KINGSBURY RESEARCH CENTER

Interim assessment and prediction

John Cronin and G. Gage Kingsbury, Northwest Evaluation Association

Paper prepared for the
Technical Issues in Large Scale Assessment Subgroup
State Collaborative on Assessment and Student Standards
Council of Chief State School Officers
February 5, 2008

Interim Assessment and Prediction

John Cronin and G. Gage Kingsbury, Northwest Evaluation Association

Enactment of the No Child Left Behind Act encoded in policy a requirement that schools ensure that all students achieve proficiency in the basic academic skills. While, schools have until 2014 to fully meet the 100% requirement of the Act, they are expected to meet annual Adequate Yearly Progress requirements each year. A series of progressive sanctions are imposed on schools that fail to meet AYP for two consecutive years.

To help teachers know the needs of their students, and to help students succeed in reaching these proficiency targets, many schools have developed or purchased interim assessments. Perie, Marion, Gong, and Wurtzel (2007) describe interim assessment in the following manner:

> "These [interim] assessments may serve a variety of purposes, including predicting a student's ability to succeed on a large scale summative assessment, evaluating a particular educational program or pedagogy, or diagnosing gaps in student's learning."

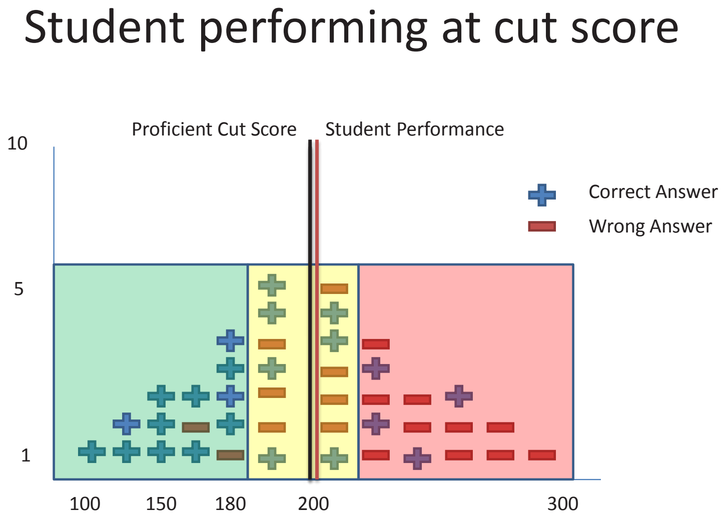
Although prediction is an important function of these assessments, the authors emphasize that interim assessment, when used appropriately, is not merely a practice assessment for the state proficiency exam, and that prediction is only one of several purposes that should be met by such tests.

Test Design and Prediction

In general, a test designed to make accurate predictions of proficiency status (for our purposes proficient or not) selects most of the items so that their difficulties are near the cut score or decision point. This is done to minimize the standard error at the cut score, which maximizes the probability of correctly predicting student cases which are near it. For a high stakes test this is quite important, because the consequences of getting a student's status wrong have may have heavy consequences for students or schools.

Figure 1 shows one of these hard cases and a test design that is designed to provide good prediction. In this case the majority of the items have a difficulty level that is near the cut score in difficulty. For a student performing near the cut score, this design maximizes the information gathered about the student's proficiency. Even though this design is sound, it will not correctly predict the status of these hard cases 80% of the time. Since the sample case is barely above the cut score, if the student was tested 100 times, just over 50% of those occasions would produce a result above the cut score.

Figure 1 – Theoretical item correct distribution for student scoring at the cut score on a test



The majority of proficiency predictions, however, are not hard. Teachers frequently comment that they can tell you within a few days of instruction which students in their class will be proficient. For most children they are probably right. Most prediction decisions are relatively easy because they are not close. Figure 2 shows a low student who is easy to predict as non-proficient. It is easy because the student not incorrectly answered most of the questions that are near the proficiency level but also incorrectly answered a large number of questions below the proficiency level.

Of course the same principle applies to students who perform well above the standard. Figure 3 shows a high performing student whose proficiency status is also easy to predict. The prediction is easy because the student not only answered the questions near the proficiency standard correctly, but also answered many tougher questions correctly. Thus the proficiency prediction is, as some would say, a no brainer.

