Festschrift Papers (III)

Validity Issues for Common District Assessments

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

In recent years, many districts have implemented systems of common district assessments. By this I mean short assessments of students' skills that can be administered, scored, and reported quickly—hence the term "short-cycle" assessments. When administered under standardized conditions, these assessments can provide administrators with frequent estimates of students' status and/or growth toward the state proficiency standard. In these cases, such assessments serve a summative function much like a large-scale assessment such as the WASL. On the other hand, the short cycle of district assessments can provide teachers both timely feedback on instruction and identify students who may need additional help. These assessments can function as formative assessments in this way.

Both purposes of assessment are important. In most districts, accountability pressures create a need for summative data. Districts need to know what instructional programs and interventions are working most effectively. At the same time, districts value information that is timely and actionable to teachers. Further, assessments are expensive to develop and purchase for only one purpose. For all these reasons, it is probably common for districts to use their district assessments for both formative and summative purposes simultaneously.

In this paper, I critically examine these uses of district assessments from a measurement perspective. Using the example of district assessments in Spokane Public Schools, I argue that there are important validity issues behind these different uses of district assessments which districts should consider. In what follows, I first describe Spokane Public Schools' recent experience developing and using common district assessments. I then turn to the measurement literature to provide a brief review of various approaches to validation of districts' inferences and uses of assessments. I invoke Kane's (1992) conceptualization of an "argument-based" approach to validity to outline a strategy for validating district assessments. Using Kane's approach, I outline Spokane's argument for how it uses data from its common district assessments. I surface some of the implicit assumptions behind this argument and then attempt to sketch appropriate sources of validity evidence or validation strategies. In some cases, I can describe validity work that I have already done, while in other cases I describe challenges or barriers to validity work that should be done.

District Assessments in Spokane Public Schools

Several years ago, Spokane Public Schools embarked on a path of developing and implementing a centrally managed district curriculum and assessment system (English, 1988). In Spokane, this is called the "Written-Taught-Tested Curriculum," and its primary purpose is to bring curriculum, instruction, and assessment throughout the district into alignment with state standards. The managed curriculum provides a common district language and framework for instructional action for the district. To the extent that the curriculum contains the state standards, and teachers faithfully teach the district curriculum, the district can assume that all students at each grade level receive a common set of challenging educational experiences consistent with the state's expectations and will master the knowledge, skills, and abilities that will be manifest on the state assessment.

A cornerstone of this district curriculum policy and theory of action is common district assessment, which in theory provides several important pieces of information. They provide frequent reports of student achievement relative to the state standards and the curriculum—how well students are performing, how well *certain groups* of students are performing, *where* student performance is lower than desirable, and *on what state standards* performance seems to be lower than desirable. They also function as an "early warning system" for students who appear to need additional assistance.

Continued from previous page...

District coordinators in each content area have developed common district assessments in each content area for students at each grade level. The district assessments are relatively new and have evolved over the past few years to meet various district needs. These include both formative and summative purposes on which I elaborate in a forthcoming section.

Approaches to Validation

Arguably, all educators want valid and reliable data from assessments. Educators want to believe that they are drawing valid conclusions from assessment data, or that they are using assessment data in valid ways. But what does it mean to validate a district's use of assessments?

Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 16). Evident from this definition, now shared widely among the measurement community, is that the object of validation is not an assessment instrument itself but the primary *inferences* and *uses* of the instrument (Messick, 1989; Test Standards, 1999). Arguably all uses of an assessment rest on inferences about what its scores mean, and not all inferences may enjoy the necessary empirical validity evidence or theoretical rationale. Some tests designed for formative purposes may not be able to support important summative purposes, and vice versa. Other tests intended to be measures of instructional effectiveness may actually not be very sensitive to instruction. Alternatively, claims that a series of local tests or other assessment results (such as grades) measure the same construct as a state assessment may turn out to be indefensible when confronted with disconfirming data.

