November 1996

Educational policy analysis archives

Arizona State University

University of South Florida

Education Policy Analysis Archives

Volume 4 Number 17 November 11, 1996 ISSN 1068-2341

A peer-reviewed scholarly electronic journal.

Editor: Gene V Glass,Glass@ASU.EDU. College of Education,
Arizona State University,Tempe AZ 85287-2411

What Does the Psychometrician's Classroom Look Like?: Reframing Assessment Concepts in the Context of Learning

Catherine S. Taylor
University of Washington

ctaylor@u.washington.edu

Susan Bobbitt Nolen
University of Washington

sunolen@u.washington.edu

Abstract

We question the utility of traditional conceptualizations of validity and reliability, developed in the context of large scale, external testing, and the psychology of individual differences, for the context of the classroom. We compare traditional views of validity and reliability to alternate frameworks that situate these constructs in teachers' work in classrooms. We describe how we used these frameworks to design an assessment course for preservice teachers, and present data that suggest students in the redesigned course not only saw the course as more valuable in their work as teachers, but developed deeper understandings of validity and reliability than did their counterparts in a traditional tests and measurement course. We close by discussing the implications of these data for the teaching of assessment, and for the use and interpretation of classroom assessment data for purposes of local and state accountability.

More than ever before, pressure is being placed on teachers to create high quality assessments of their students' learning. Work is underway in Kentucky, New Mexico, Vermont, Washington, and in the eighteen states that are members of the New Standards Project (Resnick and Resnick, 1991) to explore the viability of classroom-based assessments, projects, and portfolios as sources of state or national accountability data about student learning. These initiatives emerge from a growing belief that the teacher can be one of the best sources of

information about student learning. However, there is growing evidence that teachers have not been adequately prepared to create and conduct valid assessments. Even teacher education programs that include an assessment course may not help teachers develop the concepts and skills necessary to meet these assessment demands.

To address this problem, districts, states, and national organizations have invested considerable resources in in-service training for teachers. Organizations such as the National Council on Measurement in Education (NCME) and the Association for Curriculum Development and Supervision (ASCD) have developed training modules and training materials for classroom teachers. Groups such as the National Council for Teachers of Mathematics (NCTM) have developed documents such as *Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions* (NCTM, 1991) and *Assessment Standards for School Mathematics Standards* (NCTM, 1995) in an attempt to help teachers incorporate more appropriate assessments into their teaching practices. Still, these efforts may not be successful if the models used to educate teachers in the concepts and skills of assessment do not fit the reality of classrooms.

An example of the confusion caused by the mismatch between models based on test theory and the demands of the classroom context illustrates this problem. Preservice teachers in an assessment class had read Smith's (1991) article on the meanings of test preparation. Smith lists a number of ways teachers prepare for external standardized tests, including teaching the specific content covered on the test. Students were surprised to find that psychometricians considered this to be cheating. Were they not being admonished, by both the instructor and the course textbook to do just that--assess to see whether students were learning what had been taught. To these students, if vocabulary words were to be tested, they should be taught. If science or social studies concepts and facts were to be tested, they should be taught. Even if the test expected students to generalize a concept or skill to the new situation, the concept or skill should have been taught first! In the words of one puzzled student, "What does the psychometrician's classroom look like?"

This apparent discrepancy between the idea of "domain sampling" central to test theory and the notion that classroom assessment is intended to assess whether students learn what they are taught arises from a clash of contexts. The world of large scale external tests is very different from the world of the classroom. In this paper, we will argue that traditional tests and measurement courses and most assessment textbooks for teachers present measurement concepts in ways that better fit the world of external tests designed to measure individual differences. When teachers are taught traditional measurement concepts and expected to apply them to the context of teaching and learning, they have little chance of developing the skills and concepts they need to assess their students. We will also argue that the meanings of assessment in the context of the classroom must be considered carefully when large scale assessment programs decide to use classroom assessments for the purposes of district, state, or national accountability.

We begin by challenging traditional notions of testing and measurement in terms of their fit to the classroom. While we recognize that the principles of classical test theory may be appropriate for some contexts (e.g., administering and interpreting standardized, norm-referenced tests), we see a need for more clarity in how these models, their applications and limitations, are presented to teachers. We discuss the theoretical underpinnings of traditional measurement concepts and why they must be reframed in light of the classroom context. We examine the ways in which reliability and validity are presented in eight recently published assessment texts designed for teacher preparation and discuss why definitions of validity and reliability presented in most educational assessment textbooks fit the context of external testing better than that of the classroom.

Next we present frameworks for validity and reliability that situate these constructs in the world of the classroom teacher, and discuss how these frameworks might be used in teacher

education. We then present an overview of the assessment course we developed to help preservice teachers understand the concepts of validity and reliability as they are reframed in this paper. The work of the course was designed to help preservice teachers develop a deep understanding of the potential relationship between classroom assessment practices, subject-area disciplines, and instructional methods so that they would see valid and reliable assessment as central to their work as teachers. Evidence for the effectiveness of basing our assessment course on these frameworks is provided in the form of three studies comparing the responses of students in the redesigned course to those taking a traditional tests and measurement course in the same teacher education program.

We discuss the need for the measurement community to acknowledge the differences between the methods appropriate for external measurements and the measurement of the learning targeted by classrooms and schools. We suggest that those who prepare assessment text-books for the preparation of teachers, as well as instructors of assessment courses, clarify the philosophical positions underlying different assessment purposes and present assessment concepts in ways that are consistent with those differing purposes rather than attempting to blend frameworks that come from different philosophies about the purposes of assessment. Finally we discuss what these classroom-based conceptions of reliability and validity suggest in terms of what constitutes appropriate classroom-based evidence for large scale assessment programs.

