KINGSBURY RESEARCH CENTER

Student Test-Taking Effort and the Assessment of Student Growth in Evaluating Teacher Effectiveness

Steven L. Wise, Lingling Ma, John Cronin, Robert A. Theaker
Northwest Evaluation Association


Send correspondence to:
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97209
(503) 915-6118
The U.S. Department of Education’s Race to the Top (RTT) program has encouraged states to adopt policies aimed at measuring the effectiveness of individual teachers. To this end, student assessments are used to provide metrics that eventually lead to an effectiveness rating for each individual teacher. School systems are expected to use this information to inform professional development, compensation, promotion, tenure, and dismissal (Baker, Oluwole, & Green, 2013).

Furthermore, the RTT program specifies that the teacher effectiveness rating must be based, in part, on student growth.

From the teachers’ perspective, the teacher effectiveness component of the RTT program means that high-stakes inferences and decisions will now be made about them on the basis of how much academic growth their students exhibit. The use of tests for this purpose is controversial and new threats to validity are introduced by the use of growth measures, because the results are dependent on two tests rather than one. These threats include:

- Noise introduced by individual students who were not motivated to perform at either or both points in time, which can subsequently introduce noise to the growth measurement by introducing random inflated or deflated growth scores.
- Noise unintentionally introduced from poor or unusual testing conditions, which can cause inflated or deflated growth scores across classrooms.
- Noise systematically and intentionally introduced by manipulating testing conditions in an effort to inflate growth results.

These threats are serious and, if not addressed, may undermine the value of any findings about a teacher’s performance that are derived from growth measures. This paper will discuss these problems, overview methods for detecting non-effortful test taking, and describe methods for ameliorating the problems. In addition, the results of the data analyses will illustrate both the problems and our proposed solutions.
Test-Taking Effort and Score Distortion

The goal of educational measurement is to provide test scores that validly indicate what students know and can do. To attain this goal, test givers have traditionally focused on developing tests that are of suitable length and closely tied to the curriculum standards at the test-taking student’s grade level. While a well-constructed test is required if one is to make any meaningful effort to assess student growth, this alone is insufficient. It’s also essential that the tests be consistently administered under conditions that are likely to promote students’ best effort, and that students’ give their best effort on each assessment. If any of these conditions are not adhered to, the accuracy of the growth estimate is compromised.

Because these types of potential threats to test score validity are unrelated to the achievement construct under study, and because they affect some students more than others, they introduce construct-irrelevant variance into the test scores (Haladyna & Downing, 2004). Individual score validity (ISV) addresses the trustworthiness of individual test scores (Hauser & Kingsbury, 2009; Hauser, Kingsbury, & Wise 2008; Kingsbury & Hauser, 2007). Wise, Kingsbury, and Hauser (2009) defined ISV as the extent to which test scores are free from construct-irrelevant factors. They encouraged measurement professionals to identify the most serious construct-irrelevant threats to ISV in a given measurement context, and to develop methods for assessing the degree to which those threats affected particular scores.

An alternative definition for ISV, and one that is useful for this paper, is based on the concept of score distortion:

\[
\text{Distortion}_i = \left[ \frac{\text{What the test score indicates that student } i \text{ knows and can do}}{\text{What student } i \text{ actually knows and can do}} \right] - 1. \tag{1}
\]

Equation 1 is theoretical, as we cannot know for certain the student’s actual level of proficiency.

Conceptually, however, ISV for a student is attained when distortion of his score is equal to zero (or at
least within the limits of measurement error). Positive values of distortion correspond to instances when the test score over-estimates the student’s proficiency level; negative values indicate an under-estimation of proficiency. Positive and negative distortions correspond to different types of threats to ISV.

Positive distortion of a test score - that is, a test score that is substantively higher than the student’s actual achievement - occurs when a student score is higher than his or her actual ability would indicate because of an intended action by the student or teacher. Students can achieve scores that are higher than their actual ability without actual distortion of their test score. If a student has unusual luck guessing on occasion, or refined test-taking skills (that might improve the probability of a correct guess on a multiple choice assessment), the final score may overestimate the student’s actual achievement. This is not a product of distortion, however - it simply reflects either random variance in performance or the limitations of a particular testing methodology.

Positive distortion generally occurs if the student engages in test fraud by copying answers from a more proficient test taker during the test session, by acquiring pre-knowledge of some or all of the items on the exam, or by surreptitiously using technology (like a smartphone) to acquire answers during the test.

Because of the pressures of teacher and school accountability, test givers (such as teachers or principals) might also be motivated to engage in fraudulent testing practices. If specific items on a test are known to test givers beforehand, they could provide that information directly to students before the testing session. During a testing session, they could subtly (or not) point out to students their incorrect answers and possibly indicate the answers they should be giving. After the testing session, test givers could privately alter students’ answer sheets, either by filling in the correct answers to omitted items or by changing incorrect answers to correct.
Regardless of whether it is initiated by the student or the test giver, test fraud is intended to create positive score distortion by producing test scores that overstate what students actually know and can do. Figure 1 illustrates these dual influences on distortion. The solid arrows indicate that students and test givers can directly induce distortion (and diminish ISV) through their actions.

![Diagram](image)

**Figure 1.** The intentional influences of students and test givers on distortion.

Most efforts to diagnose and address test fraud have typically focused on positive distortion because most high-stakes tests reward students or instructors for high performance. That’s why there are high levels of test security around licensure and college entrance examinations.

Negative distortion, in contrast, is attributable to the influence of construct-irrelevant factors leading to scores that underestimate what students know and can do. Any factor that degrades test performance during the test session induces negative distortion. This can include, but is not limited to, student illness or emotional distraction, testing rooms that are too warm/cold or have noise distractions, or unmotivated students who do not give good effort throughout the test. In the past, negative distortion has been a concern because underestimates of performance may trigger negative consequences or sanctions in cases where student performance is actually higher than the reported performance. Negative distortion has not been a focus of test security efforts in the past because
deflated results would not benefit a student or teacher. As stakes around growth measurement have increased, however, negative distortion emerges as an issue.