Figure 2 – Theoretical item correct distribution for student scoring well below the cut score on a test

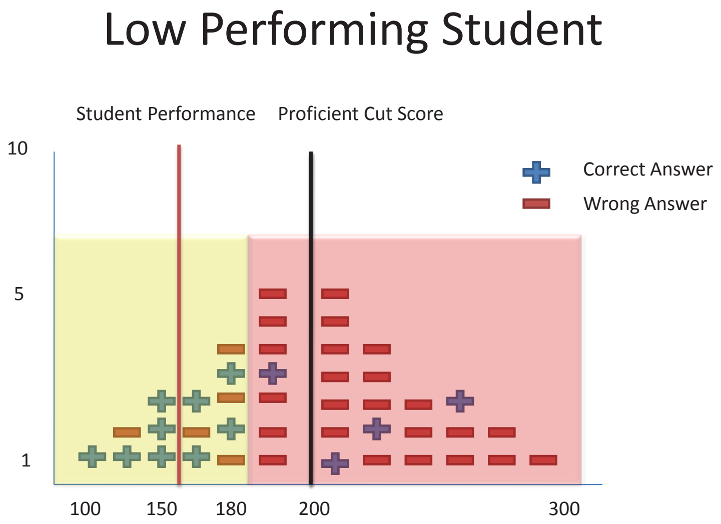
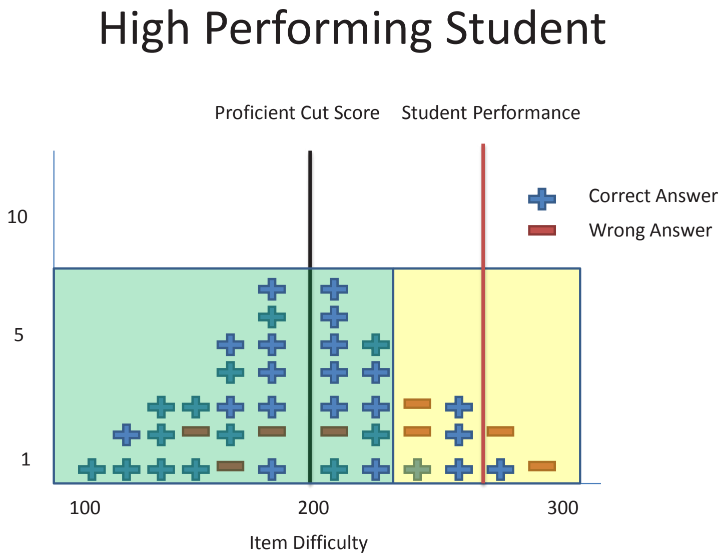


Figure 3 – Theoretical item correct distribution for student scoring well above the cut score on a test



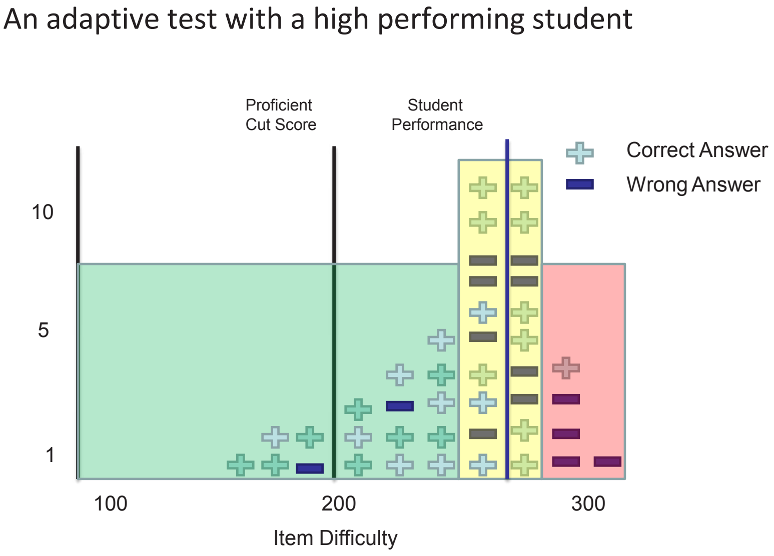
If the purpose of interim assessment was limited to prediction, this type of test design works relatively well. It gets the easy cases right and gives us a good shot at correctly predicting many of the harder cases. However, interim assessment has other purposes. If those purposes include diagnosing gaps in student learning, this design works poorly for many students.

In the case of our low performing student, this test design provides a lot of data supporting the "not proficient" prediction, but results in a fairly large standard error of measure around the student's score. In other words we have sacrificed precision in the score to gain precision in the prediction.