The Misfit of the Measurement Paradigm

The classroom context is one of fairly constant formal and informal assessment (Airasian, 1993; Stiggins, Faires- Conklin, & Bridgeford, 1986). However, few teacher preparation programs provide adequate training for the wide array of assessment strategies used by teachers (Schafer & Lissitz, 1987, Stiggins & Bridgeford, 1988). Further teachers do not perceive the information learned in traditional tests and measurement courses to be relevant to their tasks as classroom teachers (Gullickson, 1993; Schafer & Lissitz, 1987; Stiggins & Faires-Conklin, 1988). Wise, Lukin, and Roos (1991) found that teachers do not believe they have the training needed to meet the demands of classroom assessment. At the same time, teachers' ability to develop appropriate classroom-based assessments is seen as one of the six core functions of teachers (Gullickson, 1986).

Several authors have outlined what they believe are the essential understandings about assessment teachers must have in order to confront the ongoing assessment demands in the typical classroom (Airasian, 1991; Linn, 1990; Schafer, 1991; Stiggins, 1991). Many of these concepts and skills, as well as those presented in measurement text-books (e.g., Hanna, 1993; Linn & Gronlund, 1995; Mehrens & Lehmann, 1991; Nitko, 1996; Oosterhof, 1996; Salvia & Ysseldyke, 1995; Worthen, Borg, & White, 1993), are derived from a model of measurement that began in the late 1800s. Rooted in scientific thinking of the nineteenth century, test theory is based on a model of the scientific method.

With classroom instruction as the equivalent of a treatment, test theory would suggest that tools of assessment are designed to carefully assess the success of instruction for different examinees. Taking the perspective of Galton (1889), students differ in their inherent capacity to learn the content of various disciplines. The assessor is the scientist who must dispassionately assess and record each students' attainment of the defined outcomes of instruction. Students are the focus of observation and the measurement model presumes them to behave like passive objects. As Cronbach (1970) noted,

> A distinction between standardized and unstandardized procedures grew up in the early days of testing. Every laboratory in those days had its own method of measuring. . . and it was difficult to compare results from different laboratories. . .

> Standardization attempts to overcome these problems. A standardized test is one in which the procedure, apparatus, and scoring have been fixed so that precisely the same testing procedures can be followed at different times and places. . . If standardization of the test is fully effective, a man will earn very nearly the same score no matter who tests him or where. (pp. 26-27, italics added)

The classroom teacher, however, is not a dispassionate observer of students' learning. Classroom teachers have a vested interest in the outcomes of instruction--many believing that student failure is a reflection on their teaching. Both the popular press and current legislation in states such as Kentucky would suggest that the public agrees with this view of the relationship between teaching and learning. The classroom teacher, in contrast to the experimental scientist, is more like a "participant observer" (Whyte, 1943). Using the words of Vidich and Lyman (1994), the teacher is much like an ethnographic researcher. In the following quote, the authors' use of the term "ethnographic researcher" has been replaced by the term "teacher."

> The [teacher] enters the world from which he or she is methodologically required to have become detached and displaced. . . . [T]his [teacher] begins work as a self-defined newcomer to the habitat and life world of his or her [students]. He or she is a citizen-scholar as well as a participant observer. (Vidich & Lyman, 1994, p. 41)

Teachers adjust instruction for the needs of students; adapt instruction for the needs of diverse students; bring a wide range of evidence to bear on decision-making about students - extending beyond the evidence from standardized tests to observations of students' classroom behaviors, attitudes, interests, and motivations (Airasian, 1994). The purpose of classroom assessment is to find out whether students have benefited from instruction. However, unlike the dispassionate observer, the good teacher regularly adjusts the treatment, in response to ongoing assessments, in order for learning to be successful.

While the participant observer may be required to use certain methods to increase their "objectivity," they must both observe and participate in the world of the classroom. They "make their observations within a mediated framework, that is, a framework of symbols and cultural meanings given to them by those aspects of their life histories that they bring to the observational setting" (Vidich & Lyman, 1994, p. 24). The teacher's decision to attend to one source of assessment information over another reveals as much about the "value-laden interests" of the teacher as it does about the subject of her/his assessments (Vidich & Lyman, 1994, p. 25).

While this may be seen by measurement professionals as the reason objective measures are needed, qualitative researchers would respond that "The more you function as a member of the everyday world of the researched, the more you risk losing the eye of the uninvolved outsider; yet, the more you participate, *the greater your opportunity to learn.*" (Glesne & Peshkin, 1992, p. 40, italics added). Qualitative researchers would say that the very choice of what items to include in a test reflects the values and biases of the teacher. Hence the job of those who prepare teachers for classroom assessment must include an awareness of the context in which teachers teach, the goals of instruction and schooling, and the complex demands of the work of a participant observer.

If teachers are not dispassionate observers, neither are students passive objects. They are influenced by assessment processes and products (Bricklin & Bricklin, 1967; Butler, 1987; Covington & Beery, 1976; Deci & Ryan, 1987). They adapt their approach to learning and preparation for assessment in order to gain the highest possible scores (Toom, 1993). They may take on persona that will afford them the grace of teachers. Hence, neither teachers nor students fit the scientific model of standardized measurement used to frame the measurement concepts and strategies taught to teachers.

Assessment and Teacher Preparation Programs

Despite the importance of assessment in the experience of students and in teachers' ability to determine the success of instruction in terms of student learning, assessment instruction is peripheral in many teacher education programs. In programs that do include assessment courses, assessment is usually treated as a foundational course focused on a set of generalizable concepts and skills. In most programs, all prospective teachers, from the kindergarten teacher, to the APP calculus teacher, to the middle school vocal music teacher are taught in a single group. In others, assessment instruction is relegated to a 1-2 week unit in an omnibus educational psychology course. In response to the formidable range of assessment content teachers need to know, instructors may design courses that result in intellectual awareness of key concepts rather than actual competency in applying. Research on the professional development of teachers (e.g., Cohen & Ball, 1990; Grossman, 1991) suggests that intellectual awareness is not sufficient to overcome the "apprenticeship of observation" (Lortie, 1975) that dominates pre-service teachers' learning. Without significant intervention, pre-service teachers typically adopt the practices that were used with them as students or those that are used by their cooperating teachers.