**Distortion in the context of growth measurement.**

The impact of positive and negative distortion is particularly important when policy moves from a focus on achievement to a focus on growth. Growth measurement is complicated because the assessment must be accurate at two points in time. This can introduce problems that may not have been previously anticipated.

Consider three cases in which test score validity may be questionable. Assume that teachers at Jefferson School are evaluated on the growth their students achieve between terms on a standardized assessment which is administered during fall and spring. Due to unfortunate circumstances Greta, a third grade teacher, is forced to administer the spring assessment on the morning of the school’s field day. Because of their excitement and anticipation, students have a difficult time focusing on the test and their scores are inadvertently but negatively distorted. As a result, Greta’s growth score is lower than it should be and, potentially, her evaluation is negatively affected.

Teachers at Lincoln School are evaluated on the growth their students achieve on the state assessment, which is administered to every grade each spring. Fred, the school’s fifth grade teacher, surreptitiously changes student answer sheets in order to assure his students do well. His cheating leads to a positive distortion of the growth of his classroom. In this case, however, Fred is not the only teacher affected. Randall, the school’s sixth grade teacher will see his growth results deflated the following year when he inherits the students with inflated scores.

Finally, teachers at Highland School are also evaluated on the growth their students receive on a standardized assessment that is administered in the fall and spring. Teachers and administrators at Highland are acutely aware of the importance of the spring assessment and go to great lengths to be
sure students do their best on this test. They provide two weeks of coaching on practice items prior to the spring assessment, have a motivational assembly to kick off testing, and make sure students have refreshments and snacks to stay energized during the day. The fall assessment, which provides the baseline, does not receive the same attention. Students aren’t prepared in advance of the test, test conditions are treated casually, and students are told that it doesn’t really matter how they do on this assessment. As a result, many students aren’t motivated on the fall assessment, but the negative distortion serves to inflate the growth scores for the school’s staff.

These cases illustrate the power that positive and negative distortion can introduce when measuring growth. In Greta’s case, poor testing conditions resulted in negative distortion that systematically deflated a teacher’s result. In Fred’s case, his cheating not only inflated his result through positive distortion, it also deflated the result of the teacher who inherited his students. And in Highland’s case, a possibly inadvertent but systematic negative distortion of fall test results, inflated the growth scores of all teachers in the spring. The Highland case is particularly interesting because past efforts at fraud detection have primarily focused on identifying cases of positive distortion. As many schools move to the use of fall to spring tests to measure growth for high stakes, it becomes important to detect and address cases of negative distortion in fall tests that may inflate a growth result.

**Measuring Test-Taking Effort**

As computer-based tests (CBTs) have become more common, there has been increasing interest in the uses of item response time (which can be collected during a CBT) to improve the measurement of academic achievement. One research theme has focused on the use of response time to investigate examinee engagement. Early research (Bhola, 1994; Schnipke & Scrams 1997, 2002) investigated, using item response time, changes in examinee behavior as time was running out during a speeded, high-stakes test. They found that many examinees switch strategies from trying to work out the answers to
items (termed solution behavior) to rapidly entering answers to remaining items in hopes of guessing some of them correct (termed rapid-guessing behavior).

Response Time Effort

Wise and Kong (2005) observed that rapid-guessing behavior also occurred during unspeeded low-stakes CBTs. They showed that in this context rapid-guessing behavior indicates instances when a student was not expending effort toward attaining a good score\(^1\). Wise and Kong developed a measure of test-taking effort, termed response time effort (RTE), which equals the proportion of a student’s responses that were solution behaviors. An RTE value of 1.0 indicates that a student exhibited only solution behavior, a value of .90 indicates 10% of the item responses were rapid guesses, and so on.

Identification of rapid-guessing behavior requires that a time threshold be established for each item. This permits each item response to be classified as either a solution behavior or a rapid guess. Two basic principles are followed in establishing time thresholds. First, we want to identify as many instances of non-effortful item responses as possible. Second, we want to avoid classifying effortful responses as non-effortful. There is a tension between the two principles such that the first encourages us to choose a longer threshold, while the second encourages us to choose a shorter one. Thresholds are chosen to balance the two principles, with the second principle being of higher priority. Good discussions of item threshold identification methods and issues are found in Ma, Wise, Thum, and Kingsbury (2011) and Wise and Ma (2012).

The identification of rapid-guessing behavior is important because it indicates the presence of item responses that exert a negative bias on a proficiency estimate. This is due to rapid guesses being correct at a rate that is usually markedly lower than what would have been the case had the student exhibited solution behavior. Therefore, the more rapid guesses that occur during a test event, the more

\(^1\) Regardless of the stakes of the test or whether or not the test is speeded, a rapid guess indicates essentially the same thing—that the examinee was not engaged in solution behavior.
negative distortion is likely present in a test score. Thus, the presence of rapid-guessing behavior provides useful evidence that a score that has low ISV.

When a computerized adaptive test (CAT) is used, an additional indicator of low student effort is provided by item response accuracy. However, unlike rapid guessing—which can be assessed for each item response—accuracy must be considered across a set of items. The CAT algorithm is designed to select and administer items that a student under solution behavior has about a 50% chance of getting correct. Under rapid-guessing behavior, in contrast, items will be correct at a rate consistent with random guessing. For multiple-choice items with five response options, a student would be expected to provide a correct response about 20% of the time. Hence, because responses to items administered in a CAT will have consistent, differential accuracy rates under solution behavior and rapid-guessing behavior, the accuracy of a student’s responses to a set of items can be evaluated as to whether it appears to reflect solution behavior or rapid-guessing behavior. For example, if during the last half of a test a student passed only 22% of his items on a well-designed CAT with sufficient number of well-targeted items in the item bank, we might decide that he was not giving effort during that portion of the test and conclude that his score should be considered as reflecting low ISV.

