Identifying Non-Effortful Student Behavior on Adaptive Tests: Implications for Test Fraud Detection

Steven L. Wise, Lingling Ma, and Robert A. Theaker

Northwest Evaluation Association

Paper presented at the 2012 Statistical Detection of Potential Test Fraud Conference, Lawrence, Kansas

Abstract

Test fraud typically involves actions taken by individuals to purposefully distort a test score in such a way that it overstates what a student knows and can do. Teacher effectiveness, however, is increasingly being evaluated using student growth (i.e., the difference between scores at Time 1 versus Time 2). In this context, a novel type of potential fraudulent behavior becomes a concern. Test givers may be motivated to try to depress student test-taking effort at Time 1 in order to inflate the subsequent growth scores. To investigate the extent to which this represents a real practical problem, data were analyzed from a set of charter schools that has for years used fall-spring growth data to evaluate teacher effectiveness. The results showed clear evidence of lower effort in the fall, supporting the conclusion that this differential effort represents a real threat to the validity of inferences made on the basis of growth scores.

Identifying Non-Effortful Student Behavior on Adaptive Tests: Implications for Test Fraud Detection

The goal of educational assessment is to produce test scores that accurately indicate what students know and can do in a particular content domain. This goal is pursued through the administration of a test whose items adequately represent that content domain and whose items are sufficient in number to yield reliable scores. The scores from such a test are then used to either (a) make inferences about individual students (e.g., What is Michael's level of math proficiency?) or (b) make inferences based on the test scores of groups of students (e.g., How well did the students at Lakeside Middle School do on the state assessment in math?).

Obtaining a valid score for a particular student, however, requires more than just the administration of a well-developed standardized test. There are a number of potential threats to the validity of an individual test score. Some threats pertain to the behavior of the student (e.g., Did he give good effort to his test? Was he feeling ill? Did he cheat to try to get a good score?). Others pertain to the context in which the test occurred (e.g., Was the testing done late in the day? Were there noisy distractions?). Additionally, after the test event had concluded, errors might be made in scoring the test, which could result in a score that was either too high or too low.

Because these types of potential threats to 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 had affected particular scores.

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

$$\text{Distortion}_i = \begin{bmatrix} \text{What the test score} \\ \text{indicates that student } i \\ \text{knows and can do} \end{bmatrix} - \begin{bmatrix} \text{What student } i \\ \text{actually knows} \\ \text{and can do} \end{bmatrix}. \quad (1)$$

Equation 1 is theoretical, as we will not 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 suggests the possibility of at least one of several types of test fraud initiated by the student, who will frequently be motivated to attain the highest score that he can. First, the student could have copied answers from a more proficient test taker during the test session. Second, he could have acquired pre-knowledge of at least some of the test items he would receive, and made sure he knew the answers to them prior to the test session. Third, he may have brought in notes or surreptitiously used technology (such as a smartphone) to acquire the answers to questions during the test session.

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 to appear on a test were 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 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.

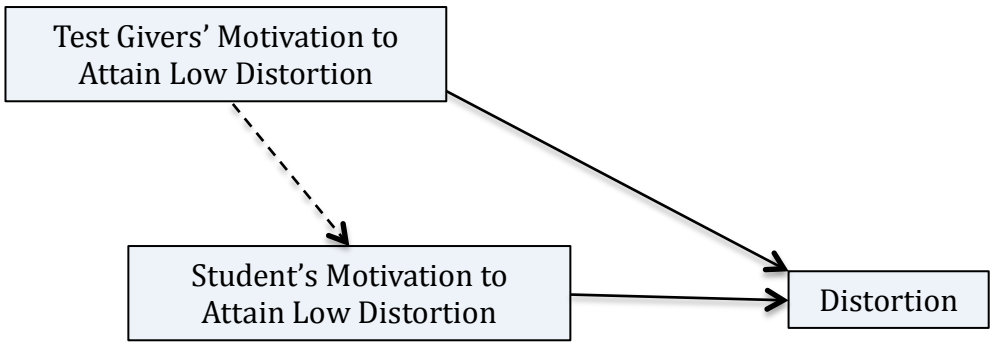


Figure 1. The intentional influences of students and test givers on distortion.

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. Unlike positive distortion, negative distortion is generally not associated with intentional test fraud.