Assessment textbooks generally reflect a view of assessment courses as survey courses, intended to present a range of assessment ideas and leaving to instructors (or the students themselves) the task of constructing a coherent picture of assessment. As Anderson, et al (1995) have noted, survey approaches to the preparation of teachers do not allow for a "rich and grounded" understanding. Ironically, textbook authors' attempts to acknowledge the classroom context may contribute to teachers' confusion and antipathy. Many textbooks (e.g., Hanna, 1993; Linn & Gronlund, 1995; Mehrens & Lehmann, 1991; Oosterhof, 1996; Salvia & Ysseldyke, 1995; Worthen, Borg, & White, 1993) combine presentations of assessment in the classroom with traditional presentations of the principles of testing and basic concepts of measurement. As we will argue in the next section, the notions of validity and reliability used in large scale external testing must be recast before they can be useful in the context of classroom teaching and learning. With the increased emphasis on appropriate assessment practices in the classroom, we must take seriously the gulf between what classroom teachers believe they need to know about assessment and what measurement professionals believe teachers need to know. In the next sections, we provide frameworks for bridging this gulf.

Definitions of Validity

Traditional Presentations of Validity

All of the assessment text books reviewed for this article acknowledged the contextual issues in the classroom; however, chapters on validity generally used the language of scientific methodology to describe this construct. Most of these texts (e.g., Hanna, 1993; Linn & Gronlund, 1995; Nitko, 1996; Salvia & Ysseldyke, 1995; Oosterhof, 1996; Worthen, Borg, & White, 1993) presented three or four "types" of validity: construct validity, content validity, criterion- related (predictive and/or concurrent) validity, and recommend that evidence for each type of validity should be obtained when using a test. Measurement professionals generally agree that for assessments to be valid, they should (a) measure the construct they are intended to measure, (b) measure the content taught, (c) predict students' performance on subsequent assessments, and (d) provide information that is consistent with other, related sources of information. Consequences of test interpretation and use, a validity issue recently raised by Messick (1989), is addressed by few published classroom assessment texts (For example, see Hanna, 1993; Linn & Gronlund, 1995; Nitko, 1996). In fact, some would disagree that "consequential validity" is a component of

the construct of validity at all (See Stuck, 1995).

Traditional presentations of these types of validity often define evidence for validity in terms of: (a) correlations between tests measuring the same construct or between a test and the criterion behavior of interest (Hanna, 1993; Linn & Gronlund, 1995; Nitko, 1996; Worthen, Borg, & White, 1993), (b) tables of specification to determine whether the content of a test measures the breadth of content targeted (Linn & Gronlund, 1995; Mehrens & Lehmann, 1991; Oosterhof, 1996), and (c) using a range of strategies to build a logical case for the relationship between scores from the assessment and the construct the assessment is intended to measure (Linn & Gronlund, 1995; Nitko, 1996; Oosterhof, 1996).

These types of validity evidence are based on two different notions of what makes an assessment valid. The evidence for the validity of an assessment is provided if (a) students perform consistently across different measures of the same construct (a notion that comes from a theory of individual differences (Galton, 1889)) and (b) links between what is measured and some framework or context external to the test (Linn & Gronlund, 1995; Messick, 1989). Taken individually, these two prongs of validity theory do not have equal value in the classroom. Classroom teachers are less interested in the consistency of student performance across similar measures than they are in whether students' learn what they are teaching (the targeted constructs). Learning, especially of skills and strategies that are taught throughout schooling, is expected to change rather than remain consistent over time.

Consistency with other, related performances is also problematic for teachers as they teach each new group of students. Given the option of looking over prior school records, teachers often claim that they do not want to be prejudiced by others' views (Airasian, 1991, p. 54). Over the course of a year, inconsistent performance may be attributed to many factors other than the validity of assessments. Students who begin to perform more poorly than expected may be informally assessed through interviews with the students and reviews of their work. Teachers may become alarmed and contact school support staff and/or parents to see if the cause lies outside the classroom. On the other hand, when poorly performing students begin to dramatically improve performance, teachers may see this as evidence of student learning and of their own success as teachers. Consistent performance across assessments is only desirable when performance is consistently good or when the content taught is constantly changing (e.g., spelling lists).

As Moss's (1996) paper suggests, the notion of the assessor as "objective observer" does not fit the context of the educational assessment as well as it does the work of experimental science. Teachers see students as the focus of purposeful action (Bloom, Madaus, & Hastings, 1981). Tests and other assessments provide information, not only about how well students have learned, but about how well they are presenting the targeted content and concepts (Airasian, 1993; Mehrens & Lehmann, 1991; Nitko, 1996; Oosterhof, 1996), how students are feeling about school, themselves, and their worlds (Airasian, 1993). Hence it is the responsibility of measurement professionals to help teachers learn how to choose and create assessment tools that will do the best job possible to make appropriate decisions about students' learning. This requires teachers to have a clear notion of validity that fits the work and the world view of teachers.

Validity in the Classroom Context

In this section, we situate Messick's (1989) dimensions of validity in the context of classroom teachers' decision-making. Messick claimed that construct validity is the core issue in assessment, and stated that all inferences based upon, and uses of, assessment information require evidence that supports the inferences drawn between test performance and the construct an assessment is intended to measure.

