13 (0.12)
	Black -White Difference	-0.03 (0.09)	-0.14 (0.08)	-0.07 (0.08)
	Hispanic -White Difference	-0.08 (0.11)	-0.16 (0.10)	-0.12 (0.11)

Note. Standards errors are in parentheses. No significant differences were found between unadjusted and adjusted coefficients.

Table 5

Estimates of Summer Learning Loss Accounting for Rapid Guessing

	Fixed effects			ICC		
	No adjustment	Filtering	Effort-moderated score	No adjustment	Filtering	Effort-moderated score
4th Grade Spring Intercept	204.13 (0.42)	204.44 (0.41)	204.26 (0.41)	0.27	0.26	0.27
Fall-to-fall linear slope	5.27 (0.12)	5.65* (0.10)	5.49 (0.11)	0.34	0.25	0.29
Fall-to-fall quadratic slope	-0.35 (0.03)	-0.38 (0.02)	-0.37 (0.03)	0.35	0.32	0.32
4th-5th Summer change	-1.68 (0.12)	-1.31* (0.11)	-1.47 (0.12)	0.55	0.51	0.57
Summer change linear slope	0.28 (0.06)	0.26 (0.05)	0.30 (0.05)	0.42	0.48	0.41

Note. * signifies difference between unadjusted and adjusted coefficients is significant at the .05 level.

Table 6

Effort-moderated Reading Scores and Demographics by Proportion of Rapid Guesses, Spring 2015

Proportion of Rapid Guesses	Effort-moderated Score	N	Proportion of Students Who Are:				
			Female	White	Black	Asian	Hispanic
All	222.322	12,113	0.495	0.643	0.246	0.037	0.074
< .10 (unfiltered)	223.801	10,828	0.512	0.647	0.237	0.041	0.075
> .10 (filtered)	209.760	1,285	0.355	0.607	0.320	0.008	0.065

Note. The <.10 (unfiltered) scores were retained in the "filtering" analyses, while the filtered scores were excluded.

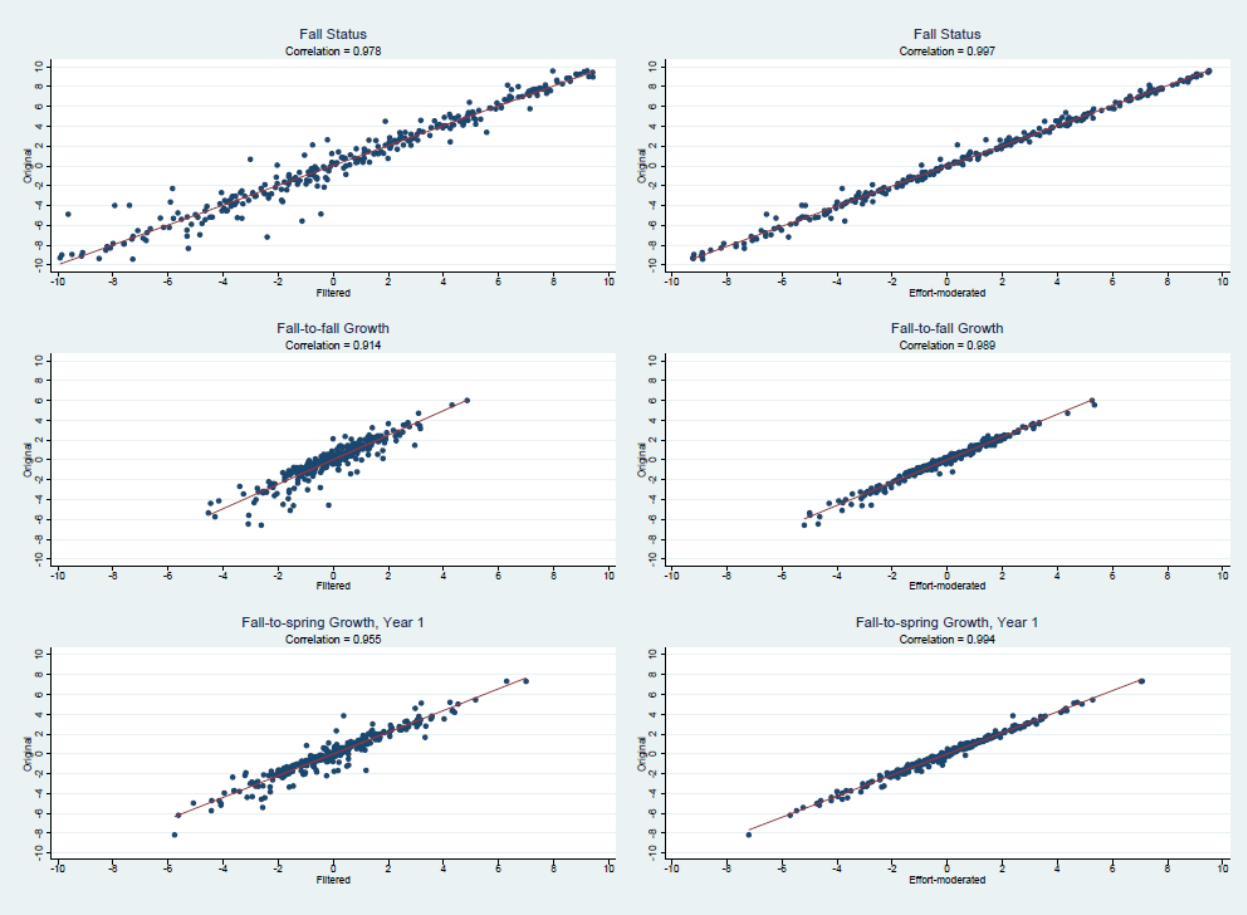

Figure 1. Comparisons of empirical Bayes estimates of school effectiveness.