Test Fraud Through Negative Distortion

There is, however, at least one scenario in which test fraud *could* be induced through negative distortion. There recently has been a growing emphasis in U.S. schools on evaluating teacher effectiveness based in part on student growth data. Growth in this context is defined as the difference between a student's proficiency level at Time 1 and that at Time 2 (assuming that the two scores share a common measurement scale). The higher the levels of growth that a teacher's students exhibit, the more positive will be the teacher's evaluation. If a teacher can somehow depress his or her students' scores at Time 1, then student growth will consequently be inflated. That is, if the teacher can induce negative distortion at Time 1, then he or she will induce positive distortion on student growth scores.

How could this be accomplished? One way would be through the manipulation of student motivation. There is a well-documented relationship between test-taking motivation and performance; if a student is unmotivated and behaves in a non-effortful manner, the resulting test score is likely to underestimate his actual level of proficiency (Wise & DeMars, 2005). In testing situations where students perceive few or any personal consequences associated with test performance, they may be particularly responsive from cues from their teacher regarding how much effort they should expend. Thus, to the degree

to which a teacher downplays the importance of the Time 1 test and emphasizes the importance of the Time 2 test, positive distortion could be induced. The dotted line in Figure 1 indicates this indirect influence, by which the test giver induces distortion through manipulation of the student's motivation.

Students exhibiting markedly lower test-taking effort at Time 1 than at Time 2 would characterize the scenario described above. Thus, differential levels of effort provide an important indicator that distortion of growth scores has occurred.

Measuring Test-Taking Effort

As computer-based tests (CBTs) have become more common, there has been an increased 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*).

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

[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.

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 occur during a test event, the more 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 the accuracy of the item responses. 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.

Thus, both response time and response accuracy can provide valuable information about the effort that was expended by students during a CAT. 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, we have developed a set of five flagging criteria for identifying test events 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.

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.

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 (i.e., fall) of an academic year and the end (i.e., spring) of that academic year. 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.

Student effort was measured using both RTE and the percentage of test events whose scores were classified as invalid using the five effort criteria described in the Appendix. Ideally, we sought to study effort for groups of students defined by common teachers or classrooms. For our initial analyses, however, we were able to obtain reliable student groupings only at the school level. Nevertheless, we believed that we might observe differential fall-spring effort even at that coarser level.

The data analyses focused on MAP fall-spring growth scores in math and reading in grades 3-8 from 39 charter schools within a single U.S. state. 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 proficiency 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 about 4.25 points, while those in reading are about 4.50 points.

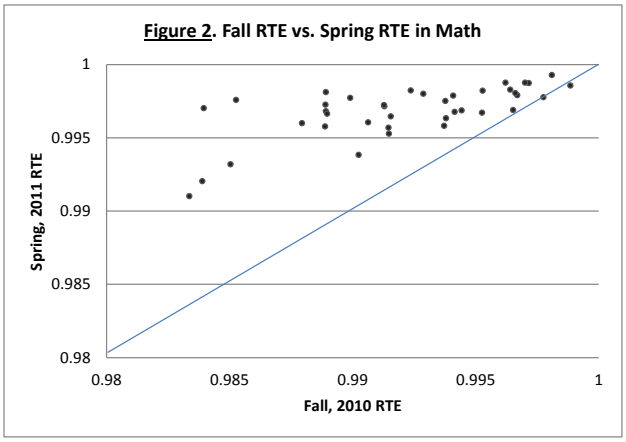
Results and Discussion

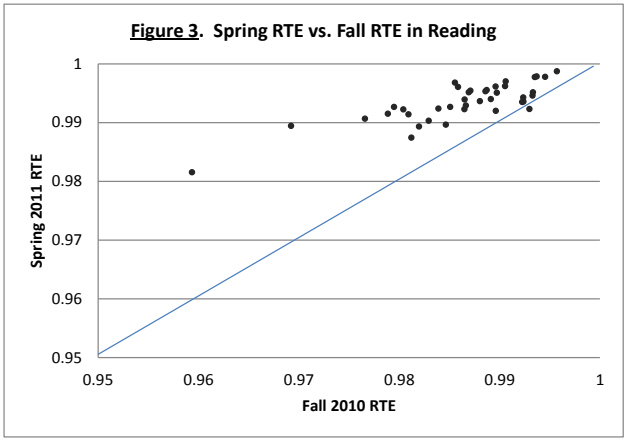
Table 1 shows descriptive statistics for all of the students across the charter schools. Two findings are particularly noteworthy. First, both mean RTE and percent invalid scores indicate that non-effortful behavior occurred more frequently in the fall than in the spring. Second, non-effortful behavior was markedly more prevalent in reading than in math. The first finding was reported by Wise, Ma, Kingsbury, & Hauser (2010) and the second has been reported in several previous studies (Wise et al., 2009; Wise et al., 2010).