The same problem applies to the high performer (Figure 3). We have very high confidence in the proficiency decision, but much less confidence about how high performing this student might actually be. If our goal is diagnose gaps in student learning, this design is not useful.

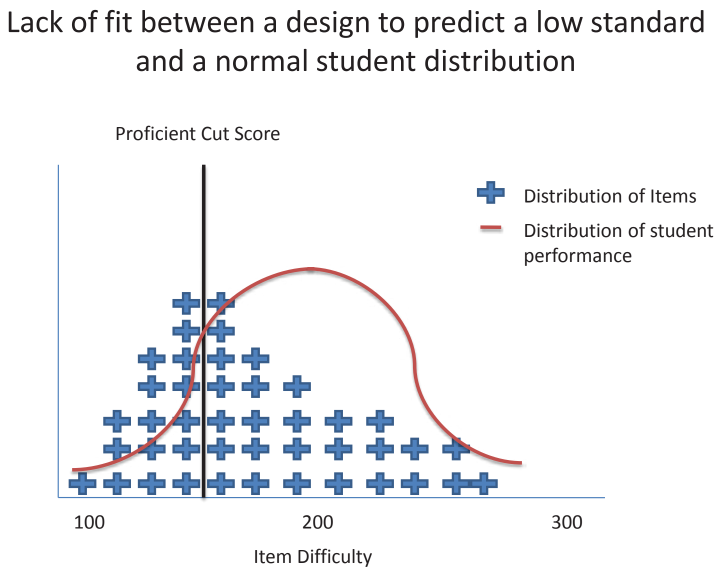
Figure 4 offers an alternative design, in this case an example of an adaptive test taken by a high performing student. The prediction that the student will perform above the standard is well-supported by data collected during the test. Because a large proportion of the items are focused near the students final achievement estimate, the test also provides more information to inform instruction than the design in figure 3, in particular we know a lot more about what this student still needs to learn. This design which focuses item content on the learner rather than the proficiency standard, predicts proficiency results well and is more likely to yield useful instructional information for the teacher or program administrator.

Figure 4 – Theoretical item correct distribution for a student performing well above the cut score on an adaptive test.



The above illustration assumes that the proficiency cut score is near the middle of the achievement distribution. Of course, this may not always be the case. As the cut score moves from the middle of the distribution, designs aimed at maximizing prediction provide less reliable information about more students' performance. Figure 5 shows a design that would be appropriate to provide good prediction relative to a lower proficiency cut score. Unfortunately, because most students perform well above the standard, this test design is not well-targeted to the achievement of students at the middle to high end of the performance distribution and would increase the standard error around more scores.

Figure 5 – Comparing a highly predictive design to a test performance distribution for an assessment with a lower than average standard.



Interim assessments have purposes that go beyond practice for the state test and prediction of the result. As a result, a good interim assessment would not necessarily mimic the design of a state assessment. **The state proficiency test needs to get the proficiency decision right and designing for that purpose has high importance. But an interim assessment should get the performance of the student right, and that goal is served better by using designs that minimize the standard error around each student's score.** Such a design does not have to greatly erode the accuracy of prediction.

Predicting a student's proficiency status has its place in interim assessment, but we are not sure that its place is at the front of the line.

Prediction Metrics

A number of interim assessments now report a student's projected proficiency status and/or proficiency category as part of their reporting package. While this information is useful, we believe information about the probability of a student's proficiency may be more appropriate. Assume an interim test is predicting whether a student is likely to reach a scale score of 200 on the state assessment. A student scoring 199 is placed in the same category as a student scoring 190 or one scoring 175. Nevertheless, the student scoring 199 has nearly a 50/50 probability of passing the state assessment, while the student scoring 175 have very poor prospects. In this case, projecting the student into a binary category (proficient or not) or even projecting the student into a performance band (advanced, proficient, partially proficient, basic) costs us important information about the student's performance.

One way to help address this issue is to report the **probability** that a student will perform at a certain level in addition to reporting the projected proficiency status or performance category. Table I shows an example of a table our organization generates from our alignment studies to provide this information.