In our research with NWEA’s Measures of Academic Progress® (MAP®), which is used to measure the academic growth of primary and secondary school students, a set of five flagging criteria for identifying test events have been developed that yield scores with low ISV. These heuristic criteria, which are described in the Appendix, are based on a combination of RTE and response accuracy, either singly or in combination. The criteria have been shown to identify many instances of non-effortful student behavior (Wise & Ma, 2012), and can be used to evaluate the degree of test-taking effort exhibited at different time periods and across various content domains.
Test Accuracy, Test Duration and Test Duration Differences

Another approach to evaluating student effort was introduced by Cronin et al. (2005) in a study which analyzed the relationship impact that answering accuracy, test duration, and changes in test duration may have on student growth scores generated from a CAT (once again MAP). Their regression model found moderate correlations ($r=.53$, $r^2=.29$) between changes in test duration on an untimed test and the percentage of correct answers on student growth scores, suggesting that drops in test duration and lower than expected proportions of correct answers produced lower growth scores. From this they tested the model by flagging students whose reduction in test duration was among the bottom decile, and tests that were in the bottom decile for percent correct. Using these criteria, they were able to detect about half of all students with growth of one standard error of less below their growth norm.

Score Distortion

Wise, Ma, and Theaker (2012) described the concept of score distortion, which they defined as the difference between what a test score indicates about a student’s level of proficiency and the student’s actual level of proficiency. Positive values of distortion correspond to instances when the test score over-estimates proficiency; negative values indicate an under-estimation of proficiency. Non-effortful test taking behavior induces a negative distortion on a student’s score. However, when considering growth (defined as $Time_2$ score - $Time_1$ score), distortion is more complicated. If non-effort occurred at $Time_1$, growth will be positively distorted because growth will be spuriously inflated. Conversely, if non-effort occurred at $Time_2$, growth will be negatively distorted.

Wise et al. (2012) investigated student effort on a computerized adaptive interim assessment in math and reading (grades 3-8) for a set of U.S. charter schools. These schools have used student growth (spring score minus fall score) as part of their teacher evaluation system for a number of years. A set of flagging criteria were used to identify test events whose scores were deemed to be invalid due to
student non-effort. These criteria were based on either response time effort (Wise & Kong, 2005) or response accuracy.

The primary finding of the Wise et al. (2012) study was a clear indication of differential mean between the fall and the spring test administrations. For nearly all of the 39 schools, the percentage of invalid scores in the fall was higher than in the spring. While the ratio of fall-to-spring percentages averaged about 2:1, there was considerable variation across schools and in some cases it was much higher, reaching a maximum ratio of around 17:1. The greater prevalence of non-effortful behavior in the fall had the effect of inducing positive distortion on growth scores, with the result that that mean growth was inflated by the differential effort. Moreover, the schools with the highest differential effort were found to be among the highest in mean student growth. These findings suggest that differential effort—both fall-to-spring and across schools—could confound interpretations of student growth. High mean student growth in a school could be explained by either high actual growth, or high differential effort, or both.

**A Case Study of Teacher Evaluation and Test-Taking Effort**

It is one thing to describe a scenario in which test givers could manipulate student effort to inflate growth scores, and another to demonstrate that it represents a real problem. That is, if a problem has little chance of actually occurring, its solution may have little practical value. To evaluate this issue, we studied data from a context in which student growth has been used as part of a teacher evaluation process for a period of time. These data should provide a basis to assess the degree to which Time 1 versus Time 2 effort discrepancies are present.

**Data Source**

This case study focused on a large charter school management organization that operates charter schools in multiple U.S. states. They have used MAP test results as part of their teacher
evaluation system for a number of years. The evaluation system for teachers includes four components, with a substantial portion being an evaluation of student achievement based on fall-to-spring growth results for MAP. Teacher compensation and retention are conditioned on the results of these evaluations.

The implementation of MAP as part of the teacher evaluation system is longstanding. As the charter school organization has increased its capacity around data use, it has refined and implemented more sophisticated approaches to measuring teacher effectiveness from test data. From the teachers’ standpoint, however, one aspect has remained constant: a sizable portion of their evaluation is based on the amount of student academic growth in MAP that is observed between the beginning of an academic year (i.e., fall) and the end of that academic year (i.e., spring). Hence, the evaluation system is consistent with the scenario described earlier in which there could be an incentive for test givers to try to depress fall scores with the goal of inflating growth scores. Occurrences of this could be identified by markedly more non-effortful test-taking behavior being observed in the fall as compared to the spring.

For purposes of this study, two distinct methodologies were pursued to evaluate non-effortful test taking behavior within classrooms. Method 1 evaluated fall 2010 and spring 2011 tests and employed both RTE and the percentage of test events whose scores were classified as invalid using the five effort criteria described in the Appendix, the criteria applied in the prior Wise et al (2012) study. Method 2 applied a variant of the methodology used in Cronin et al (2005) study. This approach analyzed data from three terms, spring 2010, fall 2010, and spring 2011 and applied three criteria low percent correct, low test time, and large change in test duration in an attempt to identify classrooms in which score distortion may have been evident. The benchmarks for each of the criteria were generated from an analysis of all spring 2011 test results. From this analysis cut scores were established at the 10th percentile for percent correct, test duration, and changes in test duration. Tests that fell at or below the 10th percentile criteria were flagged as possibly invalid.
In both cases, the data analyses focused on MAP scores in math and reading in grades 3-8 from 61 charter schools across 5 states. MAP tests are untimed, interim computerized adaptive tests (CATs), with the tests in math being generally 50 items in length, while those in reading are generally 40 items in length. MAP achievement estimates are expressed as scale (RIT) scores on a common scale that allows growth to be assessed as students are tested at different times. The standard errors of the fall-spring scores in math are typically range between 4.25 and 4.50 points on this scale. All students included in this data analysis were tested over three terms spring of 2010, fall of 2010, and spring of 2011. Tests were administered in reading, language usage, and mathematics. Classrooms were included in the analysis if there were 10 or more students who had tested in each of the studied terms.