There are multiple ways of conceptually organizing and embarking on the validation effort. One method is Messick's (1989) four-fold cross-classification of test inferences and uses by evidential and consequential bases. In the same work, Messick (1989) also outlined a variety of types of validity investigations:
We can look at the content of the test in relation to the content of the domain of reference. We can probe the ways in which individuals respond to the items or tasks. We can examine relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses. We can survey relationships of the test scores with other measures and background variables, that is, the test's external structure. We can investigate differences in these test processes and structures over time across groups and settings, and in response to experimental interventions—such as the instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and suing the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects. (p. 16)

Another method is to organize validation around the primary intended *uses* of a test (Shepard, 1993), foregrounding some validity questions and backgrounding others. One might also use the Test Standards (1999) as a guide.

In this paper, I use Kane's "argument-based" approach to organize thinking about validation of test inferences and uses. Kane (1992) suggests that validation might be framed most effectively in terms of a "practical argument" for particular inferences and uses which test users might advance for a test. According to this argument-based approach, test users first articulate as clearly as possible the major premises and claims of their argument for test meaning or use. They then surface the primary assumptions underlying these claims. Having articulated the argument and assumptions, users should then document the available validity evidence, or at least validation strategies, to support these inferences and uses.

The primary advantage of this approach is that it acknowledges the social context in which assessment operates. Educators make claims about student learning based on tests. They make arguments for using tests in some ways, but

Continued from previous page...

not others. They also question the validity of assessment data, based on different conceptions of evidence of student learning. An argument-based approach to validity has the potential to tailor validation to the particular frames of the public in question, such as the distinction between formative and summative uses of assessments which is meaningful to educators as well as somewhat analytically useful. I therefore use this distinction as the framework for validating Spokane's argument for its common district assessments.

Validating Formative Purposes of District Assessments

The purpose of formative assessment is to influence and inform instruction and action (Wiliam, 1998). Indeed, a major appeal of these systems to districts is their ability to provide content-rich assessment information to the district and classrooms on a regular basis in a way that informs next steps.

The district assessments in Spokane Public Schools have a strong formative orientation. By design, the assessments are relatively small (about 10-12 items) so that they can be administered quickly in classrooms, then quickly scored and reported district-wide so that teachers can make use of the information in a timely fashion. The district reports the results in terms of classical item difficulty statistics, disaggregated by school, so that teachers can see which skills have not been mastered by students and therefore where to focus additional instruction. By most accounts, this administration and reporting process is very effective with these assessments providing a valuable common language and metric of student achievement for school collaboration discussions. At the district level, content coordinators examine the data to identify areas of student need. Difficult items can point to areas for improved curriculum or professional development.

Wiliam (1998) suggests that formative and summative inferences of assessments should be validated in different ways. In his view, summative assessment should be validated on the basis of *score meaning* in the classical sense: What do the scores mean? How reliable are they? Do they behave as intended psychometrically? In contrast, the primary purpose of formative assessment is to stimulate and inform action: to inform next instructional steps, to guide instruction to areas of the domain where it is needed, to provide feedback to students, to motivate students to improve. For these reasons, formative assessments should be validated on the extent to which they produce the intended *consequences* (Wiliam, 1998).

Consequential validity evidence. What, then, counts as evidence of consequential validity? In Spokane, there is anecdotal evidence of the positive instructional consequences of the district assessments. The district assessments have helped teachers better understand what is expected of students at the state level, and as a result, teachers have become more consistent, aligned, and focused on the state standards in their instruction. Local data on what students know and are able to do have stimulated discussion among administrators, coaches, and teachers about instructional practice and pedagogy.

This is an important validity issue. Districts that invest in common district assessments for a strongly formative purpose should be clear about the intended instructional consequences of the assessments, and then gather evidence of the extent to which, and how, the assessments are then used to stimulate and inform ongoing midcourse corrections and differentiation in instruction.

Validating Summative Purposes of District Assessments

Arguably, consequential validity should not be only basis for validating common district assessments. Assessments lacking other important aspects of validity may produce fundamentally misleading information even if it is well-received and produces positive consequences. Evaluative pressures in districts may create needs for important summative inferences and uses of district assessments which may be designed for formative purposes. As schools become increasingly accountable for results, they will need to monitor progress of their students toward achievement goals. As far as these results are publicly reported, it will become important to ensure the technical quality of the testing process

Continued from previous page...

and data. Test results will need to be comparable despite year-to-year changes to items. Schools will need to show some measure of gain or growth for continuously enrolled students. It may also be important to have alternate forms of a test of comparable difficulty.