The simplest way to generate this table is to take a sample of students who have taken both the interim assessment and their respective state test and report the proportion of students in each scale score range (in this example a 5 point range) who went on to achieve proficiency on the state assessment. When used well, this information can be very helpful to educators and students. For example, probability information can help identify students who are well beyond the proficiency cut score (perhaps 80% probability or better). Instruction for these students can be adapted to focus on content that reflects performance beyond the proficiency cut score. Likewise, probability information can help identify the degree of intervention that might be needed for students who are near or below the proficiency cut score. For example, students with a 5% probability of passing will need more time and resources to achieve a proficiency standard than students who have a 45% probability of passing.

There is also some risk that probability information can be used in ways that may not support improved learning. One concern is that probability information can be used to support educational triage decisions. Schools could use probability data to more precisely identify students to identify "bubble students" for special attention, students whose scores offer the best prospects for leveraging a school's Adequate Yearly Progress report. While this is a risk, we think schools are better served if the data is available and teachers are well-informed about its proper use.

Table 1 – Proportion of students passing a state reading assessment based on scale score range

	Proportion of students in this score range achieving proficiency on the state reading assessment								
Bottom of Scale Score Range	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
150	16%	4%	1%	0%	0%	0%	0%	0%	0%
155	23%	6%	2%	1%	1%	0%	0%	0%	0%
160	33%	9%	3%	1%	1%	1%	0%	0%	0%
165	45%	14%	4%	1%	2%	1%	1%	0%	0%
170	57%	22%	7%	2%	3%	2%	1%	0%	0%
175	69%	31%	11%	4%	5%	3%	1%	1%	0%
180	78%	43%	17%	6%	8%	5%	2%	1%	1%
185	86%	55%	25%	10%	12%	8%	4%	2%	1%
190	91%	67%	36%	16%	18%	12%	6%	3%	1%
195	94%	77%	48%	23%	27%	18%	9%	5%	2%
200	96%	84%	60%	33%	38%	27%	14%	8%	4%
205	98%	90%	71%	45%	50%	38%	22%	13%	6%
210	99%	94%	80%	57%	62%	50%	31%	20%	10%
215	99%	96%	87%	69%	73%	62%	43%	29%	16%
220	99%	98%	92%	78%	82%	73%	55%	40%	23%
225	100%	99%	95%	86%	88%	82%	67%	52%	33%
230	100%	99%	97%	91%	92%	88%	77%	64%	45%
235	100%	99%	98%	94%	95%	92%	84%	75%	57%
240	100%	100%	99%	96%	97%	95%	90%	83%	69%
245	100%	100%	99%	98%	98%	97%	94%	89%	78%
250	100%	100%	100%	99%	99%	98%	96%	93%	86%

What is the value of prediction?

Achievement of the state proficiency cut score should, in the ideal, represent a level of performance that projects to a good prospect of future academic or life success for a student. Assuming that is the case, it is very important to identify students who are not on track to meet this objective and to implement some plan of academic intervention to support these students. Prediction can also be useful if it can help schools identify the type and amount of resources that might be needed to help students eventually reach proficiency.

While this may seem obvious, in reality, the pressure to meet AYP requirements often pushes schools to use prediction to identify students that jeopardize the school's AYP goals. In this case, prediction is serving the interest of the school and staff more than the interest of students. For example, the students most in need of additional resources and intervention may be the ones whose current achievement is furthest below the standard. This is especially true of very low performing students who may not have been identified as eligible for special education services. However the students who have the best prospects for improving a school's AYP prospects are commonly those whose current performance is closest to the cut score. Thus providing predictive information potentially has a perverse effect, by encouraging a form of triage that focuses resources on the students who can keep a school out of "AYP jail" rather than targeting resources on the students who most need them. Evidence of NCLB triage is available and growing (Neal & Whitmore-Schanzenbach, 2007; Booher-Jennings, 2005; Rosenshine, 2003). At the same time, research does not indicate that triage practices cause greater growth in the targeted students (Springer 2008).

How to address this?

While this is not an easy problem to address, there are several strategies test designers can pursue:

