Developing Measurement Theory for Classroom Assessment Purposes and Uses

Susan M. Brookhart, Helena, MT

In many fields of inquiry, the need for new theoretical developments is often best seen in areas of strain, and strain is apparent in several areas in which the conventions of measurement theory do not quite "fit" classroom assessment. Three areas of strain are analyzed in order to suggest how theoretical development might focus directly on information quality in the classroom assessment context. This article describes the context dependence of classroom assessment, its inextricable relationship with instruction, and its simultaneous formative and summative functions. Thus a case is made for new theoretical developments in the area of measurement in the classroom.

Keywords: classroom assessment, reliability, validity

Most of the information that students have about their learning—and what that means about the subject, about themselves, and about their futures—comes from classroom assessments. Similarly, most of what parents and teachers know about their children's learning comes from classroom assessments. Current discussions about classroom assessment quality use validity and reliability concepts developed for large-scale testing contexts and adapt them for classroom assessment. The purpose of this article, and this entire special issue, is to conceptualize how we might develop "indigenous" measurement theory thinking directly about classroom assessment purposes and uses instead of borrowing theory developed in the context of large-scale assessment. Scates (1943) foreshadowed some of the issues addressed here by contrasting the large-scale and classroom assessment contexts of 60 years ago. Huge advances in the theory behind large-scale assessments have been made in the last 60 years—not so for classroom assessments. The time has come to develop measurement theory for classroom assessment purposes and uses.

The need for new theoretical development is often best seen in areas of strain, and strain is apparent in areas in which the conventions and assumptions of measurement theory do not quite "fit" classroom assessment. For example, both classical and modern test theory for large-scale assessment consider context a source of irrelevant variance; the aim is to generalize across contexts. In contrast, for classroom assessment, items or tasks are dependent on, and nested within, the instructional environment. For another example, in large-scale assessment, items or tasks are usually assumed to be independent, whereas in classroom assessment they are linked together in students' classroom experiences. For yet another example, large-scale assessment theories usually make assumptions that require large sample sizes. Class sizes are small in comparison, and sometimes classroom assessments are given to even smaller subgroups or to individuals within classes.

This article describes the context dependence of classroom assessment, its inextricable relationship with instruction, and its simultaneous formative and summative functions. Analyzing these three important areas of strain will help illustrate some of the theoretical work that needs to be done in the area of classroom assessment. These three areas are not an exhaustive list of the theoretical "misfit" between large-scale and classroom assessment; they are simply three good places to start.

Shepard (2001) suggested that methods of testing, and the psychometric methodologies developed for studying them, developed in the intellectual context of the early 20th century. The hereditarian theory of intelligence and behaviorist learning theory, the curriculum of social efficiency, and scientific measurement shaped the assumptions and choices made in the development of the educational measurement as a field. Large-scale testing and "scientific" measurement provided an excellent context for some things, including the development of mathematical models for both classical and modern test theory that have become important to our understanding of how measurements "work." Psychometric theory grew, and continues to grow, and an impressive number of technical problems have been solved (Brennan, 2001). As in any field, but particularly in a highly applied field like educational measurement, additional problems continue to present themselves, of course—but the basics have been laid down. Among the important and useful concepts that have been developed are validity and reliability (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999; Feldt & Brennan, 1989; Messick, 1989; Moss, 1992; Shepard, 1993).