The district assessments in Spokane Public Schools have summative as well as formative purposes. Almost all of the district assessments are "end-of-unit" or "end-of-quarter" assessments administered at the end of instructional periods. Assessments are also administered under semi-standardized conditions (common instrument, predefined testing window, teacher scoring based on rubrics) in order to minimize variation due to administration factors and to facilitate comparison of the effectiveness of different instructional "conditions". More recently, the district has begun moving toward use of the summative district data as evidence of individual student achievement to populate a report card. This section articulates Spokane's argument for the functions of its district assessments as summative assessments of the written and taught curriculum and the validity issues and evidence surrounding them. In what follows, I use this argument as a framework for addressing inherent validity issues surrounding claims of this type and strategies for gathering validity evidence.

Inferences about Mastery of Content and Curriculum

All common district assessments are samples from a larger target domain of content knowledge, skills, and abilities. Most states make some effort to define this domain through frameworks and specification documents (Kolen & Brennan, 2004). Washington State provides a body of domain specification work at its Web site (www.k12.wa.us). This includes Grade Level Expectations (GLEs) documents which delineate exactly what students should know and be able to do at each grade level in each content area (OSPI, 2006). Teachers can internalize the GLEs to guide their own instructional planning, and districts can purchase or develop curriculum within which to embed these important learning objectives.

The "Written-Taught-Tested" Curriculum in Spokane Public Schools is a "theory of action" (Argyris & Schön, 1978) which makes strong claims about the content that its district assessments are measuring. The claim is that district assessments function as direct measures of the district's written curriculum which itself embodies the GLEs. The district's written curriculum takes the form of program guides that outline for teachers what content should be covered within a defined span of time, typically a unit lasting several weeks. These guides make clear the learning objectives—the GLEs embedded within the curriculum that will be taught if the curriculum is followed. The district assessments are designed to assess these learning objectives covered within curriculum units. A related claim is one of *alignment* between the district and state. Alignment to the state assessment system—for the district assessments to be "WASL-like"—is an important reason the district chose to develop its own curriculum and assessments rather than purchase these products from an outside vendor. The content argument is thus that the district assessments validly measure the content that students should know and be able to do within the framework of the curriculum.

These claims rest on various assumptions. One is that the district curriculum adequately captures the target domain of the state standards. Another is that the assessments adequately sample the domain of both the curriculum and the GLEs. The implication is that the district assessments are, to some degree, parallel measures of the state assessment, the WASL. Such claims should prompt a search for supporting evidence. Claims about the content of tests fall within the category of *content validity*. As Messick (1989) put it, "We can look at the content of the test in relation to the content of the domain of reference" (p. 16). Evidence of content validity can be documentation of test content and development procedures. All district assessments are developed according to the WASL item and test specifications (OSPI, 2007). All items in the district assessments are aligned (by expert judgment) to at least one GLE. All district assessments, like the WASL, have a mix of item formats: approximately half multiple-choice and half constructed-response items. The constructed-response items include short answer items (worth two points) and at least one extended response item (worth four points). All district assessments also include scoring guides to guide teachers in their scoring of student work. Arguably, these are good sources of evidence of content validity insofar as they make very clear the content of the assessments, the learning targets that the assessments are intended to measure, and the design and development of the assessments.

Continued from previous page...

Content validity evidence. Content validity is an important aspect of validity for medium-scale common district assessments. Districts that invest in common district assessments want to claim that the assessments are aligned to the state assessment system, that they are measuring the same content knowledge, skills and abilities that will be assessed on the WASL even if they are not strictly parallel forms of the WASL. Content validity may be an important issue especially in the marketplace of formative assessment systems. Outside vendors may claim that their products are aligned to state standards even if their items were not written according to state item specifications or their tests not developed according to state test specifications. This alignment may have been a *post hoc* process of aligning individual items to state standards through expert judgment. Spokane chose to develop its own common assessments specifically to build its own assessment capacity and to develop assessments specifically aligned to the Washington State standards using the WASL test and item specifications. To the extent possible, districts that go down the road of common assessments intended to prepare students for the state assessments should pay close attention to content validity.