- Emphasize in the design that an interim assessment is not merely a "practice test" for the state assessment, but that the assessment serves other important instructional purposes that are complementary but independent from the state proficiency test.
- The test design should be more focused on getting an accurate estimate of each student's performance than it is on making an accurate prediction of proficiency status or category. If one is only designing a test to make a decision about whether a student is proficient, a form that concentrates the difficulty of the test near the proficiency cut score is reasonable. That is not the only purpose of an interim assessment however. An interim assessment should focus on getting the most accurate estimate of the student's performance so that instruction can be appropriately tailored to that individual's need. If the test is focused on the prediction of proficiency, it sends the message that the test is designed to help the school meet AYP. If the test is focused on accurate assessment of the student, it sends the message that the test is intended to help the student improve on their current performance regardless of their proficiency status.
- The test design should be designed to monitor growth. State tests do an excellent job of identifying the group of students whose performance places them "at risk" to leave school without achieving proficiency. Many schools have had success in using this information to successfully address more of these student's needs. But, at least in the ideal, schools exist to help all students reach their full potential. Any student who is not making progress toward achieving that potential is at risk. Interim assessments complement the state assessment if they provide data that helps identify students who may not be showing good growth and such assessments can help schools focus not only on making sure their students are proficient, but also help them focus on helping all students reach their full potential.

Factors affecting the accuracy of prediction

There are at least five factors that affect the accuracy with which an interim assessment will predict proficiency status on a state assessment. These are:

Testing Conditions and the Motivation of the Student. These are likely to have as much or more effect on the accuracy of the proficiency prediction than the design of the test itself. This is particularly true because the conditions surrounding the administration of interim assessments, which are generally low stakes, will inherently differ from those around the state assessment itself. While the data is not entirely conclusive, a fairly convincing body of evidence exists that would indicate students perform better on tests with some stakes. NWEA completed a study of the effects of stakes in 2006 and found that in low stakes situations about 3 to 5% more students “tank” on the test (Cronin, 2006). Tanking occurs when a student achieves a score that is both substantially below a prior achievement effort, and evidence exists (typically in the form of greatly reduced item response time and deviation from normal response patterns) to suggest a lack of effort. Tanking behavior would normally increase the rate of incorrect prediction due to underestimation (Type II error). Similarly, some teachers may not attend to general testing conditions are carefully in low stakes testing conditions, and this could also contribute to situation specific underestimation of performance.

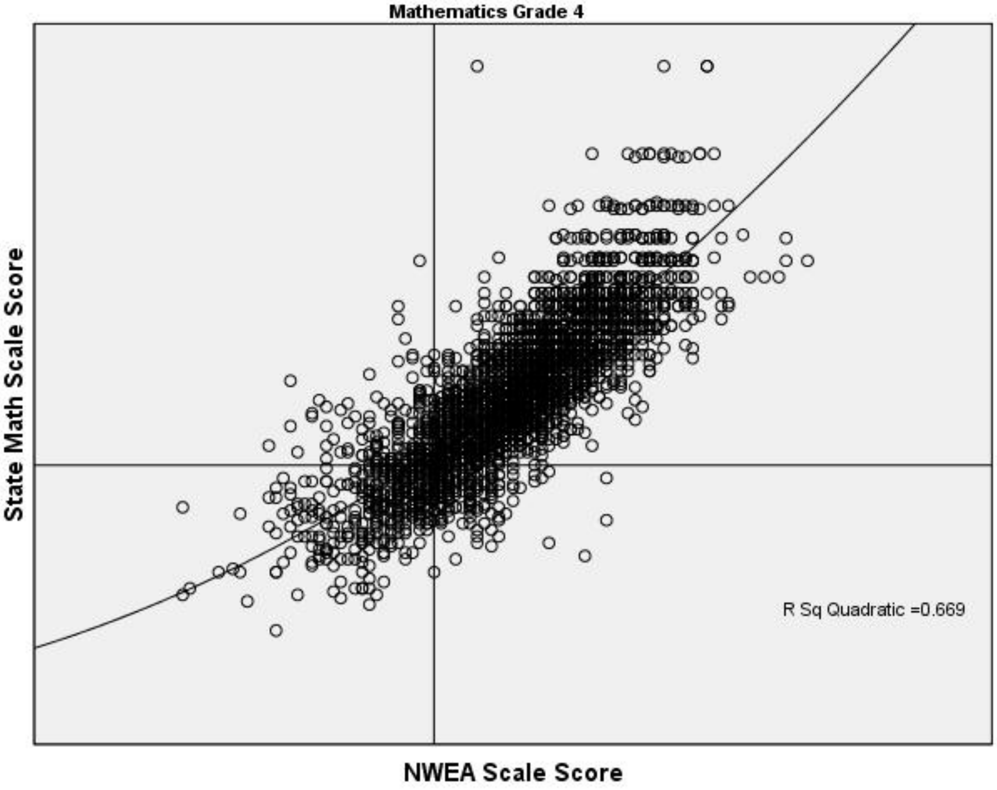
Alignment of content and item formats. The instructional purposes underlying interim assessments make it important to align them closely to state content standards. The degree of alignment is not neccesarily a strong indicator of predictive accuracy, however.