Susan M. Brookhart is a Consultant, 2502 Gold Rush Avenue, Helena, MT 59601; e-mail: brookhart@duq.edu. Her areas of specialization are classroom assessment and educational evaluation.
Historically, these developments in psychometrics that have been forged in large-scale assessment contexts have been subsequently applied to classroom assessment. This work has been important in understanding classroom assessments to date, but it has reached a point of limited returns. I think, for example, that in addition to the much lamented lack of “training” for teachers in measurement theory (Schafer & Lissitz, 1987; Stiggins, 1999; Wise, 1999), another important reason why teachers do not use sound measurement principles in their work is the lack of fit between the large-scale context, for which validity and reliability principles have been developed, and the measurement theory needs of the classroom. Stiggins (2001) approaches this problem by using the term “high-quality classroom assessment” and avoiding the terms “validity” and “reliability” in materials intended for teachers. This approach works well as an adaptation mechanism for inserting more measurement theory into teacher education, but it does not afford a way to advance measurement theory itself in the classroom context. It is time to focus directly on the classroom assessment context in its own right as a milieu for “doing” measurement theory. This should help us develop more useful ideas about classroom assessment information and information quality. This article is divided into two sections. First, I make the case that the classroom assessment context is fundamentally different from the measurement contexts that gave rise to classical and modern test theory in ways that affect the way we think about measurement. Second, using those contextual differences as a basis, I make some suggestions for new approaches to developing measurement theory for classroom assessment. The three additional articles in this issue amplify some of these suggestions and begin some of the theoretical work that needs to be done. These suggestions are not exhaustive and are not meant in any way to downplay the importance of large-scale psychometrics—indeed, for the very concepts of validity and reliability we are indebted to it—but rather to point out the misfit between the context of large-scale psychometrics and classroom assessment and to suggest more fitting and potentially more productive avenues of theoretical development for the field of classroom assessment.

Classroom Assessment Purposes and Uses

Shepard (2001) pointed out that cognitive and constructivist learning theories and reformed visions of curriculum and classroom assessment implied fundamental transformations for classroom assessments: that their content and character must be improved and that the use of classroom assessment information must become an integral part of the learning process. While I believe that Shepard (2001) is quite right, there is evidence that this and other constructivist transformations have already been accepted by the teaching profession and are already occurring in many classrooms. For example, Andrade (2000) described how teachers can work with students to create instructional rubrics that enhance student learning and thinking about their own work. Benson (2000) described how she created a “partnership class” with a high level of formative evaluation and student self-management. The areas of theoretical strain—such as the context, item, and sampling examples above—still operate even in these “improved” classroom assessment settings. Thus, constructivist transformations alone are not going to ring in the day when classroom assessment will be theoretically “sound” if judged against current concepts of validity and reliability.

We need to take another step and articulate how, even for improved classroom assessment environments, “doing” reliability and validity work must be fundamentally different. This theoretical work must acknowledge the range of classroom assessment environments, from very student-centered ones like Andrade’s and Benson’s to very teacher-centered ones. Theory about the validity and reliability of assessment information in the classroom must be able to deal with the meaning of information in its classroom context, whatever that is.

The following three sections briefly document the context considerations that I propose are the most fundamental to understanding the need for a measurement theory developed with classroom assessment purposes and uses in mind: (1) that in classroom assessment there is a psychosocial context, (2) that classroom assessment and instruction are integrated, and (3) that classroom assessment is primarily formative. These three principles operate in both student- and teacher-centered classrooms. Even in a teacher-centered classroom, the student’s point of view matters because of its effect on learning. From the student’s point of view, classroom assessment information is not merely information “about” himself or herself. Rather, it forms a major part of his or her learning life, becoming part of the lessons he or she is expected to learn, the relationship he or she has with the teacher and the subject matter, and relationships with peers (Black & Wiliam, 1998; Crooks, 1989).

The Classroom Assessment Environment

Classroom assessment occurs within a classroom environment or context or climate. The context affects the assessment; for example, instructional experiences affect how both language and tasks are interpreted and how expectations are set. Stiggins and Conklin (1992) originally described the “classroom assessment environment” in terms of teacher practices. Important to creating a classroom assessment environment were the purposes for which teachers used classroom assessments; the assessment methods used, the criteria for selecting them, and their quality; the teacher’s use of feedback; the teacher’s preparation and background in assessment; the teacher’s perceptions of students; and the assessment policy environment. We have heard a lot about the assessment policy environment recently, most notably in articles about the impact of large-scale assessment adversely affecting what teachers do in the classroom (Popham, 1999), although there are some dissenting voices (Cizek, 2001). In either case, though, the interest in effects on teacher practices highlights the importance of those practices. Stiggins and Conklin’s (1992) thesis that a teacher’s classes have an assessment “character” or environment that stems from the teacher’s general approach to assessment is now widely accepted. Haydel, Oescher, and Kirby (1998) called teachers’ beliefs about testing and assessment practices “the evaluative culture of classrooms,” and found that these beliefs were related to teacher efficacy. Students contribute to the classroom assessment environment, too. Students may be involved in the design and interpretation of assessments (Stiggins, 2001). But even in rather traditional, teacher-centered classrooms, student perceptions are inextricably tied to the classroom assessment experience (Brookhart, 1997) and ultimately the
The difference between decontextualized, one-time measurement done by outsiders and the ongoing, contextualized measurement done with a teacher changes the nature of the measure. Information from large-scale assessments is experienced as external to “school,” whereas the scores, grades, feedback, and other classroom information fold back into and become part of the classroom environment.