Inferences about Student Proficiency, Constructs, and Traits

Claims from test results about what students know and are able to do in relation to a construct or trait is perhaps unavoidable. To ask any district assessment system to provide aggregate measurements of student *status* in relation to some predefined standard of proficiency is perhaps understandable. Such a standard can be the proficiency standard on the annual state assessment or a more proximal, locally-determined proficiency standard arrived at through some form of Angoff-based standard-setting procedure. In Spokane, inferences from test results about student knowledge, skills, and abilities are common. A *de facto* purpose of the district assessments is, as one coordinator put it, "to know where our kids are. WASL results should be no surprise." Another administrator said the purpose of the district assessments is to "fill the gaps between the WASLs." The construct argument is thus that the district assessments measure the same construct(s) as the WASL tests. Again, such claims about districtwide student abilities and achievements on the basis of district assessment data rest on assumptions which should be critically examined.

Correlational evidence. One form of construct validity evidence is correlations between scores from tests believed to be measuring the same construct(s)—what Messick (1989) refers to as the "external structure" of a test. Stronger correlations represent stronger evidence of parallelism and alignment between two tests purported to measure the same or at least very similar constructs. Correlations between scores from district assessments and WASL tests will likely be moderate, which gives rise to several interpretations. First, all correlations below 1.0 provide an opportunity to better understand the concept of measurement error and to temper hopes of perfect prediction. Second, moderate correlations suggest that although the tests share considerable variation, they are measuring somewhat different constructs (Kolen & Brennan, 2004). This makes sense when we consider the nature of the two constructs being measured. Both the WASL and the district assessments are samples from a very large domain. Being a larger test, the WASL represents a larger sample that uses more items to measure the domain. The district assessments measure only the GLEs embedded within curriculum units. Thus, they measure a smaller, more defined domain than the WASL, and they are smaller assessments in which a small number of items are used to measure as many GLEs as possible. As a result, the correlation between the WASL and the district assessments is attenuated by construct underrepresentation (because the construct measured by the district assessments is smaller and more constrained) and restricted range (because the district assessments cannot measure the full range of the construct measured by the WASL).

Convergent/divergent validity evidence. Another construct validation strategy for district assessments would be to explore the convergent and divergent validity through the use of the multitrait multimethod matrix (Campbell & Fiske, 1959; Crehan, 2001). Crehan (2001) used such an approach to examine the convergent and divergent validity of data from a district-developed performance assessment and found limited evidence of validity. Unsettling is the implication that the data provided by those assessments provided misleading information for decisionmaking in that district.

Continued from previous page…

Reliability evidence. Claims about performance on items or tests about what students know or are able to do also rest on the implicit assumption that scores from the sampled items and assessments can be generalized to the target domain without error or bias. The extent to which test scores possess this property is commonly known as *reliability*. The Test Standards (1999) are clear that inferences about student ability on the basis of scores of an educational assessment require some evidence of the reliability of scores from the assessment. Reliability is thus an important issue for district assessments.

In districts where students take a district assessment only once, it will not be possible to estimate reliability by means of test-retest or alternate forms analyses. Instead, one can use measures of internal consistency reliability such as Cronbach's coefficient alpha (Cronbach, 2004). The alpha coefficient is a measure of the extent to which a test is unidimensional based on covariation among items. Strong covariation between items and comparatively small amounts of individual item variation represent evidence that the items collectively are measuring one construct or trait (DeVellis, 2003). A good Cronbach's alpha value is .80 – .90, with .70 being a minimally acceptable value.