Our organization has done alignment studies in three states in which we attempted to predict results for their English/Language Arts test from our reading test. In each case, our assessment did not include content related to language usage and writing that would have been included on the state assessment. We compared correlations in these three states to other states in which we directly aligned our reading assessment to the state assessment and found no meaningful difference in the predictive validity statistics, with an average Pearson *r* of .80 for the English/Language Arts states compared to an average Pearson *r* of .79 for the state’s testing reading alone. In these particular cases, the difference in alignment did not substantively affect the predictive ability of the interim assessment. That is not to dismiss the importance of content alignment, nor would we want to imply that two very poorly aligned assessments would be strongly predictive of one another. It may be true, however, that content alignment might not affect the quality of prediction as much as some other factors.

Measurement precision of the two tests. This is obvious and is generally the factor that gets the greatest attention. Nevertheless the more precise the measurement instruments, the greater the probability of a correct prediction. One factor that greatly affects precision is the measurement range of the assessment. If one assessment has a different measurement range than its companion assessment, the accuracy of the prediction can be affected.

Figure 6 shows the correlation between an NWEA mathematics assessment and a state mathematics assessment. In this diagram, the intersection of the horizontal and vertical lines depicts the state cut score and the corresponding NWEA scale score prediction. In this case the relationship between the scores is curvilinear, primarily because this particular state assessment had lower measurement accuracy for low-performing students. The difference in measurement precision limits the predictive accuracy at the decision point.

Figure 6 – Correlation between grade 4 performance on the NWEA MAP mathematics test and a state mathematics assessment.



Placement of the cut score. The position of the cut score potentially has a very large impact on the accuracy of prediction. Assume for a moment that we have a group of 100 fourth graders and we want to know if they are proficient at high jumping. Assume further that five feet was set as the standard for proficient performance. In this case, the placement of the cut score makes it easy to get a high degree of accuracy in my prediction. I simply predict that everyone will fail and my result is likely to be somewhere between 98% and 100% accurate. Had the standard been set at one foot, it would be equally simple to predict the results. In any case, one would expect less accuracy when cut scores are near the middle of a distribution and greater accuracy when cut scores are near one of the two ends.

Time between the interim assessment and the proficiency test. Educators are used to seeing predictive reliability coefficients that reflect the correlation between two assessment instruments that are administered at about the same time. Interim assessments are intended to influence instruction and will typically be delivered well in advance of the state proficiency test, but the impact related to instruction is not necessarily random across schools. For example, schools that create unusually large growth are likely to have more missed predictions due to Type II errors, because a greater number of students who were projected not to pass will successfully meet the proficiency standard. In such a case, poor prediction would be a good thing. Unfortunately, there may also be low growth schools that will have more missed predictions due to Type I errors, because students who were anticipated to make proficiency, failed to make enough progress to cross the bar. Thus in assessing the predictive accuracy of an instrument, one should adjust their expectations to reflect the opportunity for instruction, AND, one should be aware that the predictive accuracy of the assessment will be lower in schools that are far above or below average in student growth.

Of course, more time for instruction between the interim assessment and the state test will also generally have some corrosive impact on the accuracy of prediction.

In prior studies of alignment, NWEA assessments typically average between about 83% and 90% accuracy in predicting the performance of students.

Other issues related to prediction of state assessment results

State tests are improved from one year to the next. As the standards movement and assessment have evolved, state departments of education have become considerably more sophisticated in the tasks related to this work. This knowledge and experience leads states to make refinements to both their curriculum standards and their tests. Anytime the state test changes in substance, the predictive relationship between that tests and an interim assessment becomes, less well understood until data is available to permit updating the linkage between the two. In the meantime, educators who rely on interim assessments to predict a student's status relative to the state test are at risk to make decisions based on outdated information.

Student change on the statewide assessment may not parallel change on the interim assessment. A wide body of evidence exists demonstrating that performance on state assessments has improved more rapidly than performance on NAEP and interim assessments such as NWEA's. When this is the case, a proficiency cut score prediction generated in 2004 may significantly overstate the score needed to achieve proficiency in 2008. While the source or cause of these differences is open to debate, the effect of the differences on the ability to predict is a known fact that must be addressed if interim assessments are to have utility as predictors of state test performance.