The Integration of Assessment and Instruction

Good classroom assessments are not only measures of learning, but genuine episodes of learning themselves (Wolf, 1998). The more performance-based the assessments, the more the line between assessment and instruction blurs (Moss, 1994). Students have different amounts of help and guidance with assignments done as part of instruction. Currently, measurement theory—specifically validity and reliability—is not designed to apply to the assessment/instruction mix. In particular, the idea that students may have access to teachers’ or peers’ help with projects, so that the project grade applies not strictly to an outcome or product but to the “studenting” process leading up to that product, is a problem for the assumptions of current measurement theory that arose out of the external, large-scale assessment context. Help and coaching that students receive for assessments conducted during and as part of instruction would be seen as contributing to lack of independence of observations or “cheating” and thus a threat to reliability and validity. However, one of the basics in constructivism is that this is precisely where learning occurs—in the zone of proximal development (Vygotsky, 1978), that space between what the individual can accomplish independently and what he or she can do with assistance.

The approach often taken by measurement theorists currently is to separate assessment artificially from instruction when they think about classroom assessment. For example, a unit test or final project may be assumed to be an after-learning checkup that can be separated conceptually from the instruction that preceded it. This is not, in fact, consistent with classroom reality (Airasian, 2001), even for end-of-unit tests, much less for projects that constitute most of the instructional period. Currently, recommendations are to make assessment even more integrated with instruction (Nitzko, 2001; Shepard, 2001). This integration makes the test a part of the environment itself and a part of the fabric of students’ learning. This is fundamentally different from large-scale assessments of achievement and conventional psychometrics, where the major function of the assessment is to provide someone, typically someone besides the student, with interpretable information. Even in the best-case scenario for standardized testing programs, the students’ instructional life—and most other aspects of their lives—pause while students concentrate on demonstrating achievement for assessors they do not know.

Contrast this with the case of classroom assessment, where the teachers who planned and implemented students’ lessons are the ones who wrote or selected the assessments. Emphases in instruction are reflected in the assessments, as are particular approaches or interests that were explored in class. Thus students have a “relationship” with the assessments that they do not have with external assessments and will have a “relationship” with the results as well.

Formative and Summative Assessment Purposes

Scriven (1967, pp. 40–43) used the terms “formative” and “summative” evaluation to differentiate between the two roles that evaluation may play in education. Using evaluation in the development or improvement of some educational process is “formative”; using evaluation in decision making about the end result of an educational process is “summative.”

Several authors (Black, 1998; Black & William, 1998; Gipps, 1994; Sadler, 1983, 1989; William, 1998) have described the formative assessment process as a cyclical one. The students’ formative task is to compare ideal and actual performance, act to close the gap, compare ideal and improved performance, and so on. If students are to improve, three things must happen. First, students must develop a concept of their learning goal. The goal originally is the teacher’s, but with learning the student will internalize it. This growing concept of what “good work” is forms part of the learning itself. Second, students must develop the ability to monitor their work and compare actual with desired performance. Third, students must develop the ability to act in such a way as to close the gap, which involves setting their own learning goals. Teachers are responsible for providing the feedback that students need to do this, and teachers’ provision of that feedback itself is sometimes called “formative assessment.”