Estimates of reliability, like correlations, prompt a search for sources of unreliability, or measurement error. As described above, most district assessments are necessarily *short* (about 10 items) so that they can be administered and scored quickly. However, reliability is generally understood to increase with the number of items or tasks and the average inter-item correlation (DeVellis, 2004). In addition, items are typically written to measure different skills (GLEs). Thus, district tests may be multidimensional, rather than unidimensional, by design. Thus, test size and multidimensionality by design may place a ceiling on internal consistency reliability. Another potential source of error is inter-rater disagreement in the scoring of the open-ended items. Anecdotal evidence of variation in teachers' application of the scoring rubrics for their students' open-ended responses abounds. However, in my observations, open-ended items typically enjoy the strongest item-total correlations.

This finding suggests that open-ended items do a reasonably good job of discriminating examinees on the basis of achievement (contributing true variation) despite any inter-rater disagreement (error variation) that may exist.

Reliability estimates carry implications for summative inferences about student's level of achievement. Low reliability estimates may suggest that a large proportion of the observed variation in the scores is due to random error, or *noise*—or at least variation due to individual items that is unrelated to the primary trait being measured. In individual terms, this means that a student's observed total test score may lay at some variance from his or her true test score. In other words, low reliability produces unstable test scores. This becomes a problem when districts begin to make important decisions about students on the basis of these test scores, such as the assignment of a summative grade that will be reported publicly and become part of the student's permanent record. Low reliability will produce misclassifications. Students with test scores that overestimate their true achievement will receive higher grades, and students with test scores that underestimate their true achievement will receive lower grades. In addition, low reliability limits correlations with other measures (Carmines & Zeller, 1979). Some students who receive high marks in a content area based on district test scores will score below standard on the state assessment, and vice versa. Such results would be cause to temper strong claims about alignment with the state system.

New measurement research offers new ways of thinking about internal consistency indicated by Cronbach's alpha. Willett (1988) suggests that correlation-based reliability estimates might obscure important dynamics of student growth. Reliability describes the extent to which two measures produce the same rank ordering of examinees. However, that can be misleading when examinees are growing (both in negative and positive directions). By this thinking, low reliability estimates may be evidence of considerable intra-individual growth. Cronbach (2004) himself suggested that the alpha coefficient is less appropriate for the kind of mixed-format performance assessments currently in use today which are multidimensional by design. Marzano (2000) makes a similar argument in his application of measurement theory to

Continued from previous page...

formative classroom-level assessment. He suggests that most teachers rarely design classroom assessments to measure only one trait or construct. Thus, internal consistency may not be the most useful way to think about the reliability of these assessments.

These are important issues for district assessments. Consumers of common district assessments at all levels want to claim that their tests validly and reliably measure the psychological construct(s) they are intended to measure. Such claims may be tenuous without the kinds of evidence of reliability and construct validity outlined above. Minimal or weak evidence of validity and reliability in these areas may be cause to temper such claims.

Inferences about Growth in Student Achievement

Related to the issue of student performance status is *growth* in student achievement. Districts that invest in common district assessments, especially when those assessments measure the same students in a content area several times a year, might reasonably desire some form of information about growth in achievement. At the district or school level, pressure to improve performance may place a premium on data that could show students growing toward proficiency. Many administrators would like to use some measure of growth to more rigorously evaluate the effectiveness of various instructional programs and treatments (Lissitz, 2006), while many teachers would like to see growth over time in their students' learning—especially in regard to the state proficiency standards—as a result of instruction. Anecdotal evidence of these desires abounds. The appeal is understandable. Students in a grade level are assessed on their proficiency in a content area several times a year.

It is tempting to ask: Where are the students in relation to the state standards? How many are on track to pass the upcoming state assessment?

Fortunately, these desires for growth data come at a time when the supply of expertise and knowledge in this area is growing. An explosion of empirical work is currently happening in the area of growth research and longitudinal data analysis (Lissitz, 2006; Lloyd, 2007; Singer & Willett, 2003), and this work carries powerful implications for schools' efforts to measure and determine what can be done to cause all students to reach state standards. One immediate and important implication might be to provide educators with a more precise understanding of growth and the technical requirements for valid growth inferences. Measurement researchers restrict their use of the term "growth" to data that meet two important specifications: (a) the same examinees are observed on the same construct on repeated occasions; and (b) all measurements of examinees fall along a continuous score scale (Kolen & Brennan, 2004; Lissitz, 2006; Singer & Willett, 2003; Willett, 1988). In what follows, I discuss each of these requirements and efforts to provide valid growth inferences in the context of Spokane Public Schools' common district assessments.