**Data Analysis and Results**

**Method 1**

Table 1 shows descriptive statistics for response time effort (RTE), the percentage of invalid scores, and growth scores. Mean RTE values indicate that non-effortful behavior was more prevalent in the fall than in the spring across all three content domains. It is consistent with the results from prior data examined at the school level (Wise, Ma, & Theaker, 2012). The percentage of invalid scores were measurably different between two terms. In the fall, more test events were flagged as invalid events by the ISV criteria (1.4 to 3.2 percent more).
Table 1. Descriptive Statistics for RTE, Percentage of Invalid Scores, and Fall-to-Spring growth

<table>
<thead>
<tr>
<th>Content Area</th>
<th>N</th>
<th>Fall Mean RTE</th>
<th>Fall Percentage of Invalid scores</th>
<th>Spring Mean RTE</th>
<th>Spring Percentage of Invalid scores</th>
<th>RIT Growth Mean</th>
<th>RIT Growth SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>10014</td>
<td>0.992</td>
<td>3.0</td>
<td>0.996</td>
<td>1.6</td>
<td>9.16</td>
<td>7.92</td>
</tr>
<tr>
<td>Reading</td>
<td>22377</td>
<td>0.985</td>
<td>8.2</td>
<td>0.991</td>
<td>5.0</td>
<td>8.22</td>
<td>9.26</td>
</tr>
<tr>
<td>Language</td>
<td>22407</td>
<td>0.989</td>
<td>5.7</td>
<td>0.994</td>
<td>2.9</td>
<td>7.84</td>
<td>8.30</td>
</tr>
</tbody>
</table>

Additionally, as found in prior studies, examinee efforts varied across different content domains. In general, examinees were more likely to engage in effortful behavior in mathematics than in reading or language tests (Wise, Ma, Kingsbury, & Hauser, 2010). The average RTE values in math tests were greater than 0.99 in both terms, whereas the average RTE values dropped below 0.99 in reading or language tests in the fall term.

Scatterplots of the spring mean versus fall mean RTE values at the class level are displayed in Figures 2a, 2b, and 2c. Note that data points were not tightly clustered around the reference line in those scatterplots. Nor did most of them fall above the reference line, which is different from the pattern reported in Steve, Ma, & Theaker (2012) when the school-level data were examined. Figures 2a, 2b, and 2c suggest that greater variation existed at the class level. Some classes showed considerably lower RTE in the fall term, especially in reading (e.g. 0.89) and language classes (e.g. 0.89).

No strong linear relationships were observed at the class level, and measurable differences in mean RTE values seemed to exist between two terms. Classes in the upper left quadrant demonstrated low mean RTE values in the fall but higher values in the spring. For example, there were a few math classes with fall RTE as low as 0.94 and spring RTE increased to 0.99. Some reading and language classes, fall RTE were as low as 0.89 and fall RTE increased to 0.99.
The prior study (Wise, Ma, & Theaker, 2012) showed that lower mean RTE were found in the fall for almost every school. However, among classes lower mean RTE values were found in many, but not all, in the fall. In general, as indicated in both scatterplots and histograms, examinees were more likely to show effortful behavior in the spring, despite some outliers following a different pattern. Compared to mathematics tests, much greater variations existed in reading and language tests and the average RTE values were considerably lower in those two tests.

**Figure 2a.** Fall RTE versus Spring RTE in Mathematics
Figure 2b. Spring RTE versus Fall RTE in Reading

Figure 2c. Spring RTE versus Fall RTE in Language

Histograms of mean RTE differences between two terms are displayed in Figure 3. Histograms show the distributions of the RTE differences between the fall and spring terms across three content domains. In 81% of mathematics classes, differences in RTE values were within [-0.01 to 0.01] and effortful behaviors in those classes were not measurably different between two terms. Only 64% of
reading classes and 69% of language classes showed similar RTE values between two terms. Examinees were more likely to demonstrate similar levels of test engagement between fall and spring in a mathematics test, than in a reading or language test.

RTE values were more likely to be higher in the spring than in the fall, regardless of content domain. In other words, non-effortful behavior occurred more frequently in the fall than in the spring. The pattern of differences in average RTE values was somewhat similar across the three content domains. The average RTE values were higher in only 2% of mathematics classes in the fall, whereas 17% were higher in the spring. In the meantime, 7% of reading classes had higher RTE values in the fall, whereas 28% of them had higher values in the spring. As with the reading test, 5% of language classes had higher values in the fall but 26% had higher values in the spring. Histograms in Figure 3 serve as a clear indication that the examinee effort was considerably higher in the spring.

**Figure 3.** Distributions of Spring-Fall RTE Differences across Three Content Domains

The percentage of invalid test events from mathematics, reading, and language are displayed in Figure 4a, 4b, and 4c. They show that some reading and language usage classes had as many as 30T
invalid tests in the fall. Results further show that invalid test events were more likely to occur in the fall than in the spring across all content domains.

**Figure 4 a.** Percentages of Invalid Test Events in Mathematics at the Class Level

![Percentage of Invalid Test Events in Math](image)

**Figure 4 b.** Percentages of Invalid Test Events in Reading at the Class Level

![Percentage of Invalid Test Events in Reading](image)

**Figure 4 c.** Percentages of Invalid Test Events in Language at the Class Level

![Percentage of Invalid Test Events in Language](image)
Consistent with the findings from those scatterplots, those distributions in Figure 5 indicated that there were more invalid test events in the fall than in spring. The proportion of invalid test events during the spring term decreased in 30% of mathematics classes, 43% of reading classes, and 37% of language classes. As a result, a smaller proportion of classes had more invalid test events in the spring, 13%, 14%, and 12% for mathematics, reading, and language respectively.
The relationships between the fall-spring decrease in invalid test events and fall-spring growth are displayed in Figures 6a, 6b and 6c for each content domain (decreases in the proportion of invalid test events are expressed in positive numbers). Classes which experienced a greater decrease in invalid test events exhibited more growth. These differences are potentially quite large. For example, a mathematics class which exhibited a 20% decline in invalid test events would, by the regression, show 2.80 RIT additional growth, an amount nearly equivalent to the test’s standard error of measure. A 20% decline in invalid tests would project to even larger additional gains in reading and language usage.
**Figure 6a.** Fall-Spring Mean Growth in Math versus Decrease in Invalid Test Events at Class Level

Regression Equation: Fall-Spring Growth in Math = 8.90 + 0.14 * Percent Decrease in Invalid Test Events.