Whether the teacher uses student-centered methods or not, the student occupies a central role in this formative process because only he or she can take the actions necessary to improve. The teacher cannot learn “for” the student. The formative assessment process itself is a part of the learning because it is in this process that students come to understand the standard of quality work with which their work is compared. This differs from summative assessment, in which the assessor is the only one who necessarily has to understand the standard, and it changes the relationship between the measurers and the measured. “Summative assessment requires that teachers (or other assessors) become members of a community of practice, while formative assessment requires that the learners become members of the same community of practice” (William, 1998, italics in original).

British authors tend to call classroom assessment “formative” and external assessment “summative” (Black & William, 1998; Gipps, 1994). In the United States, in the context of grading as currently practiced, it is reasonable to call many classroom assessments “formative” as well. Any assessment information that is used in the calculation of a report card grade goes home to parents and is stored in school records as a judgment of student achievement that can be superseded, but never erased. Using formative classroom assessment for summative purposes can create tension for the relationship between teacher and student. Gipps (1994) considered two different summative processes. “Summing-up” means creating a picture of achievement based on accumulating assessments that...
were originally formative, as for Records of Achievement in the United Kingdom. “Checking-up” means tests or tasks at the end of learning, assigned specifically to collect information for summative judgments, as for final exams in U.S. high schools or “unit tests” in classrooms. In either case, students would participate in these classroom assessments knowing that their performance will not only inform their own future work, but also will inform parents and teachers who will use that information to make judgments and draw conclusions about them.

Some argue whether it is even possible to have a “summing-up” process that uses information originally intended as formative assessment for a summative purpose. Students often pay less attention to feedback and more to the grade or score that “counts” in the final grade and thus learn less from the feedback than they might otherwise (Crooks, 1988; Sadler, 1989), and this is a major issue in the area of classroom grading practices. Others argue that summative assessment, such as a test at the end of a teaching episode or unit, can have positive effects if it is aligned with instruction and “deeply criterion-referenced, incorporating the intended curriculum, which should be clearly salient in the perceived assessment demands” (Biggs, 1998, p. 107).

Students can use information from classroom assessments to compare their performance with their concept of good performance and make some inferences about what improvements they need to make in their work. For several reasons, all classroom assessment can be formative. Students see classroom assessment as a demonstration of what they were “supposed to learn.” They recognize the language of the assessment as the language of instruction. Therefore, they can understand what the results of the assessment mean for them. Successful students even use in formative ways information from end-of-unit tests on material they may not revisit in class (Brookhart, 2001). This cyclical formative process, in which the student is not the “subject” whose achievement is measured but an active participant in the process, such that the meaning of the assessment information itself cannot be fully understood apart from the student who is using it, distinguishes the classroom assessment context from the large-scale psychometric context.

**Summary**

Three ways of looking at classroom assessment have been discussed: the classroom assessment environment, the integration of assessment and instruction, and the pervasive formative purpose of classroom assessment. Each is an aspect of the central issue that in classroom assessment the nature of the measurement information itself is different enough from the nature of the large-scale information that has traditionally been the object of psychometric study to warrant theoretical consideration in its own right. In the classroom, an “environment” exists in a different sense from the “administration” environment of large-scale tests. Assessment must be integrated with instruction, which implies that the meaning of the items or assessment tasks will be dependent on the environment. All classroom assessment is formative, even that which is also used summatively.

Taken together, these factors render the measure itself literally a part of who the student is in the classroom. As Thomas and Oldfather (1997) quote a student: “That’s my grade. That’s me.” This is terribly serious for students in classrooms and has all sorts of implications for future learning choices, self-worth, and a host of other cognitive, affective, and conative understandings (Covington, 1992). Classroom assessment information and uses become a part of the daily realities of the classroom and can lead to both positive and negative consequences.

**A Measurement Theory for Classroom Assessment**

Because classroom assessment differs from other assessment in the ways documented in the previous section, it is reasonable to ask what measurement theory work might look like if we started with the classroom assessment context in mind. Table 1 organizes these thoughts by contrasting some validity and reliability demands in large-scale and classroom assessment. The two-column format emphasizes the contrasts between the two but is not meant to suggest a complete disconnect. Most of these issues, I think, are more accurately seen as two ends of a spectrum. Obviously the issues framed under classroom assessment are expressed in language developed for large-scale assessment. The intent of this contrast is to illuminate areas of strain, where measurement theory developed for the large-scale assessment end of the spectrum fits poorly with the classroom assessment side, thus arguing the need for some development at that end.