The same examinees are observed on repeated occasions. A significant threat to observing the same examinees over time is mobility. Student mobility is a serious challenge in many schools. In Spokane, students move around considerably, both within and to beyond the district. In recent years, the district has not had a mechanism for collecting data for every student in the district, so it has used sampling methods. The district has collected district assessment data by means of independent random samples by which each of 100 teachers receives a new random list of students whose work is requested for central data collection. While this design had the advantage of a producing a large representative district-wide sample (stratified by classroom), it did not provide repeated measurements for the same examinees. One solution to this problem is to centrally select one district-wide random sample at the beginning of the year and follow it as a *panel* over time. This may not be a serious challenge in districts that have the resources or data collection mechanism to collect data from every student.

Continued from previous page...

All measurements of examinees fall along a continuous score scale (Kolen & Brennan, 2004; Lissitz, 2006; Singer & Willett, 2003; Willett, 1988). This is an important measurement issue which is not always well understood in education where growth has been defined broadly enough to include different kinds of analyses.

Consider a personal example: My five-year-old son stands up against a wall and I measure his height as 40 inches. Several months later, we repeat the procedure and find that his height now exceeds 40 inches. That is literally true growth in height along one continuous scale that spans the entire range of the construct of height. The implication for educational assessments would therefore be tests of varying difficulty, given at different ages, sharing one continuous scale. This is called a *vertical scale* which makes possible inferences about *absolute* growth.

Now consider a common alternative: My son is a student in a fifth grade class which receives a month of math instruction and then takes an end-of-unit assessment of the math content covered within the unit. My son scores at the mean of the distribution. His class then receives another month of instruction and then takes a second end-of-unit assessment of equal size and format to the first but which covers somewhat different material. My son scores at the 85th percentile. Can we reasonably infer that he has "grown" in math achievement? The two tests represent different rulers of achievement. We can much more validly infer that my son has grown *relative* to his classmates than we can infer that he has grown in math skill in any absolute way along any hypothetical continuous scale.

From the perspective of growth inferences, the district assessments in Spokane conform more to the latter example than to the former. Within a content area, the district assessments are discrete measurement points that capture what students are expected to have learned within the latest curriculum unit. Each unit, especially in mathematics, may focus on different content. Each assessment within a content area also has a slightly different number of total possible raw points. As a result, the assessments are like different rulers of different length and calibration.

However, it may be still be possible to "link" these assessments together on a continuous scale of proficiency in a content area. There are two issues. One is the construction of the continuous scale of achievement based on the district assessments. Once the same students in a grade level are observed on each of the district assessments, it may be possible to use IRT scaling techniques with single-group common person equating (Bond & Fox, 2007; Yu & Popp, 2005) to determine the extent to which the items from the different district assessments form a continuous scale of content area achievement. Such a common score scale for the district assessments could provide a foundation for educators to observe growth in the same students' content area achievement over the course of one year.

The other issue is to locate the WASL state proficiency standard on this locally developed continuous scale. What does a particular score on a district assessment, or some combination of performances on the district assessments, mean in relation to scores on the WASL? Is the WASL a more difficult assessment than the district assessments? An easier assessment? Or about the same? If WASL results included item parameters, it might be possible to use a single-group common person concurrent calibration strategy to construct a scale that includes the state proficiency standard. Then the district assessments would be on the same scale as the WASL and would provide better information throughout the year about where students are in relation to the state standard. However, the WASL results do not include item parameters, and so a weaker alternative may be to employ some kind of linking strategy (Kolen & Brennan, 2004) based on only total scores rather than items.