**Figure 6b.** Fall-Spring Mean Growth in Reading versus Decrease in Invalid Test Events at Class Level

Regression Equation: Fall-Spring Growth in Reading = 7.66 + 0.16 * Percent Decrease in Invalid Test Events
Figure 6c. Fall-Spring Mean Growth in Language versus Decrease in Invalid Test Events at Class Level

Regression Equation: Fall-Spring Growth in Language = 7.07 + 0.23 * Percent Decrease in Invalid

Finally Figures 7a, 7b and 7c show to what degree growth scores were affected when invalid test events were deleted. The effect of removing the records was meaningful but small, depending on subject a 20% decline in invalid scores would yield a roughly one point decline in growth score. This creates an interesting conundrum. While classrooms with higher rates of invalid tests clearly showed higher growth, removing the invalid tests did not seem to entirely eliminate the effect. In other words, the classrooms with high levels of invalid tests would seem to continue to have higher levels of growth even after invalid records were removed.
Figure 7a. Change in Mean Growth Scores in Math versus Decrease in Invalid Test Events

Regression Equation: Change in Mean Growth in Math = -0.05 – 0.04 * Percent Decrease in Invalid Test Events

Figure 7b. Change in Mean Growth Scores in Reading versus Decrease in Invalid Test Events

Regression Equation: Change in Mean Growth in Reading = -0.06 – 0.04 * Percent Decrease in Invalid Test Events
Regression Equation: Change in Mean Growth in Language = -0.05 – 0.05 * Percent Decrease in Invalid Test Events

Method 2

Method 2 re-analyzed the same dataset in an effort to identify classrooms that may have inflated fall-to-spring growth scores because of unusual changes on several indicators. Data from three test terms were analyzed, spring of 2010, fall of 2010, and spring of 2011 and students were included only if they had data from all three terms. For purposes of this analysis classrooms were combined when assigned to a single teacher and teachers were included in the analysis if their classrooms had a total of 30 or more tests across all subjects. The following indicators were used to identify tests with possible validity issues related to student effort.
- Criterion 1 - The student had an accuracy rate at the 10th percentile or below for the fall 2010 term.
- Criterion 2 - The student had a test duration equivalent to the 10th percentile or below for the fall 2010 term.
- Criterion 3 - The student had a change in duration (shorter) between the spring 2010 and fall 2010 terms that was at or above the 90th percentile.

Norms for these indicators were established through an analysis of all MAP test results across all partners served by NWEA during the spring 2011, fall 2011, and spring 2012 terms. Each subject and grade level was normed separately on each of the three criteria. This created a normalized set of indicators that would permit collapsing data across grades levels and subjects.

These criteria were applied to the individual student records in the dataset. The dataset was then aggregated by teacher for all subjects and classes taught. Based on those aggregations, we identified cases in which 20% or more of the teacher’s students violated the criteria, which would have been double the expected rate. There were 674 teachers evaluated in the sample.

Settings in which 20% or more of the students violated one or more of the criteria offer evidence that students may not have been fully engaged with the MAP assessment. This does not by itself offer compelling evidence that teachers’ intentionally or unintentionally engaged in gaming the system to inflate their students’ growth, although gaming is one possible explanation. For purposes of this portion of the study, our interest was in determining the impact that evidence of poor student effort may have on the growth results for a sample of teachers.

---

2 These criteria were established by extracting growth results for all students in the NWEA Growth Research Database who had test results for the Spring 2011, Fall 2011 and Spring 2012 terms. Thresholds were separately established for each subject and grade level, which makes it reasonable to combine subjects and grade levels for purposes of evaluating effort.

3 On adaptive tests, nearly all students answer close to 50% of the items correctly. Significant deviation from this pattern can be a sign that students aren’t giving full effort on the test.
Table 1 shows the results for Criterion 1. Teachers who had higher concentrations of students with unusually low test accuracy showed significantly larger drops in score between spring 2010 and fall 2010 than the remaining teachers. This group also showed significantly higher rates of growth between the deflated term, fall 2010, and the following term, spring 2011. The effect size associated with the changes was moderate for the spring 2010 – fall 2010 term, and large for the fall 2010 – spring 2011 term.

Figure 1 depicts the differences in growth over the three terms for the two groups. It shows that the teachers with unusually large concentrations of low accuracy rates show greater drops in score over the summer but, interestingly, much larger score gains in the spring. The underlying reasons for this are not entirely clear. One possibility that students who engaged in lots of random guessing might achieve larger score gains when they are engaged with a subsequent test. Another possibility is that these teachers paid more attention to testing conditions and offered more encouragement than other teachers during the subsequent term. In any case, the overall growth shown by classrooms with high concentrations of students with unusually low accuracy rates was higher than growth for the other teachers.
Table 2 - Results for teacher on Criterion 1 – 20% or more students with 10th percentile or below test accuracy

<table>
<thead>
<tr>
<th></th>
<th>Classroom with &gt;20% violating criteria</th>
<th>Classrooms with &lt;20% violating criteria</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>57</td>
<td>617</td>
<td>674</td>
</tr>
<tr>
<td>Mean Classroom Growth Spring 2010 - Fall 2010</td>
<td>-2.62</td>
<td>-1.36</td>
<td>-1.47</td>
</tr>
<tr>
<td>Percentile Rank Associated with Mean Classroom Score</td>
<td>24%</td>
<td>52%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td>1.91</td>
</tr>
<tr>
<td>Effect Size Difference 2 groups</td>
<td></td>
<td></td>
<td>0.66</td>
</tr>
<tr>
<td>Mean Classroom Growth Fall 2010 - Spring 2011</td>
<td>13.03</td>
<td>8.01</td>
<td>8.44</td>
</tr>
<tr>
<td>Percentile Rank Associated with Mean Classroom Score</td>
<td>90%</td>
<td>51%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td>3.44</td>
</tr>
<tr>
<td>Effect Size Difference 2 groups</td>
<td></td>
<td></td>
<td>-1.46</td>
</tr>
</tbody>
</table>

Figure 8 – Difference in Spring 2010 – Fall 2010 – Spring 2011 growth
It is now very common to use value-added measures to evaluate teacher performance, and these generally depict teacher effectiveness by producing z-scores that are used to rank teachers relative to each other. Because these rankings may determine a teacher’s ultimate evaluation rating, the impact that student effort may potentially have on them is relevant. In terms of Criterion 1, the mean (negative) growth score associated with teachers with high rates of Criterion 1 violations would have placed them at the 24th percentile of growth, while the mean growth score of the remaining teachers was equivalent to the 52nd percentile. Because of the larger negative gain score, teachers started the fall-spring growth period with lower fall scores than other teachers and lower scores than would have been expected. Despite their lower scores, students of teachers with high levels of Criterion 1 violations, showed considerably larger gains in fall 2010 – spring 2011 than had been expected. The mean fall - spring gains for this group were equivalent to the 90th percentile of the entire sample, while the remaining teachers scored mean gains equivalent to the 51st percentile. The difference in fall to spring gain, which was equivalent to more than 5 scale score points, offset the fall drop (of about 1.3 points) by a large margin. Thus the additional gains are not entirely explained by any advantage that the teachers with high rates of Criterion 1 violations may have obtained from an artificially depressed fall score.