The methodology used to set the prediction equation may cause predictive accuracy to vary . Prior to 2006, NWEA used a variety of regression based methodologies to do cut score prediction. We reevaluated our methods in 2007, because the regression based methods tended to overpredict pass rates by about 3 to 8%. We have since moved to an equipercentile based methodology that produces a lower rate of overprediction, but this method still tends to predict slightly more students will pass the state test than actually do. We are not really sure why this is the case.

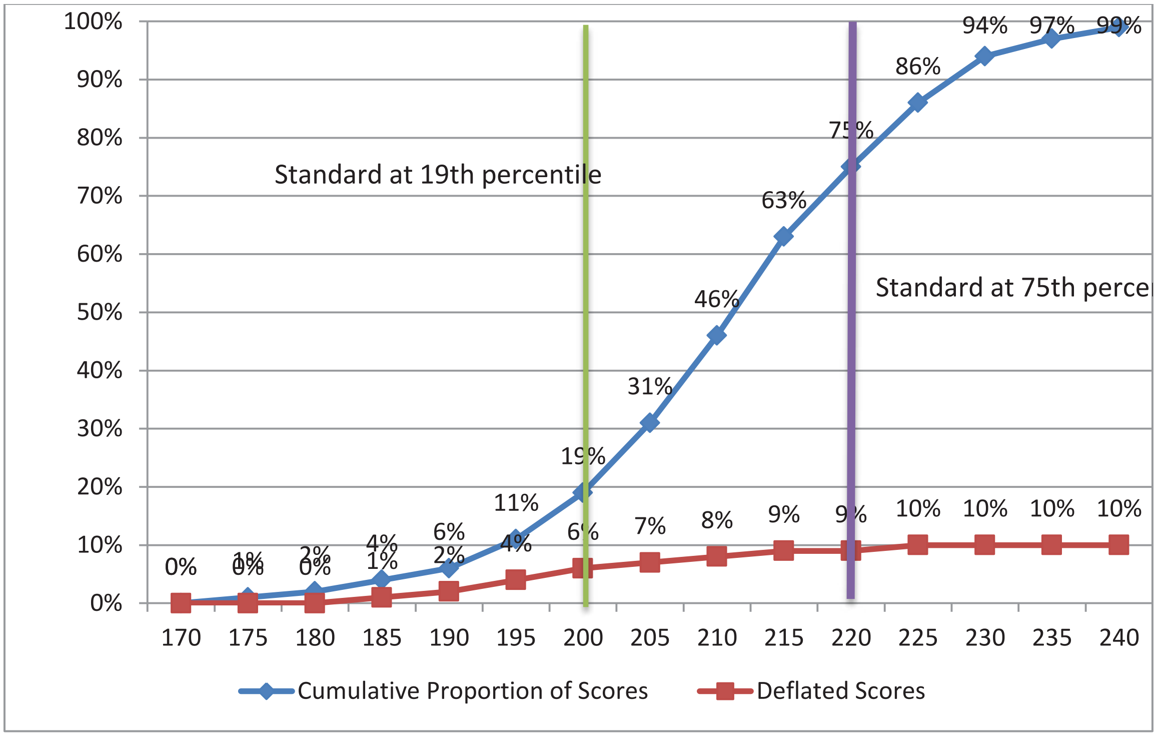
Cut score placement and student attention may have an impact on predictive accuracy. Previous studies of state standards peg the range of the majority of cut scores to be between about the 15th and 60th percentile of student performance (Cronin et al, 2007; National Center for Educational Statistics, 2007). A certain number of students, we estimate between about 4 and 12% depending on the stakes associated with the test, may not engage with a test, and therefore will perform substantively below their real achievement level.

Most of the students who don't engage well with the test will cluster near the low end of the performance continuum. Figure 7 shows a hypothetical cumulative frequency distribution of 4th grade mathematics students, alongside a distribution of students in the population who had deflated estimates due to lack of motivation or other factors that influenced their performance on the date of the test.

Assume that a proficiency standard is set at the 19th percentile of performance. If about 6% of the population's scores are deflated, it would mean that nearly 1/3 of the cases performing below the 19th percentile are potentially false negatives. Many of these students may be identified for participation in interventions. Unfortunately, if the intervention needed was simply a bit more motivation on the day of the assessment, or perhaps a hearty breakfast, then educational interventions may be wasted on that group of students. Had the standard been set at the 75th percentile, the proportion of cases that represent potential false negatives is lower (though the number of cases is higher).