> 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 the 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 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 . . . interventions such as instructional . . . treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. (Messick, 1989, p. 16)

Validity, then, is a multidimensional construct that resides, not in tests, but in the relationship between any assessment and its context (including the instructional practices and the examinee), the construct it is to measure, and the consequences of its interpretation and use. Translated to the classroom, this means that validity encompasses (a) how assessments draw out the learning, (b) how assessments fit with the educational context and instructional strategies used, and (c) what occurs as a result of assessments including the full range of outcomes from feedback, grading, and placement, to students' self concepts and behaviors, to students' constructions about the subject disciplines.

Messick stated that multiple sources of evidence are needed to investigate the validity of assessments. In the classroom context, this means that teachers must know how to look at their own assessments and assessment plans for evidence of their validity, they must know where to look for alternative explanations of student performance, and they must consider the consequences of assessment choices on their students and themselves. In short, teachers should develop a "habit of mind" related to their assessment processes. After situating each dimension in the context of teachers' work, we suggest general approaches that assessment instructor might use to help teachers use that dimension in their own assessment practice.

Validity Dimension 1: Looking at the content of the assessment in relation to the content of the domain of reference. Before teachers can look at their assessments in this way, they must be able to think clearly about their disciplines, understanding both the substantive structure (critical knowledge and concepts) and the syntactic structure (essential processes) of the disciplines they teach (Schwab, 1978). They must be able to determine which concepts and processes are most important and which are least important in order to adequately reflect the breadth and depth of the discipline in their teaching and assessments. As Messick (1989) states, one of the greatest sources of construct invalidity is over- or under-representation of some dimension of the construct. Once they have clearly conceptualized the disciplines they teach, teachers must know how to ascertain the degree to which the types of assessment tasks used in the classroom are representative of the range and relative importance of the concepts, skills, and thinking characteristic of subject disciplines.

In addition, because the process of assessment is as much a function of how assessments are scored as it is a function of whether the tasks elicit student learning related to the structure of the discipline, teachers must examine the degree to which the rules for scoring assessments and strategies for summarizing grades reflect the targeted learnings. As with breadth and depth of coverage within assessments, teachers must be able to evaluate whether scoring rules give too little or too much value to certain skills, concepts, and knowledge leading to questions about the validity of the interpretations teachers make from resulting scores.

To obtain evidence for this dimension of validity, teachers can be taught to stand back from their teaching, frame the learning targets of instruction carefully, and plan instruction and assessment together, in light of the overall targets of instruction. Without a clear picture of what

is to be accomplished in a course or subject area, teachers cannot adequately assess whether their assessments (selected or self-developed) are valid. Once teachers develop a framework of learning targets (learning goals and objectives), they can learn how to carefully analyze whether assessment and instructional decisions link back to this framework. They can be given opportunities to look at scoring rules developed for open-ended student work and determine whether these rules relate directly to these targets of learning.

Validity Dimension 2: Probing the ways in which individuals respond to the items or tasks and examining the relationships among responses to the tasks and items. Teachers do not often have the luxury of "item tryouts" when developing their assessments. Before giving students an assessment, teachers must examine the degree to which the assessments have the potential to elicit the learning the students are expected to achieve. This means they must examine the assessment tasks and task directions to determine whether students are really being asked to show the learning related to the targets. Teachers must know to ask themselves, "Have the directions for the task or the wording of the items limited my students' understanding of the expectations of the task?"

Teachers should be encouraged to use assessment strategies that will allow them to probe their students' thinking and processes. This becomes increasingly important as the emphasis on higher level thinking and processes increases (Stiggins, Griswold, & Wikelund, 1989). In performance assessments, for example, examinees are often asked to explain their thinking and reasoning as part of the assessment task. Teachers commonly ask students to show their work in mathematics and science assessments. These classroom assessment practices lend themselves to probing the ways in which individuals are responding. This probing not only provides information about the validity of the assessments, but can provide better pictures of students' learning.

Teachers must know how to look across students' responses to a variety of assessment tasks to determine whether patterns of students' responses support the use of the assessments. The mechanisms for this type of examination have historically been quantitative item analysis techniques. However, few teachers use these quantitative techniques in actual classroom practice (Stiggins & Faires-Conklin, 1988). Teachers can be shown how to scrutinize student work qualitatively, looking for patterns in responses that reveal positive and negative information about the assessments. If items and tasks have not yet been used with students, teachers must know how to examine the demands of a range of items and tasks and ask themselves, "Are students who can show understanding of a concept in one assessment format (e.g., an essay), likely to show equal understanding in a different format (e.g., a multiple-choice test)?"

In order to probe examinee performance within and across different measures, teachers can learn to develop multiple measures of the same targeted learning. They may not only discover different ways to assess a given construct, but they may discover for themselves that particular types of assessment are more or less suited to certain learning targets.

Validity Dimension 3: Investigating differences in assessment processes and structures over time, across groups and settings, in response to instructional interventions. To investigate these validity issues, teachers must know how to examine the relationship between the instructional practices used and the assessments themselves. They must ask themselves, "Did I or will I actually teach these concepts well enough for students to perform well?" They must also evaluate the adequacy of various assessment strategies for the unique needs of their students. They must be able to judge whether an assessment can be used in many different contexts or whether differing contexts, groups, and instructional strategies require the development of different assessments.

Examination of this dimension of validity can be obtained when teachers are asked to look carefully at the relationship between an instructional plan and the demands of an assessment. If the work demanded in an assessment was not an adequate focus of instruction, teachers can

decide ahead of time whether to adjust instruction to fit the learning targeted in the assessment or whether to adjust assessments to fit the learning targeted in the instruction.