Criterion 2 was applied to identify classrooms in which larger than expected concentrations of teachers had students with low test durations. Of the 674 teachers in the sample, 149 had 20% or more of their students violate this criterion. The impact of this on scores was similar, but less pronounced than the Criterion 1. Teachers with high concentrations of students with Criterion 2 violations had larger negative gains between spring 2010 and fall 2010 than other teachers, leading to a small effect size difference of .26. Once again their fall 2010 to spring 2011 gains were considerably larger, showing a medium effect size difference, and this difference more than offset the original drop in score.
The impact of these differences on teacher rankings was smaller than was the case with Criterion 1. Teachers with high concentrations of students in violation of Criterion 2 showed gain scores at the 39th percentile for the spring 2010 – fall 2010 term, while other teachers showed mean gain scores at the 52nd percentile. For the fall 2010 – spring 2011 term, gain scores for the teachers with high concentrations of students with low fall 2010 test durations were larger, and their mean growth was equivalent to the 71st percentile, compared with mean growth at the 49th percentile for all other teachers in the sample.

Table 3 – Results for teachers on Criterion 2 – 20% or more students with 10th percentile or below test duration

<table>
<thead>
<tr>
<th></th>
<th>Classroom with &gt;20% violating criteria</th>
<th>Classrooms with &lt;20% violating criteria</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>149</td>
<td>525</td>
<td>674</td>
</tr>
<tr>
<td>Mean Classroom Growth Spring 2010 - Fall 2010</td>
<td>-1.85</td>
<td>-1.36</td>
<td>-1.47</td>
</tr>
<tr>
<td>Percentile Rank Associated with Mean Classroom Score</td>
<td>39%</td>
<td>52%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td>1.91</td>
</tr>
<tr>
<td>Effect Size Difference 2 groups</td>
<td></td>
<td></td>
<td>0.26</td>
</tr>
<tr>
<td>Mean Classroom Growth Fall 2010 - Spring 2011</td>
<td>10.33</td>
<td>7.90</td>
<td>8.44</td>
</tr>
<tr>
<td>Percentile Rank Associated with Mean Classroom Score</td>
<td>71%</td>
<td>49%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td>3.44</td>
</tr>
<tr>
<td>Effect Size Difference 2 groups</td>
<td></td>
<td></td>
<td>-0.71</td>
</tr>
</tbody>
</table>

Criterion 3 was applied to identify cases in which students showed 90th percentile or above declines in test duration between spring 2010 and fall 2010. The decrease in fall test duration would be an indication that students showed less effort on that test than the one taken in the prior spring. The
results here were consistent with the other cases. Teachers who had 20% or more of their students violate this criterion showed spring 2010 – fall 2010 drops that were significantly greater than the remaining teachers. The effect size of this difference was moderate. The difference between the two groups fall 2010 – spring 2011 growth scores were not as large, but once again teachers with larger concentrations of students violating the criteria showed larger growth and a small effect size difference.

There were once again differences in teacher rankings between the two groups. Teachers with higher concentrations of students in violation of Criterion 3 showed mean negative gain scores between the spring 2010 and fall 2010 terms equivalent to the 33rd percentile, while the remaining teachers showed mean gains averaging at the 58th percentile. Teachers with high levels of Criterion 2 showed gain scores at the 39th percentile for the spring 2010 – fall 2010 term, while other teachers showed mean gain scores at the 52nd percentile. For the fall 2010 – spring 2011 term, gain scores for the teachers with high concentrations of students with low fall 2010 test durations were once again larger, their mean growth was equivalent to the 71st percentile, compared with mean growth at the 49th percentile for all other teachers in the sample. In this case, however, the additional gains merely offset the lower fall scores.
Table 4 – Results for teachers on Criterion 3 – 20% or more students with 90th percentile or above difference in Spring 2010 – Fall 2010 test duration

<table>
<thead>
<tr>
<th></th>
<th>Classroom with &gt;20% violating criteria</th>
<th>Classrooms with &lt;20% violating criteria</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>254</td>
<td>420</td>
<td>674</td>
</tr>
<tr>
<td>Mean Classroom Growth Spring 2010 - Fall 2010</td>
<td>-2.14</td>
<td>-1.06</td>
<td>-1.47</td>
</tr>
<tr>
<td>Percentile Rank Associated with Mean Classroom Score</td>
<td>33%</td>
<td>58%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td>1.91</td>
</tr>
<tr>
<td>Effect Size Difference 2 groups</td>
<td></td>
<td></td>
<td>0.56</td>
</tr>
<tr>
<td>Mean Classroom Growth Fall 2010 - Spring 2011</td>
<td>9.24</td>
<td>7.95</td>
<td>8.44</td>
</tr>
<tr>
<td>Percentile Rank Associated with Mean Classroom Score</td>
<td>62%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td>3.44</td>
</tr>
<tr>
<td>Effect Size Difference 2 groups</td>
<td></td>
<td></td>
<td>-0.38</td>
</tr>
</tbody>
</table>

Finally, results were analyzed relative to a fourth criterion. Teachers were identified if the total number of criteria violated by their students constituted 60% or more of their student count. This was done to see if applying criteria in combination produced stronger differences in growth than were found applying them separately.

