Aligning District and State Assessments to Measure Growth in Achievement

By Jack B. Monpas-Huber, Ph.D.

Why, as depicted by the WASL results in Table 1, do students seem to be less proficient in mathematics after third grade?

Table 1
*WASL Mathematics Proficiency Rates, Grades 3-7, 2006-2008**

	Spokane			State		
	2006	2007	2008	2006	2007	2008
Grade 3	66.7	74.3	75.2	64.2	69.6	68.6
Grade 4	62.4	62.8	60.7	58.9	58.1	53.6
Grade 5	57.9	63.7	69.2	55.8	59.5	61.2
Grade 6	54.0	57.9	55.9	45.9	49.6	49.1
Grade 7	44.4	53.9	52.4	48.5	54.6	50.5

*Values are percents of students meeting or exceeding the state standard.

When Spokane Public Schools examined these results in the spring of 2007 and asked this question, it immediately looked for an *instructional* explanation: In the aggregate, were Spokane's district curriculum and instructional practices less aligned with state standards beyond 3rd grade? Were students not getting the learning experiences they needed to achieve the state standards for math proficiency?

For answers, Spokane turned to data from its own district assessments. Like many districts, Spokane had implemented a system of district interim benchmark assessments designed to measure students' mathematics achievement several times within the year prior to WASL. While many districts use commercially available assessments such as NWEA-MAP, Spokane had committed to developing its own district assessments in order to build assessment capacity and deep understanding of state standards. Consistent with the district's mission of alignment between state and district expectations, these assessments were WASL-like *by design*. They were developed using the WASL test and item specifications. They were, however, smaller than the WASL in order to be administered, scored, and the results reported back quickly. In spite of this difference in size, there was good reason to assume that the district assessments were measuring the same domain of knowledge, skills, and abilities as the WASL and reporting consistent information.

Were the data from these district assessments showing the same apparent decline in math achievement after third grade? If so, why? Were teachers teaching the district curriculum with *fidelity*? And to the extent that they were, was the curriculum *rigorous* enough to adequately move students to the state standards?

Many other districts probably asked similar questions of their district assessments about the effectiveness of their instructional programs. Such questions are fair when the primary purpose of state assessments is to provide feedback to districts and schools about the effectiveness of their instructional systems, and when districts invest in district assessment programs in order to have "multiple measures" of student achievement and rely less on one test given once per year.

Validity of District Assessments

Data from district assessments, combined with data from the state assessment, offer promise for good progress monitoring and program evaluation. However, how districts interpret and use data from multiple measures also raises important issues of validity and technical quality that districts would be wise to consider. Any time educators use any kind of assessment data to make decisions about students, such data need to be valid and reliable so that people can trust that the results are stable and measuring what they are intended to measure (Messick, 1989).

For Spokane, such validity questions about the district assessments were at least as important as the instructional implications of the results. Were

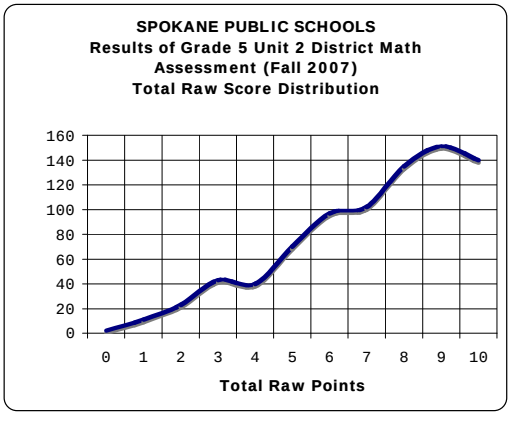
(Continued on next page)...

the district assessments reporting consistent information as WASL, and would they be predictive of WASL performance? The fact that the district assessments and the WASL were developed from the same test maps and item specifications provided strong evidence of *content validity*. But to what extent were they really measuring the same thing? What counts as evidence? At stake was no less than whether the district assessments were reporting true achievement of state standards.

The 5th Grade Mathematics Study

To try to answer some of these questions, Spokane conducted a longitudinal study. The district selected a sample of approximately 1,000 students to study over the course of their fifth grade year and ultimately into their sixth and seventh grade years. The sample was a stratified random sample of 10 students from each classroom to ensure that the sample represented the broader population of fifth grade students. Besides WASL data from grades 4 and 5, the district collected all data from the district math assessments on these fifth grade students. Ultimately, the study amounted to 10 waves of math data on these students. Item-level data were collected in order to examine the qualities of the items as well as the tests.

Figure 1
Results of Spokane District Math Assessment (Fall 2007)



The descriptive results of the study were interesting. The results shown in Figure 1 are typical of the results from the district assessments. Data from the district assessments tended to show negatively skewed distributions in which most students earned near-perfect scores. This pattern of results was good news for several reasons. It suggested that most students were learning the standards that were measured by the assessment. It also suggested that the teachers were teaching the district curriculum and that the tests were sensitive to instruction. However, these results raised other questions and implications. Did a maximum score on the district assessment really mean a student had mastered the standards measured by the assessment and would bring the same ability to the WASL? Was the curriculum rigorous enough? Such questions were cause for serious discussion among Spokane curriculum and assessment personnel.

Gathering Evidence of Construct Validity

How stable were the results of the district assessments? Were they measuring the same constructs as the WASL? At the same time that curriculum specialists were analyzing the results of the district assessments, assessment personnel were analyzing the reliability and validity of the assessments in order to answer questions like these. Two primary issues emerged which other districts that have developed in-house assessments may wish to consider.