Validity Dimension 4: Surveying relationships between assessments and other measures or background variables. Teachers must know how to judge the degree to which performance on the assessment and the score resulting from the assessment are directly attributable to the targeted learning. They must determine whether performance is influenced by factors irrelevant to the targeted learning such as assessment format, response mode, gender, or language of origin. This becomes increasingly critical as classrooms become more diverse and whole group teaching becomes more difficult. In general terms, teachers must know how to adapt an assessment format to meet the needs of diverse students while still obtaining good evidence about student learning related to the targets of instruction. Finally, teachers must know how to create scoring mechanisms for open-ended performances that are clearly related to the learning targets and that are precise enough to prevent biased scoring.

When teachers develop assessments, they can be asked to examine whether factors other than the targeted learning will influence students' performances. They can be asked to examine scoring rules to see whether the rules provide an unfair advantage or disadvantage to students who have certain strengths or weaknesses unrelated to the targeted learning.

Validity Dimension 5: Tracing the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. Teachers must consider the influence of classroom assessments on the learners themselves. The nature of the assessments, feedback, and grading can all influence student learning, students' self concepts and motivation (Butler & Nisan, 1986; Covington & Omelich, 1984), and their perceptions of the disciplines being taught. Teachers who assess their students' knowledge of science by giving them only multiple-choice tests of isolated facts, for example, communicate that science is a collection of facts about which everyone agrees. Those who assess students' inquiry strategies and their ability to make generalizations from observations or to systematically test their own hypotheses, communicate something different about the structure of the discipline of science.

To examine this dimension of validity, teachers can be asked to assess whether a given assessment reflects the syntactic and/or substantive structure of the discipline they teach (Schwab, 1978). Does the assessment target students' deep understanding of important concepts within the discipline or does it test surface knowledge? Does the assessment ask students to show their ability to use the processes through which professionals within the discipline construct new knowledge and ideas?

Teacher also can be asked to determine whether methods used to summarize grades for a marking period give adequate weight to those performances most directly related to the learning targeted. Teachers can be asked to look at their methods of feedback (formative assessments) and determine whether they are likely to motivate learning or to stifle learning; to assess whether feedback will lead to improvement, be largely insubstantial (Sommers, 1991), or be perceived by students as too late to make a difference in their grades (Canady, & Hotchkiss, 1989).

The five dimensions of validity described here can be taught in ways that emphasize their importance and usefulness in teachers' everyday work. Later we will briefly describe a course designed for this purpose and present evidence of its effectiveness. We recognize, however, that validity rests, in part, on teachers' ability to gather reliable information about student learning. Traditional presentations of reliability, based on test theory, are not immediately transferable to the work teachers do. In the next section, we describe traditional treatments of reliability in assessment textbooks and present an alternative framework.

Dimensions of Reliability

Traditional Presentations of Reliability

Measurement professionals place most of their emphasis in assessment on reliability--often at the expense of the validity of assessments. A common claim in test theory is that "for an inference from a test to be valid, or truthful, the test first must be reliable." (Mehrens & Lehmann, 1991, p. 265). This assumption is based on a mathematical model of test theory wherein observed scores are composed of true scores and measurement error. The less error in a test (i.e., the more reliable) the more truthful the test score. Hence, an unreliable assessment is automatically less valid.

Textbooks usually discuss reliability in terms of consistency (Airasian, 1993; Hanna, 1993; Linn & Gronlund, 1995; Mehrens & Lehmann, 1991; Nitko, 1996; Oosterhof, 1996; Salvia & Ysseldyke, 1995; Worthen, Borg, & White, 1993). When gathering evidence for the reliability of tests, the focus on consistency is related to either score reliability or rater reliability. Score reliability means that if a test were administered to an examinee a second time, the examinee would receive the same or about the same score. One way that measurement specialists try to ensure score reliability is through the standardization of tests. When assessments are standardized, all examinees complete the same items and/or tasks. If examinees are retested, they should complete the exact same tasks under exactly the same conditions. This would help to ensure that consistency of performance.

Another element of score reliability discussed in textbooks is that of generalizability. The longer the test (the more items and tasks) the more opportunities students have to show their learning. If students do better than they should on one item or task, they are just as likely to do more poorly than they should on another item or task. If a test is long enough, positive measurement error should cancel negative measurement error. Hence, the student is likely to earn a score that would be replicated if s/he took a parallel test. Writers who have expanded their discussion of reliability to include performance-based assessments focus on the number of performances necessary to obtain scores for examinees that can be generalized to the domain of interest (Linn & Burton, 1994).

Discussions of reliability in many textbooks; however, are based on the notion that assessment takes place at a single time and that summary decisions are made about examinees based on single testing events. In the classroom, teachers are engaged in on-going assessment over time and across many dimensions of behavior (Airasian, 1993; Stiggins, Faires-Conklin, & Bridgeford, 1986). Like motivation researchers, teachers see giving students choices about assignments as a way to increase student motivation and engagement (Deci & Ryan, 1985; Nicholls, 1989; Nicholls & Nolen, 1993). While individualization of instruction may result in better achievement and motivation, it means that standardization is very difficult. In addition, few teachers have the time or the inclination to administer parallel test forms to see whether students' scores are consistent; and psychometric techniques developed for looking at internal consistency of exams are not appropriate for many forms of classroom assessment. Some teachers give students opportunities to revise their work after feedback, both for the purposes of assessment and to enhance student learning (Wolf, 1991). Hence, the notion of a test with multiple items is only one of many possible assessment episodes in the classroom. Teachers do, however, collect many sources of information about student learning--not only through tests but through a range of formal and informal assessments: homework, classroom work, projects, quizzes. If this information is relevant to their learning targets, teachers could make reasonable generalizations about student learning.

The second dimension of reliability relates to the judgments made about students' work. Rater reliability refers the degree to which raters agree when assessing a given student's work. Studies have documented that when raters are well trained and scoring criteria are well developed, raters can score student work with a high degree of consistency across raters (e.g.,

Hieronymus & Hoover, 1987; Shavelson & Baxter, 1992). In the classroom, however, a single judge (the teacher or a teaching assistant) is often responsible for evaluating all student work. Teachers rarely exchange student work or have another evaluator look at student work.