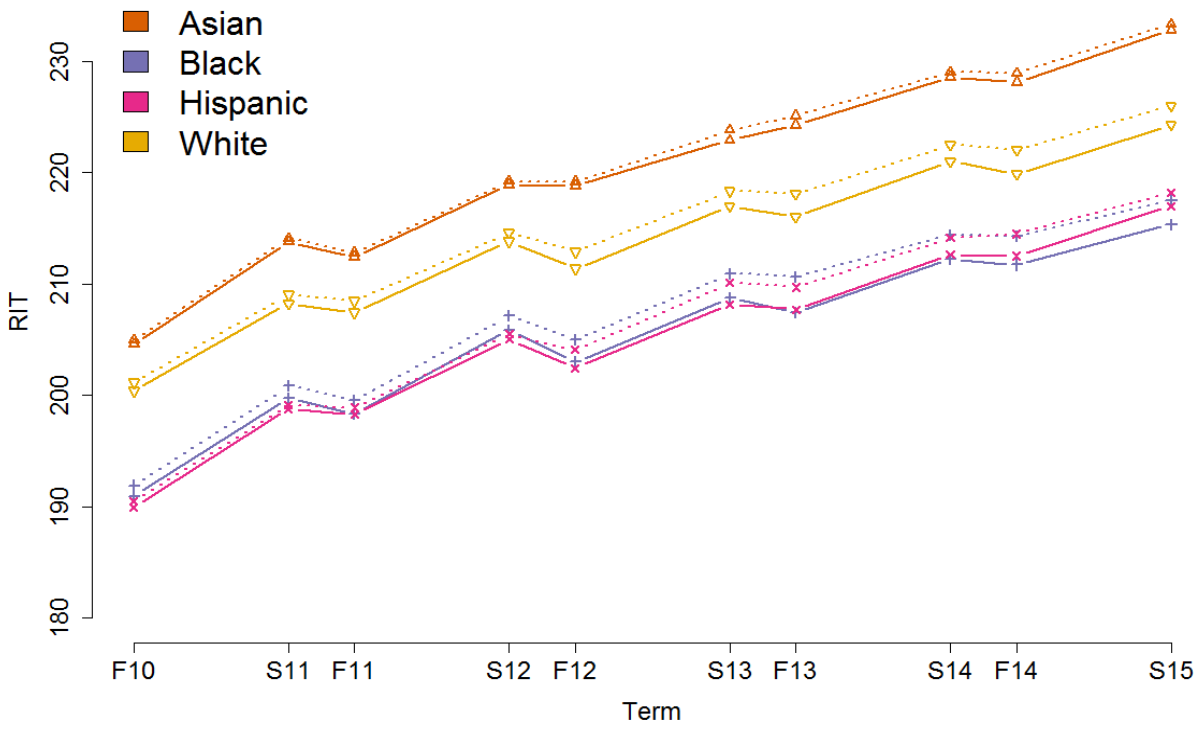

Figure 2. Comparisons of average student trajectories by race/ethnicity and whether filtering was used (where the solid line represents all students, and the dotted line represents only engaged students). Term represents the school term and year (F10 = Fall 2010). The dotted lines represent the trajectory of filtered score means for each race/ethnicity, and the solid lines represent the trajectory of unadjusted score means.

Appendix

Appendix Table 1.
Coding of the Design Matrix for the Compound Polynomial model for Research Questions 1 and 2

Grade/Term	X_0 (Intercept)	X_1 (Fall linear slope)	X_2 (Fall quadratic slope)	X_3 (Fall-spring Int.)	X_4 (Fall-Spring linear slope)
4th Fall	1	0	0	0	0
4th Spring	1	0	0	1	0
5th Fall	1	1	1	0	0
5th Spring	1	1	1	1	1
6th Fall	1	2	4	0	0
6th Spring	1	2	4	1	2
7th Fall	1	3	9	0	0
7th Spring	1	3	9	1	3
8th Fall	1	4	16	0	0
8th Spring	1	4	16	1	4

Appendix Table 2.
Coding of the Design Matrix for the Compound Polynomial model for Research Question 3

Grade/Term	X_0 (Intercept)	X_1 (Spring linear slope)	X_2 (Spring quadratic slope)	X_3 (Spring-Fall Int.)	X_4 (Spring-Fall linear slope)
4th Fall	1	-1	1	1	-1
4th Spring	1	0	0	0	0
5th Fall	1	0	0	1	0
5th Spring	1	1	1	0	0
6th Fall	1	1	1	1	1
6th Spring	1	2	4	0	0
7th Fall	1	2	4	1	2
7th Spring	1	3	9	0	0
8th Fall	1	3	9	1	3
8th Spring	1	4	16	0	0

ABOUT THE COLLABORATIVE FOR STUDENT GROWTH

The Collaborative for Student Growth at NWEA is devoted to transforming education research through advancements in assessment, growth measurement, and the availability of longitudinal data. The work of our researchers spans a range of educational measurement and policy issues including achievement gaps, assessment engagement, social-emotional learning, and innovations in how we measure student learning. Core to our mission is partnering with researchers from universities, think tanks, grant-funding agencies, and other stakeholders to expand the insights drawn from our student growth database—one of the most extensive in the world.

nwea