The following sections present an overview of these issues. There may well be other issues that follow from the heavy influence of context, integration of assessment and instruction, and formative nature of classroom assessment, but I see these as some of the most fundamental. The goal of this section is for readers to begin to see legitimate approaches to developing “classrometric” measurement theory. Some of the ideas are explored in more depth in the articles to follow. All of the authors join me in challenging the measurement community to work toward the depth of understanding of classroom assessment information that large-scale measures now enjoy.

**The Measure and the Measured**

The first comparison in Table 1 describes the difference in the nature of the relationship between the measure and the measured described in the section on classroom context and formative assessment. Psychometric theory was developed to ensure the quality of assessment information when the measure provides information about an attribute of an object, event, or person. Validity describes the extent to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests (AERA, APA, & NCME, 1999). Reliability describes the stability of measures over factors that should be irrelevant to the measure (time, forms, raters, etc.). Implied in these conceptions is a sense that the measure itself is in some ways external to the inferences or actions. The test functions as a “dipstick” into the “oil tank” of the student’s achievement. The validity goal is a meaningful inference about student performance and/or effective use of that information for a specified purpose. Purposes to which achievement information from large-scale assessment are put include selection, placement, and guidance decisions for students and accountability reporting for schools and districts.

“Classrometric” theory should take into account that classroom assessments provide information about students that immediately becomes part of their learning environment and their own psychology (Covington, 1992; Thomas
### Table 1. Contrasting Large-Scale and Classroom Assessment Concepts

<table>
<thead>
<tr>
<th>Concepts in Large-Scale Assessment</th>
<th>Concepts in Classroom Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity: The measure is external to the inferences made and actions taken.</td>
<td>Validity: Inferences made and actions taken are internal to the measurement process.</td>
</tr>
<tr>
<td>*.: Students are “subjects” upon whom observations are made.</td>
<td>*.: Students are observers jointly with teachers; “those measured” make the inferences and take the actions in the formative assessment process.</td>
</tr>
<tr>
<td>*.: The validity goal is a meaningful inference about student performance and/or effective use of that information for a specified purpose.</td>
<td>*.: Students’ awareness of and benefit from assessment information are part of the “information” itself.</td>
</tr>
<tr>
<td>Validity: The measurement context is construct-irrelevant.</td>
<td>Validity: The measurement context is construct-relevant.</td>
</tr>
<tr>
<td>*.: Content specifications describe a domain.</td>
<td>*.: Assessment is part of instruction. A good assessment is an “episode of genuine learning.”</td>
</tr>
<tr>
<td>*.: Administration can be standardized.</td>
<td>*.: Content specifications reflect both the domain (learning objectives) and instruction (modes, activities).</td>
</tr>
<tr>
<td>*.: Scores can be equated or linked across contexts and forms of assessment.</td>
<td>*.: Teacher beliefs, teacher instructional practices, and teacher understanding of both the subject matter and students (including cultural and linguistic differences) are relevant validity concerns.</td>
</tr>
<tr>
<td>Reliability is consistency over irrelevant factors.</td>
<td>Reliability is sufficiency of information.</td>
</tr>
<tr>
<td>*.: Occasions, time, items and/or tasks are facets of error variance.</td>
<td>*.: The reliability goal is stable information about the gap between students’ work and “ideal” work (as defined in students’ and teachers’ learning objectives).</td>
</tr>
<tr>
<td>*.: The reliability goal is stable ranking of students on a score scale (NRT) or stable categorization of students along an achievement continuum (CRT).</td>
<td></td>
</tr>
</tbody>
</table>