Reliability in the Classroom Context

For reliability to have meaning for teachers, the concept has to make sense for the classroom and school context. Two dimensions of reliability relevant to the classroom are:

Reliability Dimension 1: Determining the dependability of assessments made about students. The concept of reliability can be reframed to fit the classroom context if the reality of the classroom and a broader and inclusive meaning of reliability are acknowledged. The American Heritage Dictionary (Houghton Mifflin Company, 1981) definition of reliable is "dependable." While measurement professionals have equated dependable with consistent, the former term is more appropriate for the classroom. Assessment may occur frequently in the classroom using measures that could not stand up to psychometric standards of reliability (e.g., research reports, written essays); however, it is possible that grading decisions made at the end of a marking period can be much more reliable than the individual assessments themselves. Even writers who are fairly cautious about performance-based assessments and portfolios admit that the classroom context could provide more reliable assessment information simply because teachers have more information from which to make judgments (Dunbar, Koretz, & Hoover, 1991). Hence, for assessments to be reliable, teachers must ensure that they have sufficient information from which to make dependable decisions about students. Given this framework, evidence for the validity of assessments used to make decisions should be the foremost consideration for teachers. Reliability of assessment decisions depends on the quality of the assessments. If attention is given to evidence for validity, then teachers can begin to ask themselves whether there is sufficient information from which to make dependable decisions. A wide range of assessments can serve the purpose of a long test--the more sources of assessment information, with demonstrable evidence for validity, the more likely dependable decisions can be made.

Teachers can be asked to look across diverse sources of assessment information planned for a given unit of instruction and determine whether there is sufficient information from which to make dependable judgments about students' learning related to the learning targets for the unit. Teachers can use grading policies to organize their thinking about the sources of information available for making judgments about student learning. Rather than using "averaging" techniques in grading, teachers can be shown how to use their professional judgments to look at the range of evidence about student learning and make a "holistic, integrative interpretation of collected performances." (Moss, 1994, p. 7) Reliability, then, becomes a judgment based on sufficiency of information rather than test-retest consistency.

Teachers can also be taught to develop public performance criteria that all students must apply to their work, even if they make their own choices about what work to do (see Figure 1 for an example). This level of standardization can allow for individual choice in projects and other types of performances while still ensuring that students' work will demonstrate their learning related to the targets of instruction. This will also help with rater consistency, the second dimension of reliability.

Figure 1

Directions and Criteria for Literature Project

This project will give you a chance to do some literary analysis. You will be working as a literary critic. In doing so, you will show your understanding of:

- how authors communicate major themes through their writing
- how authors communicate authors' their perspective or purpose in their writing
- how authors use language to create images, mood, and feelings
- how to judge the effectiveness of an author's work

You may choose a short story or a collection of three or more poems by a single author. In your writing be sure to include:

- a main message or theme you see in the story or poems
- what you believe is the author's purpose or perspective
- a description of at least two figurative language strategies the author used to communicate mood, images, and/or feelings
- specific examples from the story or poems to support your claims about theme, purpose, perspective, and figurative language
- an overall judgment about whether the author was effective in communicating themes and his/her perspective/purpose and in using figurative language strategies
- at least three reasons to support your overall judgment

If you choose to choose to use poems, make certain that the poems share a similar theme or message. Remember that authors often have more than one theme or message in their work, but be sure to focus your thinking on only one. Begin your paper by introducing the story or poems and the author. Organize your writing so that it will build a case for your positions and ideas about the writing. Look back at the literary reviews we have studied in class to give you ideas about how to organize your writing.

You must tell me what story or poems you have chosen to write about on _____. You will turn in an outline or web for your paper on _____. The first draft is due on _____. Your final draft, the outline/web, and marked first draft are due on _____. Be sure to give the source of the literary work(s) at the beginning of the paper.

Reliability Dimension 2: Determining the degree of consistency in making decisions across students and across similar types of work. Teachers generally use three types of assessment that could be affected by the consistency of their judgments about students' learning. They create short answer and essay items for tests; they assign projects and performances; they give several similar assignments (such as writing prompts) for which they have the same expectations. In these three situations, consistency of teachers' judgments depends on (a) whether the rules for scoring short answer items and essays are consistently applied across students, (b) whether the rules for scoring extended performances are applied consistently across students, and (c) whether rules for scoring frequently occurring types of assessment are applied consistently across similar tasks.

Teachers can be taught to develop public scoring criteria that they then apply to all students' performances. This can assist them in making consistent judgments across different students' performances. Teachers can be taught how to create generic scoring rules that can apply to multiple similar short answer or essay items (see Figure 2) so that they assess a range of

responses to short answer or essay items based on the same criteria.

Figure 2

Generic Scoring Rules for Historical Essay

Performance Criteria

- Essay is clearly and logically organized.
- Position is clearly stated near the beginning of the essay.
- At least three arguments are given for the position.
- Arguments clearly support position.
- Specific supporting evidence is given for each argument.
- All supporting evidence is accurate and supports arguments.

Scoring Rubric

- **4 points** The essay is clear and logical in taking a position on a historical issue and in supporting the position with arguments and evidence. The essay thoroughly and effectively presents the position, arguments, and supporting evidence such that the reader can understand and entertain the writer's point of view. All supporting evidence is accurate.
- **3 points** The essay is clear and logical in taking a position on a historical issue and in supporting the position with arguments and evidence, although more evidence is needed for at least one argument. The essay presents the position, arguments, and evidence such that the reader can understand the writer's point of view. All supporting evidence is accurate.
- **2 points** The essay takes a position on a historical issue and supports the position with arguments and evidence, although more and/or stronger arguments and evidence are needed. The essay could be organized more effectively to communicate the position , arguments, and evidence. Some information presented may be inaccurate
- **1 point** The essay takes a position on a historical issue but provides little or no support for the position. Organization may or may not communicate the writer's ideas clearly. Some information presented may be inaccurate

If teachers learn how to frame the items and tasks given to students in a way that allows them to make consistent assessments and if they use scoring rules consistently across students and similar tasks, they are more likely to ensure that their evaluations of student's responses are consistent.