Table 1. Descriptive Statistics for RTE, Percentage of Invalid Scores, and Fall-to-Spring Growth

Content Area	N	Mean RTE		Percentage of Invalid Scores		RIT Growth Mean	RIT Growth SD
		Fall	Spring	Fall	Spring		
Math	13,416	.993	.997	3.4	1.4	10.39	7.75
Reading	13,463	.987	.994	7.3	4.1	8.17	8.81

The accuracy of solution behaviors and rapid-guessing behaviors was consistent with our expectations with data from an adaptive test. The accuracy rates for item responses classified as solution behaviors in math and reading were 51.0% and 51.3%, respectively. The accuracy rate for rapid guesses in math (whose items had 5 response options) was 21.0%, which is close to the value expected by random responding. Similarly, the accuracy rate for rapid guesses in reading (whose items had 4 response options) was 26.3%.

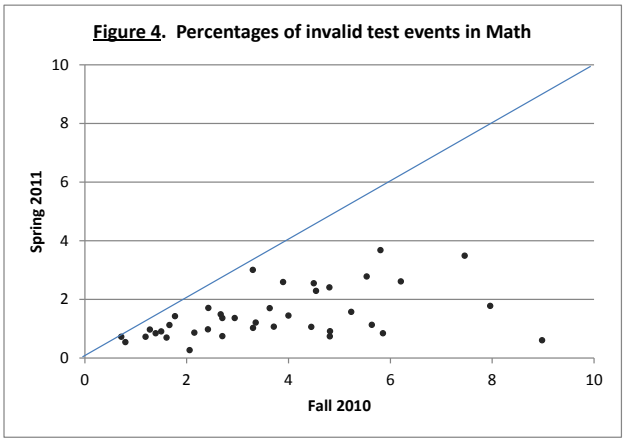


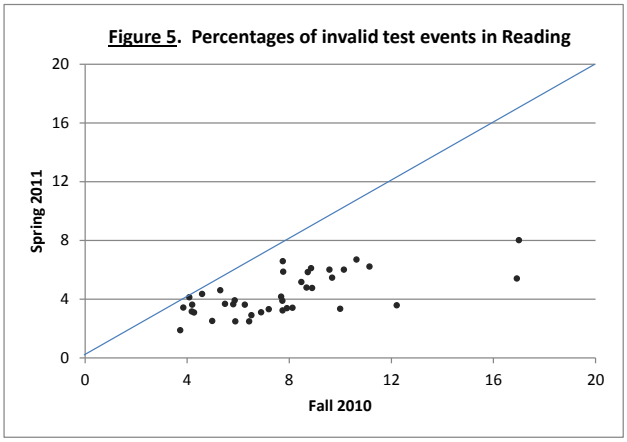


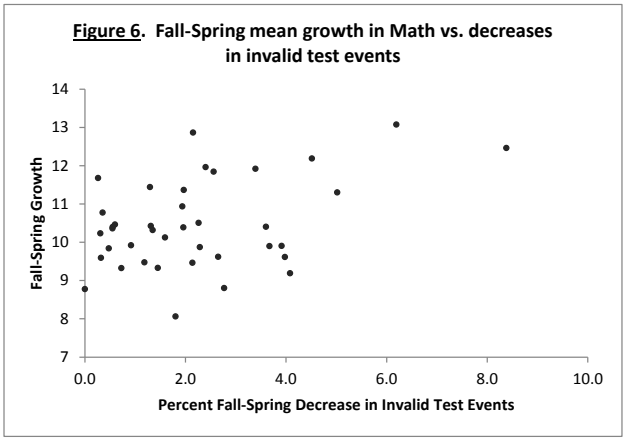
A scatterplot of spring versus fall mean RTE values in math is shown in Figure 2. If effort had been similar in the fall and spring, the data points would be expected to fall near the main diagonal of the graph. Figure 2 shows, however, a somewhat surprising data trend indicating lower mean RTE in the fall for nearly every school. The results for reading, shown in Figure 3, indicate the same basic trend. In addition, there was substantial