& Oldfather, 1997). Validity is still the extent to which inferences and actions based on test scores are appropriate and warranted, but the main actions of interest are relatively immediate, internal changes in the students who are measured (see Moss, this issue). The two most common classroom assessment purposes are: (1) checking on students’ achievement of specific instructional goals (which is summative if used for the purpose of assigning a grade, but is also used formatively); and (2) formative assessment in the ongoing instructional process during daily classroom activities (for adjusting and continuing instruction). In the preceding section, it was shown that classroom assessment purposes are dependent on their context, especially the assessment environment and the instruction in which both students and teachers participate, and that the students’ formative understanding of results shapes the meaning of the assessment information. Indeed, the questions or tasks and how they are posed become part of how the students conceptualize their learning targets (Sadler, 1983, 1989; Stiggins, 2001). If that is so, then inferences made and actions taken are internal to the measurement process. It is the students who are “measured,” but it is also students who make the inferences and take the actions in the formative assessment process. Therefore, students’ awareness of and benefit from assessment information are part of the “information” itself.

Similarly, since assessment information will be conveyed through teacher feedback, the teachers are also internal to the measurement process itself. Their feedback, as well as the score or grade, is part of the assessment information. Underlying all discussions of formative assessment are “assumptions about the psychology of learning” (Black & William, 1998, p. 16). Giving feedback assumes an understanding of what causes errors. The validity of a formative assessment is determined in part by the validity of the model of learning on which the teacher’s feedback is based (Black, 1998, p. 123). Teachers’ feedback varies in valence (positive or negative, focusing on strengths or shortcomings) and purpose (evaluative or descriptive), and therefore feedback varies in formative usefulness to the student (Tunstall & Gipps, 1996). Usefulness of feedback, in turn, is part of the consequential evidence for the validity of the classroom assessment information (Messick, 1989). Sometimes, the evidence of a student’s work speaks directly to the student in the absence of—or in spite of—teacher feedback.

The validity goal in classroom assessment becomes understanding the role of that assessment information—grade, score, teacher oral and written comments, student observations of his or her errors or wonderful moments—in the ongoing classroom learning environment. What did the performance mean to the student? Did the student accurately interpret his or her performance? Was the communication from the teacher clear; that is, did the student understand the grade and/or teacher feedback in the way in which she intended? Was that feedback helpful? Did the student pay attention to it? Did the student trust the information? How helpful was the information from the assessment for fostering future learning or continued progress toward goals? Students actually would be the best judges of this if they were taught how to reflect on it (Sadler, 1989; Stiggins, 2001).
These validity questions about post-assessment use of information depend in part on some validity questions that are more relevant to planning the assessment. Content validity evidence is already noted to be of particular importance for classroom assessments (Airasian, 2001; Nitko, 2001). Evaluating the content of a classroom assessment includes evaluating not only the match to instructional objectives, but also to classroom instruction. It may also include a consideration of what happened during that instruction. For example, a teacher may put on a test a problem of the sort that students found particularly troubling, so they can demonstrate that they can do it. Or the teacher may avoid such a problem so as not to trip up students unnecessarily. Each approach to the content of the test may be valid for some purposes but not for others. Notice that these content considerations require the input and judgment of the teacher, because it is the teacher who would understand what happened during instruction.

Garcia (1994) wrote that the equity challenge for classroom assessment is different from that for large-scale performance-based assessment. The major difference is that the sensibilities, understanding and interpreting of cultural meanings, and knowing what topics and situations to ask about and what modes of response would be idiomatic shift from the teacher (for classroom assessment) to the test developers (for large-scale assessment). This is not a trivial difference in the validity requirements.

Moss (this issue) looks at validity theory in light of some of these issues. She points out that the very job of a teacher is to make interpretations of students’ capabilities at any point in time obsolete by continuing to support their learning and development. She begins a description of a theoretical stance that sees valid information in the classroom as information that teachers and students find useful to support the learning process.

**Construct-Relevant and Construct-Irrelevant Variance**

For standardized achievement tests, what counts as construct-relevant variance can be defined by a table of specifications that describe a domain in terms of subject matter and cognitive level. The measurement context is construct-irrelevant, which is one of the reasons that administration must be standardized. The test is supposed to mean the same thing for fifth graders in North Carolina as it means for fifth graders in Alaska. Scores can be equated or linked across contexts and forms of assessment.