We have claimed in this paper that the frameworks we have set forth can be used to design assessment courses for teachers that not only better prepare them for the assessment tasks they will fact, but that help teachers develop habits of mind in which valid and reliable assessment is seen as central to the teaching-learning process. To support this claim, we briefly describe a course based on the validity and reliability frameworks and present evidence of its effectiveness.

Assessment Frameworks in Action

The assessment course described here was taught at a large northwestern university, that provided a certification program for approximately 250 elementary and secondary teachers per year. Courses were ten weeks in length and a given class included pre-service teachers from all academic subjects and the arts for kindergarten through twelfth grade. During the quarter in which the assessment course was taught, students spent at least 20 hours per week in their field placement sites in addition to their course work as a transition into full time student teaching the following quarter.

During the summer of 1991, the decision was made to redesign the tests and measurement course for the teacher preparation program. Prior to that time, didactic procedures were used to cover standardized test interpretation, item writing and item analysis techniques, and statistical procedures for obtaining estimates of validity and reliability of tests. Students were assessed on their ability to write test items in various formats, and tested on their knowledge of measurement principles and concepts.

The redesign of the course was part of an overall restructuring of the teacher preparation program and was based on exit surveys indicating that students did not value the course (R. Olstadt, personal communication, May, 1991) as well as recommendations from the literature about what assessment courses for teachers should address (Airasian, 1991; Linn, 1990; Stiggins, 1991). In redesigning the course, the two most significant shifts were that (a) all assessment concepts were to be taught in the context of instructional practices and (b) the major emphasis of the course was to be on assessment validity and reliability rather than simply assessment techniques and memorization of abstract concepts.

We began with a model proposed by Linn (1990), and expanded it to include the use of process portfolios (Valencia, 1990; Wolf, 1991). We chose process portfolios because they are an interactive teaching tool in which successive iterations of work build upon one another to create a "prepared accomplishment" (Wolf, 1991), in this case a well developed plan that integrates instructional planning and assessment development using clearly defined learning objectives as the unifying force. We then planned assignments that would give students the opportunity to develop specific assessment literacy skills and strategies and that would require students to examine their own work in terms of validity.

In what follows we briefly discuss the work of the course and how the requirements of the assignments designed to help teachers develop the classroom-based definitions of validity and reliability given above. A more thorough description of the course is presented in Taylor and Nolen (1996) and Taylor (in press). In Taylor and Nolen (1996), each classroom course assignment is discussed in terms of its function in helping students think about one or more of the dimensions of validity, including excerpts from the students' self-evaluations that highlight the depth of their learning. In Taylor (in press), the types of decisions that had to be made to effectively use portfolios as an instructional and assessment tool are presented.

The Process Portfolio

The portfolio provided both a means for instruction and learning during the course (process portfolio), and the product used to assess students' learning at the end of the course (showcase or assessment portfolio). The use of process portfolios allowed students to benefit from peer and teacher feedback (formative assessment) on the first draft of each assignment prior to its submission for grading purposes. Instructor feedback was intended to focus their thinking so that subsequent versions of their work reflected a better understanding of the course objectives. With better understanding, students could improve the quality of their own work.

At the end of the course, students pulled all of their work together in an assessment

portfolio that "showcased" their learning for the course. They then wrote self- evaluations of their learning. In what follows the components of the of the portfolio are described.

The Structure of the Assignments for the Course

To teach all five dimensions of validity and both dimensions of reliability, it was necessary to help students investigate assessment concepts in a meaningful context. The centerpiece of the course was a set of related assignments designed to guide students through the development of a unit of instruction so that they could engage in the thinking and skills necessary to make valid connections between learning objectives and instruction, instruction and assessment, and learning objectives and assessment.

For their assignments, students described a plan for a subject they would be likely to teach, and produced documents that were reasonable representations of the types of work good teachers do. Table 1 shows the assignments for the course and the dimensions of reliability and validity each was intended to help students learn.

Table 1
Configuration of the Portfolio Components for the Assessment Course

Title	Description	Validity Dimension	Reliability Dimension
Subject Area Description	A description of the content foci and the instructional units in a subject area for an 8 to 12 week period including content coverage and major concepts targeted.	1	
Subject Area Goals and Objectives	4-6 discipline based 4-6 objectives for each goal with discipline- based rationale for a subject the student planned to teach	1	
Instructional Unit Description	A description of instructional activities that would target 4-6 of the subject area objectives for 2-4 weeks of the period; with activities rationale indicating how each activity would help students learn the relevant objective(s)	1, 3	1
Item Sets: • Checklist or Rating Scale • Performance Assessment • Essay Items • Traditional Items	Four separate item sets as examples of the various types of assessment items and tasks that are used in classroom assessment (observational checklist or rating scale, performance assessment, essay items, traditional items (multiple choice, true-false, completion, matching, short-answer); each with the validity rationale	All	All

Sample Feedback	Mocked-up student work for one unit assessment with written or dialogue of oral feedback; philosophy and rationale about giving feedback	5	
Grading Policy	A description of the types of work that would be included in the grade, how different work would be evaluated, and how absences and late work would be handled; also included an example grade summary for one student	1	1,2
Self Evaluation	Description of own learning of selected course objectives, including discussion of concepts of validity, reliability, bias, and fairness referring to own work to show exam- ples of own learning	All	All

Students were required to write rationales for all assessment decisions made during the development of components of the plan. Writing rationales forced students to articulate the validity and reliability issues that arose within each component of the plan, as well as giving the instructors a means to assess the conceptual learning that complemented the technical work displayed. The process of writing rationales also seemed to lead to deeper understanding of the concepts (Taylor and Nolen, 1996).