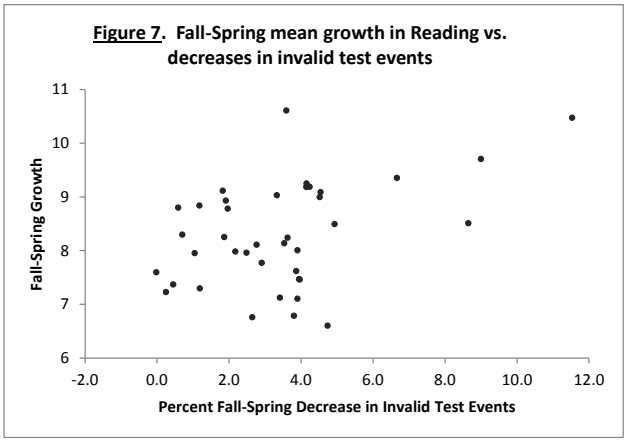
variation in the fall RTE values, with the graphs showing some schools with markedly lower fall RTE.

The percentages of invalid test events in math and reading are shown, Figures 4 and 5, respectively. Again, similar results were found for each content area. There were clear indications that effort was lower in the fall for virtually every school. In math, the difference in the preponderance of invalid test events was as high as 8 percentage points. In reading, the difference for one school exceeded 11 percentage points. These results suggest that lower fall effort was the norm rather than the exception in both math and reading.







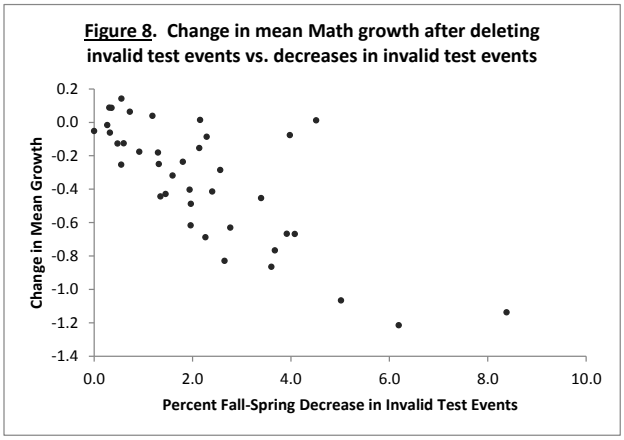


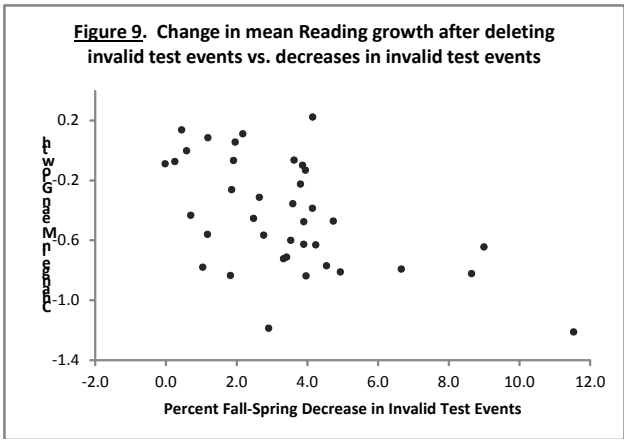
The relationship between the fall-spring decrease in invalid test events and fall-spring growth is shown in Figure 6 for math and in Figure 7 for reading. In both content areas, there was an association between the two variables. Each of the three schools exhibiting a decrease in invalid test events that exceeded 5 percentage points showed above average growth. Similarly, all four of the schools showing at least a 5-point decrease in invalid test events in reading exhibited above average growth.

The results shown in Figures 4-7 provide the clearest answer to the general research question that this case study was intended to answer (i.e., is there evidence of disparate fall-to-spring student effort that induces positive distortion on growth scores?). The finding that nearly all of the schools showed lower effort in the fall suggests that (a) there may be a general tendency of test givers to downplay the relative importance of fall testing or (b) students may be relatively disinclined to devote strong test-taking effort early in an academic year. Whichever explanation is more valid, the finding that the largest decreases in effort was associated with above average growth suggests that the degrees of differential fall-spring effort observed in these schools may have a meaningful impact on mean growth scores.