But we have just claimed that in classroom assessment, the assessment is inextricably linked with instruction. The meaning of the questions or tasks depends on the instruction, so the same question might mean something different in two different classes. For example, a math test question asking for a particular proof might require original application for students in one mathematics class and recall of yesterday’s chalkboard work in another. In addition, questions or tasks are supposed to be part of the instruction, because students are supposed to be learning from taking the test or doing the assignment. Content specifications for classroom assessments should reflect both the domain as set forth in the instructional objectives (“what the student should know and be able to do” after the lessons) and also the modes of instruction.

This makes the measurement context itself “construct-relevant.” Teacher beliefs, teacher instructional practices, and teacher understanding of both the subject matter and students are relevant validity concerns. Validity information is sought that answers the question of how well the score or grade indicates achievement of the construct(s) represented by the instructional goals. But the instructional goals are not just pages in the district’s curriculum. Instructional goals are heavily influenced by teacher beliefs about the subject matter and about what constitutes appropriate instruction (McMillan, this issue).

Particularly important for evaluating the validity of a classroom assessment is defining the construct in its instructional and contextual sense. Is writing part of the skill or irrelevant? Did we discuss this in class? Is this problem the same as one the students have already worked (in which case it measures recall)? Is this problem similar to ones the students have already worked (in which case it measures transfer)? A close transfer or a stretch? How far did you want them to go?

Shepard (2001) says that deeper understanding is the underpinning for cognitive transfer and proposed richer, more authentic, and problem-solving tasks for classroom assessment. Parkes (2001) suggests that cognitive transfer is part of the person-by-task variability problem that has been documented for performance assessments. Conventional psychometrics would address this problem by standardizing the tasks, narrowing the domain, or adding tasks (“Spearman-Brown style thinking,” Parkes, 2001, p. 145). These conventional approaches are clearly not compatible with classroom assessment purposes. If we start by thinking in the classroom assessment context, the amount of transfer of learning is relevant to the classroom instruction of which the assessments are an integral part. Teachers lead students to explore applications more widely or narrowly in their instruction. Parkes (2001, p. 155) points out that in conventional psychometrics, aiding student transfer with deliberate strategies like the use of concept maps would not be valid if transfer were part of the construct. A classroom-centered approach might ask, “Why not?” If the integration of assessment and instruction is taken seriously, having the test or assignment become a vehicle for transfer of learning—an episode of learning itself where the students experience the next leap in their ability to apply knowledge—might be an intended consequence.

Much of what constructivists propose to assist with learning is an attempt to develop what (for want of a better term) I could call “group intelligence.” Current classroom instructional methods include teaching questioning techniques, scaffolded instruction, reciprocal teaching, and other methods generally described as “cognitive apprenticeship.” These strategies were developed in the 20th century as constructivism gained ground (Shepard, 2001) and as schools were called upon to be more and more inclusive (Covington, 1992).

What all these cognitive apprenticeship methods have in common is that they externalize the internal abstracting and understanding functions of successful learners. Thus even if a cognitive transfer method, such as teaching students how to use concept maps or outlines, makes a difference, it has contributed to learning. Classroom assessments are the measure of that, and the children themselves are at once the agents of the change and the thing that is changed. If assessment is viewed as formative, “in classrooms where assessment practices are equitable, teachers realize that there are no negotiable answers because assessment is a process of finding out what knowledge a child has. Errors can then be viewed as windows on a child’s development rather
than mistakes that must always be corrected.” (Gooden, 1997, p. 29).

Other validity questions that are directly related to the classroom assessment’s place in the classroom instructional environment are legion. How is the information used (and how useful is it) for teacher decisions? Does the assessment lead to sound and productive decisions? What about the use and usability of information for student decisions, studying, and future learning? Can students use the information for self-evaluation? How? Do they do that? What about problem solving and higher order thinking? What provision for effort is made in the interpretation of classroom assessment information? What about contextual factors—compatibility with group setting, group administration, students helping each other, and so forth? Moss (this issue) writes that because classroom assessment is internal to the teaching-learning process—designed and conducted by the teachers (and sometimes students) themselves—and is used for multiple purposes, validity criteria should emphasize whether and how the assessment contributes successfully to teaching and learning and how well it serves the multiple purposes and audiences.