A total of 103 teachers were found with violations that constituted 60% or more of the student count. Their classrooms showed significantly larger declines in spring 2010 – fall 2010 scores than the remaining teachers and the effect size difference was large. Once again fall 2010 to spring 2011 gains for the group with higher concentrations of violations were also larger than the gains for the remaining
These differences had a substantive effect on how teachers were ranked. Teachers in the group with higher concentrations of violations had negative gain scores at the 22nd percentile between spring 2010 and fall 2010, compared with the 55th percentile for the remaining teachers. The fall 2010 to spring 2011 difference mean gain was at the 80th percentile for the group with higher concentrations of violations and the 49th percentile for the remaining teachers. Once again the fall to spring gains for the group with high rates of Criterion 4 violations more than offset the additional drop in fall scores.

Table 5 – Results for classrooms on Criterion 4 – Criteria violations totaled 60% of the class count or more

<table>
<thead>
<tr>
<th></th>
<th>Classrooms with violations &gt;= 60% of class count</th>
<th>Classrooms with violations &lt; 60% of class count</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>103</td>
<td>564</td>
<td>667</td>
</tr>
<tr>
<td>Mean Classroom Growth Spring 2010 - Fall 2010</td>
<td>-2.81</td>
<td>-1.25</td>
<td>-1.49</td>
</tr>
<tr>
<td>Percentile Rank Associated with Mean Classroom Score</td>
<td>22%</td>
<td>55%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td>1.91</td>
</tr>
<tr>
<td>Effect Size Difference 2 groups</td>
<td></td>
<td></td>
<td>0.82</td>
</tr>
<tr>
<td>Mean Classroom Growth Fall 2010 - Spring 2011</td>
<td>11.38</td>
<td>7.95</td>
<td>8.44</td>
</tr>
<tr>
<td>Percentile Rank Associated with Mean Classroom Score</td>
<td>80%</td>
<td>49%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td>3.44</td>
</tr>
<tr>
<td>Effect Size Difference 2 groups</td>
<td></td>
<td></td>
<td>-1.00</td>
</tr>
</tbody>
</table>

Discussion

Student growth is now frequently included in evaluations of teachers. While states commonly use spring to spring results for teacher evaluation, most states that have adapted Race to the Top programs allow school systems to choose a local assessment as a secondary measure. Many school systems have
chosen to use MAP as that measure, and fall to spring results are generally used for that portion of the evaluation. The results from both Method 1 and Method 2 indicate that student effort can have impact on growth results, and that there is evidence that fall results for many teachers or classrooms may be artificially deflated because of low effort on the part of a significant number of students in these classes. Results from Method 1 would indicate that growth scores in some classrooms may be inflated by as much as one standard error or more (3 RIT), by deflated fall results due to lack of effort. Results from Method 2 suggest that low levels of effort on fall tests may have a substantive effect on how value-added results ultimately rank teachers. Taken together, both analyses would suggest that student effort, which is not adequately addressed by most of the value-added models, can have a substantive and systematic impact on growth results.

One impact is obvious, if fall-to-spring results are used to evaluate teachers, teachers have an incentive to game the system by facilitating conditions in which many students may not give their best effort on a fall assessment. We don’t assert that the methods here provide a strategy for identifying and establishing causation in cases where teachers may be manipulating conditions as we don’t believe an appropriate level of certainty can be obtained from a remote evaluation of the data. Further, in many cases, poor testing conditions may not be the responsibility of the teacher. For example, if the school principal requires that a teacher administer an untimed test in the 30 minutes before lunch, we couldn’t fault the teacher if students weren’t highly motivated to give enough time to the test to do well. Nevertheless, bringing transparency to conditions in which there are low levels of effort shown by students can be very useful. Transparency alone may deter gaming behavior, and having data about the quality of testing conditions can help school systems improve their test administration practices. One possible solution is to use spring-to-spring assessment results to measure growth since each teacher’s post-test becomes another teacher’s pre-test. This is not a fully adequate solution. Spring-to-spring testing leaves as much as five months of the calendar year outside the control of the teacher who will
ultimately be responsible for a classroom’s results and introduces noise to the analysis of growth which can be difficult to address. Approaches which use a minimum of three terms (spring – fall – spring) or preferably multiple years of data seem to offer more promise.

The other impact is less obvious but equally important. Just as fall-to-spring growth results may be overstated when fall scores are deflated, a teacher’s growth results can also be negatively affected if test conditions are inadvertently degraded in spring, or if students systematically engage in non-effortful behavior on their own. While there is an argument to be made that teachers are responsible for motivating their students to perform on tests, and that poor motivation may be a reflection of poor teaching, we believe it is important to determine whether a poor growth result is likely the impact of some problem in instruction, or a problem related to lack of effort. After all, the two problems require different solutions. Once again, making the level of student effort transparent can protect teachers who may see results that are unfairly deflated because they were obtained in poor testing conditions that were out of their control. Further it can remove noise from value-added analysis that can help improve the reliability of results obtained.

We were surprised and a little puzzled that teachers with low rates of student effort on the fall assessment in the Method 2 analysis showed fall-to-spring growth that was generally much greater than the growth obtained from classrooms that had more even effort. Indeed, their growth was greater than any artificial deflation that might have been present in the fall results. The reasons for this aren’t clear and should be the subject of future study. One possibility, one that we don’t consider highly likely, is that teachers who were gaming the system in fall, would also naturally game the system in spring. We considered that possibility, but an analysis of average test times for teachers on Criterion 4 would not suggest systematic gaming is taking place. Table 6 does show, as expected, that classrooms with high levels of Criterion 4 violations showed average test durations that were shorter in fall. Their tests in spring were also longer, average just over 10 additional minutes in duration. But the tests for the other
group were also longer by nearly seven minutes, and the test durations for the teachers with high levels of Criterion 4 violations still had shorter mean test times in spring 2011 than the comparison group of teachers. This would not indicate any systematic attempt to manipulate spring 2011 conditions to inflate growth scores.

Table 6 – Mean test durations for teachers, disaggregated by status relative to Criterion 4.