Demonstrating that differential effort can lead to positive distortion leads to a question regarding the degree to which this distortion biases the growth scores. Figures 8 and 9 show the impact on mean growth of deleting the scores from test events identified as invalid. There is a negative relationship, particularly for math. High degrees of differential effort appear to be associated with larger decreases in mean growth. These findings support the claim that higher growth is due, at least in part, to higher differential effort.

It might be tempting to interpret the vertical scales in Figures 8 and 9 (change in mean growth) as indicators of the amount of distortion induced by differential effort for a given school. There are, however, two caveats to consider. First, Figures 8 & 9 show what would happen if the scores from invalid test events were deleted. We do not know for certain the degree to which the deleted scores were distorted by differential effort. Second, we are not suggesting that only the scores from deleted test events were subject to distortion. Our effort flagging criteria are conservative in detecting non-effort, which means that we will classify a test event as invalid only if there is clear evidence of non-effortful behavior (Wise & Ma, 2012). This implies that there will typically be some amount of non-effortful behavior that will go undetected.





The data analyses from the case study illustrate that differential effort does appear to occur in practice at the school level, and that this induces meaningful amounts of positive distortion in growth scores. This implies that test fraud could successfully be accomplished by test givers through a manipulation of the amount of relative effort students give to their fall and spring assessments.

General Discussion

This paper introduced and discussed the concepts of test score distortion. It was shown that most types of cheating (by either test givers or students) are intended to induce positive distortion on test scores, with the goal of overstating what students actually know and can do. Negative distortion, in contrast, is generally caused by the influences of one or more sources of construct-irrelevant variance that exert a negative bias on test scores. When student growth is considered, however, it is important to realize that positive distortion on growth scores could be induced through a negative distortion of Time 1 scores through a manipulation of student effort.

How, specifically, could this be done? Given the high-stakes nature of many assessments, it might appear difficult to lower a student's test-taking effort. It should be noted, however, that many assessments deemed "high-stakes" are actually "little-to-no-stakes" from the perspective of the student. For example, the scores from state assessments given according to the No Child Left Behind legislation have carried enormous consequences for school accountability. From the student's perspective, in comparison, the scores from these assessments have carried few personal consequences. The student's grades are unaffected by performance on the state assessment, and scores on the assessment usually are unavailable until months after the test event.

Given the absence of personal (i.e., external) consequences associated with many standardized assessments, one might reasonably turn the question around and ask why students give *any* effort to these assessments. In the absence of personal consequences, a student's test-taking motivation is driven by internal factors, such as a desire to please teachers and parents, academic citizenship, competitiveness, and ego satisfaction (Wise &

Smith, 2011). In this context, cues from the teacher about the importance of a particular test and the amount of effort expected from students can have a strong influence on the amount of subsequent effort that occurs. Certainly, students are generally not eager to take tests, and it may require only subtle cues from the teacher that a test event is not highly important for them to decrease the amount of effort expended. Teachers are used to devoting a lot of time encouraging their students to give their best efforts to standardized assessments, and it require little more than inaction on their part during fall testing to induce a sizable positive distortion on fall-spring growth scores.

It is important to consider what constitutes test fraud. In this paper, test fraud means that someone acted intentionally to induce a positive distortion on test scores. However, a teacher who de-emphasizes the importance of fall assessment performance relative to that emphasized for spring testing may not view his actions as fraudulent. A principal who encourages her teachers to view the fall testing merely as a baseline (consequently de-emphasizing its importance), and thereby indirectly diminishing test-taking effort might not view her actions as fraudulent. But from the standpoint of score distortion, these actions constitute fraudulent behavior even if test givers do not perceive it as such.

Given the emerging emphasis of student growth data on teacher evaluations, it is likely that the problem discussed in this paper will constitute a serious threat to the validity of inferences made about teacher performance. Teachers will increasingly be tempted by the evaluative benefits of discouraging effort on fall assessments. Moreover, teachers who would normally resist the temptation may adopt the practice anyway, because they believe that other teachers are doing so. That is, they may be trying to re-

level what they perceive to be an uneven playing field. The pervasive differential effort exhibited by schools in the case study may be evidence that this has occurred in the charter schools.