In any case, when a proficiency standard is set below the middle of the performance continuum, the proportion of false negatives among the identified "at risk" population is very likely to increase. This reinforces the need for schools to bring multiple sources of data to the table when identifying students for intervention based on their likely proficiency status. In addition, if schools are aware of this risk, they can take steps to more carefully monitor interim assessments. Test publishers can do their part by researching and implementing strategies to identify scores that may have suspect validity due to student guessing behavior or abnormally low item-response time on a computerized assessment.

Figure 7 – Hypothetical distribution of students and deflated mathematics scores on a fourth grade mathematics assessment.



Discussion

Interim assessments are capable of predicting proficiency on state assessments with an accuracy rate of 80% to 90%. Whether the rate of accuracy is closer to 80% than 90% is influenced by a wide variety of factors especially the placement of the state proficiency cut score, the interim assessment's design, the stakes associated with the two assessments and the match between the interim assessment's content and format and the state assessment. Of these factors, the placement of the cut score is likely to have the greatest effect on predictive accuracy.

A single-minded pursuit of greater predictive accuracy may compromise other important objectives that states may have for an interim assessment. For example, the best predictive design for an interim assessment is one that measures how students perform on content near the cut score. Unfortunately this design results in less accurate measurement of students who perform far above or below it. Because the proficiency cut score in many states is not set at the middle of the distribution, there is also a risk that a proficiency-focused interim assessment may be a poor measurement fit for the majority of students. The broader definition of interim assessment would be better met by using an assessment designed to measure each student well, while also providing reasonable prediction of statewide assessment results. This could be accomplished by using and adaptive test or a slightly longer broad-range test.

The approach to interim assessment that bests meets the broader purposes of the form is one that gets the most accurate measure of the student's performance rather than the most accurate measure of the student's proficiency. It is substantially easier to use a test that was designed to give accurate results for each student to predict proficiency than to use a test designed to estimate proficiency to determine the instructional needs of each student. Interim assessments that provide highly accurate student results should demonstrate a high degree of predictive accuracy, but they will be more useful for schools that are trying to assess where to start students in their curriculum or for schools that are trying to identify gaps in their current instructional programs.

References

Booher-Jennings, J. 2005. Below the bubble: “Educational Triage” and the Texas Accountability System. *American Educational Research Journal 42* (2): 231-268.

Cronin, J. (2006, April). *The effect of test stakes on growth, response accuracy, and item-response time as measured on a computer-adaptive test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cronin, J., Kingsbury, G. G., Dahlin, M., Adkins, D., and Bowe, B. (2007, April). *Alternate methodologies for estimating state standards on a widely used computer-adaptive test.* Paper presented at the Annual Conference of the American Educational Research Association, Chicago, IL.

Cronin, J., Dahlin, M., Adkins, D. and Kingsbury, G.G. (2007, October). *The Proficiency Illusion.* Washington, D.C.: Thomas B. Fordham Institute.

National Center for Educational Statistics (2007). *Mapping 2005 State Proficiency Standards onto the NAEP Scales*(NCES (2007-482). U.S. Department of Education: Washington, DC.

Neal, D. and Whitmore-Schanzenbach, D. (2007). *Left Behind by Design: Proficiency Counts and Test-Based Accountability.* http://www.aci.org/docLib/20070716_NealSchqanzenbachPaper.pdf (accessed August 18, 2007).

Perie, M., Marion. S., Gong, B. and Wurtzel, T (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief.* Dover, NH: the National Center for the Improvement of Educational Asssessment Inc.

Perie, M., Marion, S.F., and Gong, B (2007). *Moving Towards a Comprehensive Assessment System: A Framework for Considering Interim Assessment.* Dover, NH: The National Center for the Improvement of Educational Assessment Inc.

Rosenshine, B. (2003). High-stakes testing: Another Analysis. *Educational Policy Analysis Archives 11* (24), http://epaa.asu.edu.epaa/v11n24 (accessed September 8, 2007).

Springer, M. (2008). Accountability Incentives: Do Schools Practice Educational Triage? *Education Next 8* (1) 75-79.

Partnering to Help All Kids Learn® | NWEA.org | 503.624.1951 | 121 NW Everett St., Portland, OR 97209

July 2014

NWEA™