One issue is reliability. In classical test theory, reliability is often expressed as a single statistic—Cronbach's coefficient alpha—which is based on test length and redundancy of items. Spokane's district assessments—probably like most locally developed district assessments—were necessarily short in order to administer, score, and report results more quickly. However, shorter tests are in general less reliable. Reliability becomes an issue when the total test score matters for some purpose such as correlation or prediction. A test that does not correlate very strongly with itself will not correlate very strongly with anything else. Lower reliability also means the total test score reflects other factors besides true math ability and will fluctuate if the test is administered multiple times.

A second issue is dimensionality. Classical test theory assumes that a test measures only one dimension, such as math computation. Arguably this

(Continued on next page)...

is less of an issue in reading where students can be asked to perform similar kinds of skills but with more sophistication to comprehend more challenging texts. It is more of an issue in mathematics when students are asked to perform very distinct kinds of operations or use different kinds of content such as algebra and geometry. It was an issue for Spokane's assessments insofar as each district math assessment measured a somewhat different domain of state standards than the previous, and each assessment was designed to measure several state standards. Possibly this was not so uncommon among locally developed district assessments. Dimensionality becomes an issue when one wants to use the total test score to make an inference about student mastery of a particular domain but the total test score reflects several dimensions rather than one and possibly dimensions that were not intended. Factor analysis can be a useful tool for assessing what dimensions the items seem to be measuring *based on the data* rather than the test developer's *a priori* assertions of what the items are measuring. It was not uncommon for factor analyses of Spokane's district assessments to show items clustered on dimensions other than those they were intended to measure based on the test map.

Districts that choose to develop their own in-house districts assessments and are serious about technical quality might therefore want to consider exploring these issues. They may want to consider writing focused tests designed to measure one primary dimension of learning rather than multiple, and using multiple items to measure the same performance expectation rather than one item to measure several different performance expectations.

Linking District Assessments to State Assessment

Another issue to consider with locally developed district assessments is how the assessments are *scaled*. Large-scale assessments such as the WASL are provided a scale in order to report consistent information about difficulty and student ability each year despite inevitable differences in the difficulty of different test forms and student abilities each year. Equal interval scales also facilitate arithmetic operations and statistical analyses.

Like most districts that invest in district assessments to measure state standards, Spokane wanted to make inferences from its assessments about students' performance on the WASL. One challenge that stood in the way was different scales. The WASL used the familiar equal interval scale with 400 as the proficiency standard, while the district assessments used the total raw score. One way to overcome this challenge was to put the district assessments on the same scale as the WASL so that they shared 400 as the same level of proficiency. After taking a district assessment, students could receive a scale score with 400 as WASL proficiency. The question then became how to *link* or *equate* scores from district assessments to WASL so that they share the same scale. This question prompted a review of the literature on scaling and equating (Dorans, Pommerich, & Holland, 2007) which revealed a variety of different approaches.

One promising approach to scaling uses the Rasch model, which makes use of information from the items as well examinee variation in total test scores. Bond and Fox (2001) provide an accessible introduction to the Rasch model and its applications to a variety of measurement issues. This piece includes a chapter on equating scores of the same students (a "common person" design) from two different tests designed to measure the same construct. Districts that choose to develop in-house assessments to emulate the state assessment might want to consider exploring this literature to scale district assessments in their own right or to link them to the state assessment.

Broader Issues for Districts

Districts that invest in district assessment systems for purposes of program evaluation face an important choice between two approaches. One is to use outside instruments such as MAP. The other is to develop in-house district assessments that emulate the state assessment.

(Continued on next page)...

The advantage of locally developed assessments is the development of assessment capacity as teachers become "students of the standards" (Eaker, 2008). Local district personnel develop deep understanding of state standards and how to collect credible evidence of student achievement of those standards. If the assessments are developed according to the same test map and item specifications as the state assessment, they have evidence of *content validity*. However, one challenge of this approach is the time and expense of training and freeing teachers to do this work. Locally developed assessments may also not be subject to the same rigorous validity studies for evidence of technical quality. It may also be difficult to link scores from the district assessment to the state assessment in a clear way.

Commercially available district tests, such as MAP have a different set of issues. A huge advantage of these kinds of assessments is that they likely enjoy the benefit of a very large item bank, which is beneficial in several ways. The items likely enjoy very high quality and have known difficulty and discrimination statistics based on piloting. This makes possible computer adaptive testing which helps provide more reliable measurement of student abilities. Such tests also enjoy the benefit a continuous scale of growth for measuring student achievement across grade levels. However, it may be more difficult for these kinds of instruments to claim *content validity* when they are not developed according to the same test map and item specifications as the state assessment.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Second Edition. Mahwah, NJ: Lawrence Erlbaum.

Dorans, N. J., Pommerich, M., & Holland, P.W. (Eds). (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

Eaker, R. (2008). Statement made in conference session on the role of common assessments in professional learning communities. Seattle, WA.

Messick, S. (1989). Validity. In *Educational Measurement*, R. Linn (Ed.). Washington, DC: American Council on Education.

-Jack B. Monpas-Huber, is Director of Assessment and Student Information Shoreline Public Schools and held similar postions in Spokane and Northshore. He serves as Secretary to the National Association of Test Directors.