<table>
<thead>
<tr>
<th></th>
<th>Classrooms with violations &gt;= 60% of class count</th>
<th>Classrooms with violations &lt; 60% of class count</th>
</tr>
</thead>
<tbody>
<tr>
<td>average spring 2010 test duration</td>
<td>45.82</td>
<td>45.41</td>
</tr>
<tr>
<td>average fall 2010 test duration</td>
<td>37.30</td>
<td>45.41</td>
</tr>
<tr>
<td>average spring 2011 test duration</td>
<td>47.74</td>
<td>52.06</td>
</tr>
</tbody>
</table>

In conclusion, there is evidence that student effort has a significant influence on growth results, and that unusually low levels of student effort may have a significant impact student growth within a classroom or among students assigned to a teacher. These impacts can have an important impact on value-added results, causing meaningful changes in the way teachers are ranked relative to each other and, in school systems where test results are used as part of the evaluation, ultimately change the evaluation ratings of significant numbers of staff. Additional research is required to more fully understand the impact this may have on teacher evaluation and efforts should be undertaken to report metrics that make the level of student effort on tests more transparent. Ultimately, if student growth results are used in the evaluation of teachers, research must be undertaken to introduce some control for the impact low effort may have on results.
References


Bhola, D. S. (1994). *An investigation to determine whether an algorithm based on response latencies and number of words can be used in a prescribed manner to reduce measurement error*. Unpublished doctoral dissertation, University of Nebraska-Lincoln.


Appendix

Effort Flagging Criteria for Identifying Low ISV Scores

When multiple-choice items are being administered, there are two types of behaviors that indicate a student has become disengaged from his test and is exhibiting non-effortful test-taking behavior. First, he may respond to items very rapidly (i.e., faster than it should take him to read the item and thoughtfully consider the problem it poses). Second, his answers may be correct at a rate that is consistent with what would be expected by chance through random guessing. Rapid responses typically exhibit chance-level accuracy. Additionally, however, chance-level accuracy can sometimes occur in the absence of rapid responding. For these reasons, both response time and answer accuracy are used in the criteria for flagging MAP test events as exhibiting low ISV due to effort.

Rapid-guessing behavior is identified using response time effort (RTE; Wise & Kong, 2005), which is based on the conceptualization that each item response can be classified as reflecting either rapid-guessing behavior or solution behavior (Schnipke & Scrams, 1997, 2002). This classification is done using pre-established time thresholds for each item using the normative threshold method (Wise & Ma, 2012) set at 10 percent. This means that the threshold for an item is set at 10 percent of the average time students have historically taken to answer the item. RTE for a test event equals the proportion of the student’s responses that were solution behaviors. This leads to the first effort flag:

Flag A: If the student gave rapid guesses to at least 15% of the items (overall RTE ≤ .85).

Flag A specifies the amount of rapid-guessing behavior that can be tolerated over the entire test event. Test-taking effort, however, is not all-or-none. Students sometimes exhibit non-effort during only a portion of a test event. This complicates the identification of non-effortful behavior, and the overall indicators may not be sensitive to detecting lesser degrees of non-effort. For example, if a
student gave good effort on the first 43 items of a 50-item CAT and then gave rapid guesses to the remaining items, his non-effort on the last 7 items would not be enough to trigger Flag A.

One solution to this problem is to consider rolling subsets of the items from a test event. For example, for subsets of size 10, we would consider items 1-10, then 2-11, then 3-12, and so on, until the end of the test. In general, for a $k$-item test and subsets of size $r$, there will be $(k-r)+1$ rolling subsets. Using rolling subsets of size 10, we developed two additional RTE-based flags for considering low effort on a more local level:

**Flag B:** If the student exhibited low RTE (local RTE ≤ .70) on at least 20% of the rolling subsets.

Item response time is useful for identifying rapid–guessing behavior. Inspection of MAP data indicates, however, instances in which a student exhibited low accuracy in the absence of rapid guessing. This suggests that some students can become disengaged from during a test event without resorting to rapid-guessing behavior.

Low-accuracy responses should be evaluated carefully, because they could also be due to the student receiving items that were much too difficult for him. In principle, this alternative explanation should not pose a problem for a CAT because the CAT algorithm strives to select and administer items that a given examinee has a .50 probability of answering correctly (which requires that the item pool is capable of providing items that are well targeted to each student). With MAP, however, it occasionally occurs that close targeting is not possible. This tends to happen in lower grades for very low proficiency students, which means that items of low enough difficulty were not available to administer during those MAP test events.

---

4 MAP is usually administered to students several times each academic year and, once an item has been administered to a student, that item cannot be re-administered to the same student for 24 months. After multiple MAP administrations to low proficiency students, this can lead to a shortage of available closely targeted items.
Low accuracy responses should be used as indicators of low effort only if it is established that they were not due to poorly targeted items. To accomplish this, a pool adequacy requirement is imposed specifying that low response accuracy will only be considered for test events in which at least 60% of the time during the CAT, the student received an item whose difficulty was no more than three RIT points away from the student’s momentary proficiency estimate. This led to the development of two additional flags related to response accuracy:

**Flag C:** If the student passed fewer than 30% of the items (overall accuracy ≤ .30) and at least 60% of all of the administered items were within three RIT points of the student’s momentary proficiency estimate.

**Flag D:** If the student exhibited low accuracy (local accuracy ≤ .20) on at least 20% of the rolling subsets and at least 60% of all of the administered items were within three RIT points of the student’s momentary proficiency estimate.

Finally, the joint occurrence of rapid responses and low accuracy on any of the rolling subsets was considered to be particularly indicative of low effort. This led to the final effort flag:

**Flag E:** If the student passed no more than two items (local accuracy ≤ .20) and gave three or more rapid guesses (local RTE ≤ .70) on any 10-item subset, and at least 60% of all of the administered items were within three RIT points of the student’s momentary proficiency estimate.

In the effort analysis of MAP data, a student’s test event was classified as invalid on the basis of low ISV if any one of the five effort flags was triggered.

---

5 The standard errors of student scores in math are typically about 3.0 RIT points, while those in reading are about 3.2 RIT points.
Learn more about our research at NWEA.org/Research.

NWEA has nearly 40 years of experience helping educators accelerate student learning through computer-based assessment suites, professional development offerings, and research services.