Recommended Solutions

Student growth data provides valuable information to teachers about the instructional needs of individual students. That is its primary purpose. When growth scores are also used to evaluate teacher effectiveness, there is a very real risk that test givers will attempt to distort the growth scores so as to put themselves in the best possible evaluative light. In the process, however, they undermine the validity of the primary instructional use of the growth data. It is therefore important that this threat to growth score validity be effectively addressed. This can be done through a combination of detection, deterrence, and mitigation.

When a CAT is used, the five effort flagging criteria used in this paper can be useful in detecting scores that are invalid due to low ISV. This would permit one to audit the results from a school, a classroom, or a student. If markedly lower effort was exhibited in the fall testing term, an investigation should be initiated into the causes of the discrepant effort, and whether or not those causes constitute test fraud. If a non-adaptive CBT is used, then the two flagging criteria relying on response time that are not based on score accuracy could be used to detect invalid scores—though probably not as effectively as when a CAT is used. When paper-and-pencil tests are used, it would be more difficult to detect low effort. Post-test self-report measures regarding effort might be used, but the validity of their scores might be questionable because it would be difficult to ascertain how truthfully

students would respond. This, then, represents an additional advantage of computer-based tests over paper-and-pencil tests.

If procedures were put into place to detect differential effort, then test givers should be made aware of them, as well as of potential sanctions that could result from their occurrence. The knowledge that differential effort can be detected should serve as a strong deterrent. This awareness should be augmented with professional development for test givers that clearly defines this type of test fraud and what constitutes unacceptable practices.

The remaining problem concerns how to effectively deal with instances of differential effort that do occur. Simply deleting the growth scores based on invalid test events is unsatisfactory for two reasons. First, we would like to get valid instructional information for as many students as possible, and deleted scores would not provide this information. A second, more cynical reason is that the deletion of scores from invalid test events could lead to different fraudulent strategy. A teacher may be motivated to selectively discourage effort only from the students who he believes are likely to show the lowest growth. Strategic deletion of their data would provide an additional way for a teacher to inflate his students' growth.

To mitigate the problem of low student effort, statistical methods may be useful in both salvaging growth scores for as many students as possible and in facilitating more valid assessment of teacher effectiveness. Wise and DeMars (2006) investigated a, IRT-based method for adjusting scores for the effects of non-effortful test-taking behavior. This method involves classifying each item response as solution behavior or a rapid guess, and basing proficiency estimation for a student only on solution behaviors (i.e., ignoring rapid

guesses). This method is based on the assumption that for a test event containing both solution behaviors and rapid guesses, the solution behaviors all reflect effortful behavior. Bowe, Wise, and Kingsbury (2011) found evidence that this assumption does not always hold. Thus, additional research is needed to develop methods for statistically managing the effects of low student effort on test scores.

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.

Bowe, B., Wise, S. L., & Kingsbury, G. G. (2011, April). *The utility of the effort-moderated IRT model in reducing negative score bias associated with unmotivated examinees*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1),* 17-27.

Hauser, C., Kingsbury, G, G., & Wise, S. L. (2008, March). *Individual validity: Adding a missing link.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Hauser, C., & Kingsbury, G, G. (2009, April). *Individual score validity in a modest-stakes adaptive educational testing setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Kingsbury, G. G., & Hauser, C. (2007, April). *Individual validity in the context of an adaptive test.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ma, L., Wise, S. L., Thum, Y. M., & Kingsbury, G. G. (2011, April). *Detecting response time threshold under the computer adaptive testing environment.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Schnipke, D. L. & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Education Measurement*, *34(3),* 213-232.

Schnipke, D.L. & Scrams, D.J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1-18.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38.

Wise, S. L., Kingsbury, G. G., & Hauser, C. (2009, April). *How do I know that this score is valid? The case for assessing individual score validity.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

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

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

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010, May). *An investigation of the relationship between time of testing and test-taking effort.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, and C. W. Buckendal (Eds.) *High-stakes testing in education: Science and practice in K-12 settings (pp. 139-153).* Washington, DC: American Psychological Association.

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 who 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[2].

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[3]. 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.

---

[2] 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.

[3] 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.

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.