Growth research could have considerable practical implications for curriculum and instruction. If it were possible to give the actual WASL test on the first day of school, what would the distribution look like? How many students would already be at standard? How many students would be close, and farther away? What assumptions would be challenged by this result? How well does a linear model really represent the learning process? The ability to measure growth more precisely through the use of a continuous scale could stimulate considerable discussion about teaching and learning.

Continued from previous page...

Discussion

The purpose of this paper was to raise a series of validity issues for common district assessments whose use is increasing in public education. Worth repeating is that not all uses and inferences of common district assessments may be valid. However, complicating the matter, as this paper has tried to suggest, is the mix of conflicting pressures for both formative and summative purposes that converge at the district level. Clearly, small, frequent tests should be able to provide useful information to teachers. However, small tests developed for formative purposes may not be able to support important summative inferences and uses. To complicate matters, purposes of assessments may evolve over time in response to changing organizational priorities. Possibly district assessments themselves, like the state assessment, may evolve over time in response to changing needs.

Districts that choose to invest in common district assessments would do well to think through some of these validity issues and consider gathering appropriate validity evidence.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Argyris, C., & Schön, D. A. (1978). *Organizational learning: A theory of action perspective*. Reading, MA: Addison-Wesley Publishing Company.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage Publications.

Crehan, K. D. (2001). An investigation of the validity of scores on locally developed performance measures in a school assessment program. *Educational and Psychological Measurement*, 61(5), 841-848.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. (CST Technical Report 643). Los Angeles, CA: Center for the Study of Evaluation.

DeVellis, R. F. (2003). *Scale development: theory and applications*. 2nd ed. Thousand Oaks, CA: Sage Publications.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

English, F. (1988). *Curriculum auditing*. Lancaster, PA: Technomic Publishing Co., Inc.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. Second Edition. Springer.

Lissitz, R. (Ed.). (2006). *Longitudinal and value-added models for student performance*. Maple Grove, MN: JAM Press.

Lloyd, J. E. V. (2007). On the quantitative analysis of individual change: Unpacking the meaning of "change" and "commensurability". Manuscript submitted for publication.

Marzano, R. J. (2007). Applying the theory on measurement of change to formative classroom assessment. Retrieved November 10, 2007, from http://www.marzanoandassociates.com/html/resources.htm#papers

Marzano, R. J. (2000). Analyzing two assumptions underlying the scoring of classroom assessments. Retrieved November 10, 2007, from http://www.marzanoandassociates.com/html/resources.htm#papers

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications.

Continued from previous page...

Office of the Superintendent of Public Instruction. (2007). *Test and item specifications for the Washington Assessment of Student Learning (WASL)*. Retrieved on November 17, 2007 from http://www.k12.wa.us/assessment/WASL/testspec.aspx.

Office of the Superintendent of Public Instruction. (2006, September). *Mathematics K-10 grade level expectations: A new level of specificity*.

Popham, J. (1987). Measurement-driven instruction. *Phi Delta Kappan*.

Shavelson, R. J., Gao, X., & Baxter, G. P. (1993). Sampling variability of performance assessments. (CSE Technical Report 361). Los Angeles: Center for the Study of Evaluation.

Singer, J. S., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press.

Wiliam, D. (1998). The validity of teachers' assessments. Paper presented Working Group 6 (Research on the Psychology of Mathematics Teacher Development) of the 22nd annual conference of the International Group for the Psychology of Mathematics Education, Stellenbosch, South Africa.

Willett, J. B. (1988). Questions and answers about the measurement of change. In E. Rothkopf (Ed.), *Review of research in education* (1988-89) (pp. 345-422). Washington, DC: American Educational Research Association.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Yu, C. H., & Popp, S. E. O. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment, Research & Evaluation* 10(4). Retrieved on November 18, 2007 from http://pareonline.net/getvn.asp?v=10&n=4

-Jack B. Monpas-Huber, *Assessment & Program Evaluation, Spokane Public Schools.*
Please direct correspondence to: Jack B. Monpas-Huber, Ph.D., Director of Assessment and Program Evaluation, Spokane Public Schools, 200 North Bernard Street, Spokane, Washington, 99201. E-mail: JackM@spokaneschools.org