Reliability and Error

In conventional psychometrics, reliability is consistency over irrelevant factors. Occasions, time, items, and/or tasks are facets of error variance. The reliability goal is stable ranking of students on a score scale for norm-referenced scoring or stable categorization of students along an achievement continuum for criterion-referenced scoring. Shepard (2001) wrote that classroom assessments do not need to be “as reliable” as large-scale assessments because errors in judgment on one day might be corrected with the addition of more information on the next. I submit that, while classroom assessments will not be “as reliable” as large-scale tests by the definition of reliability above, this is more an issue of misapplying the reliability goals of large-scale testing to classroom assessment than it is an issue of poor-quality classroom assessments. Daily classroom assessment needs to be of high quality, because errors in judgment on one day will inform instructional decisions on that very day. By the next day an instructional opportunity will have been missed—or worse, if the teacher has acted on misinformation, damage may have been done. However, the reliability may not be well gauged by methods developed for large-scale assessments, and by these measures even good-quality classroom assessment information may seem to be “unreliable.” If the classroom purpose is formative and situated, then the reliability goal is stable information about the gap between students’ work and “ideal” work, as defined in the teachers’ instructional objectives and in the students’ concepts of the achievement targets for which they are aiming in a particular assignment—which may or may not be the same thing! Reliable rank ordering of students would not be the criterion against which scores should be judged. Reliable categorization of students onto a developmental continuum of quality of work and accurate diagnostic (and thus often multidimensional) information about specific shortcomings are more important. The student will use information about that work to compare current achievement with ideal and plan for future work. The students’ conception of ideal work or “good” work will grow, too, as they get better at doing it and as their conceptions deepen. Smith (this volume) proposes a reliability judgment of sufficiency of information for these purposes.

Summary and Overview of Special Issue

Readers should not interpret the articles in this issue as an attack on large-scale psychometrics, which has given us the basis on which we are building. Clearly there is a relationship between psychometric theory and what I have dubbed “classrometric” theory. This special issue emphasizes the need to work directly on psychometric theory within the classroom assessment context. Instead of always beginning with the large-scale context and “applying” or transferring to the classroom assessment context—therefore systematically underemphasizing some things (like the formative feedback loop) and overemphasizing others (like sample size)—it is time to take what we know about how to “do” measurement theory and work intentionally with the classroom context. It is time, and it is important. We have a much better chance of being relevant and of learning useful things if we study assessment in its habitat. Working in the classroom context, we may even discover some aspects of measurement theory that we can apply back to the large-scale context.

In this issue, Moss looks at what kinds of validity considerations regarding classroom assessments might support the work teachers and students do in the classroom. She finds interpretive practices for drawing and warranting conclusions appropriate when standardization is not appropriate, as in most classroom applications. Smith distinguishes among the reliability concerns for classroom assessments, course grades, grade point averages, and class rank. He proposes sufficiency of information as the relevant reliability principle for classroom assessments. McMillan focuses on teachers’ classroom assessment decision making. He describes the tension between teachers’ internal beliefs and values and external influences and the implications of this tension for classroom assessment theory and practice.

The authors of this set of articles all hope that this special issue helps change the direction of theoretical thinking about classroom assessment. We believe this shift is timely. Classroom assessment has become recognized recently as an important topic (Shepard, 2001; Stiggins, 2001). The role of the classroom teacher is acknowledged in the current large-scale testing context (Crocker, 2003). The next challenge is to value and elevate the status of classroom assessment and to understand why it needs original theoretical work instead of always receiving secondary application of theoretical work from large-scale assessment—applications whose assumptions are strained in classroom assessment and therefore applications which by definition consign classroom assessment to be “second-class” assessment. “Classrometric” thinking will help with the development of educational measurement as a discipline and should raise the quality of information in the classroom. Thus while our main audience for this special issue is the measurement community who will do this theoretical work, practical improvements will logically follow. The quality of information available in the classroom is one of the greatest resources for teaching and learning.

References