When all components were completed, students collected them into a final showcase portfolio. They wrote a single page reflection on each document and a self-evaluation of their learning in the course. In addition to these core assignments, other assignments were given to broaden students' understanding of assessment concepts. They included:

1. "Thought papers" in which they discussed their thoughts about collections of course readings (from the text book and a course reader).
2. A letter to parents explaining norm-referenced testing and score types
3. A written interpretation of one student's scores from a norm-referenced test.

The assignments listed above formed the core of the course as it evolved over the next twelve quarters. Based on student work and feedback, we adjusted the portfolio components, norm-referenced test interpretation assignments, and the number of thought papers required. We clarified instructions and experimented with various scoring schemes for the final portfolios. The focus of this paper is on the classroom assessment components of the portfolio; therefore, the latter three assignments are not discussed further here.

In what follows, we briefly discuss each of the components of the portfolio in the order the components were assigned. We also discuss the links between components and their links to the validity and reliability frameworks.

Subject area description, goals and objectives. Students began by writing a brief (one page) description of a subject area they planned to teach the quarter following the assessment course. The description included a general outline for one quarter or trimester, including the units of study and the major concepts and processes to be taught. The purpose of this component of the plan was to help students envision a subject area as a whole rather than as piece-meal units or text-book chapters. From this vision of the subject area, they were more able to articulate the overall learning goals of the course.

Once the general description was completed, students wrote four to six learning goals and four to six relevant objectives for those goals. We hoped that this level of objective writing would lead our students to clarify, for themselves, the most central learnings in the disciplines they planned to teach. This conceptual clarity is necessary if teachers are to develop assessments that reflect the disciplines studied (Validity Dimension 1).

Finally, students wrote a rationale describing how their goals and objectives reflected the substantive and syntactic structures of the discipline they intended to teach. This requirement built upon the educational psychology course they had taken the previous quarter in which they explored the concepts of disciplinary structure (Schwab, 1978) and pedagogical content knowledge (Grossman, Wilson, & Shulman, 1989). Students revisited this component throughout the quarter as they developed a deeper understanding of their goals and objectives through the assessment development process.

*Unit description.*Once students had completed their subject area descriptions, they described a brief unit of study that would fit within the quarter or trimester they had described in the subject area description. This component proved vital to students' understanding of how to establish the validity of assessments. Without the instructional unit as an anchor, it would be difficult to address aspects like the validity of methods of assessment for the methods of teaching used (Validity Dimension 3). Students developed units that were unique to their individual interests and that they were likely to use; therefore, the units were also a "hook" that kept students engaged in the work of the course..

Students selected up to six subject area objectives as the focus for the instructional unit. Then they wrote a brief narrative of the activities they would use to teach the objectives each day of the unit, linking the objectives to each activity, and providing a rationale for why the given activity or activities would lead to the targeted learning. This helped them to judge the fit of the assessments to the discipline as well as the fit of assessments to the unit of instruction (Validity Dimensions 1 and 3)

Unit Assessments. For the next part of the portfolio, students used a variety of techniques to create assessments for their instructional units. Students fully developed four different types of assessment for their units:

1. *An observational checklist or rating scale.* The assignment for the observational checklist or rating scale required students to identify one or more unit objectives and one or more situations from the unit for which observation would be an appropriate form of assessment. The checklist or rating scale was to have at least 10 items that were of clearly observable behaviors that could show the learning described in the objective(s).
2. *A performance-based assessment.* This assignment included a description of a performance that was either an integral part of the instructional unit or that could be used for students to show the learning objectives that were the target of the instructional unit. Students wrote directions (oral or written) that were sufficient for their students to complete the performance and show the learning, as well as a checklist, rating scale, or rubric(s) to evaluate the performance.
3. *Two essay items.* The assignment for the essay items required students to think about two essay prompts that could be written in the instructional unit through which students could show learning related to one or more of the unit objectives. Essay prompts had to be explicit enough that students would know what they were to do to successfully write the essays. Essays were to be brief (extended essays were considered performance assessments). Students also had to write scoring rules (checklists, rating scales, and/or rubrics) for each essay.
4. *A set of "traditional" test items.* This assignment was for a set of at least 10 items that assessed one or more unit objectives. The set had to include at least three multiple-choice

items, one matching item, two completion item, two true-false items, and two short answer items. The item set could be organized as a quiz, part of a unit test, or into one or more daily worksheets (for younger students). Students had to develop a scoring key for the select items and scoring rules (key words, checklists, rating scales, or rubrics) for the supply items.

Students were asked to develop assessments that fit with their instructional methods and that assessed their unit objectives. Students then had to write a rationale for each item or task that answered several questions:

1. How will the item/task elicit students' learning related to the targeted unit objective(s)? (Validity Dimensions 1 and 2)
2. How does the item/task reflect concepts, skills, processes that are essential to the discipline? (Validity Dimensions 1 and 5)
3. How does the item/task fit with the instructional methods used in the unit? (Validity Dimension 3)
4. How do the rules for scoring the item/task relate to the targeted unit objective(s)? (Validity Dimension 1)
5. Is the mode of assessment such that all students who understand the concepts will be able to demonstrate them through the assessment? (Validity Dimension 4)

By thinking about each item or task and its relationship to the discipline and the unit methods, students went beyond simply practicing item or task writing techniques and had to consider whether the assessment represented the construct (Validity Dimension 1) and whether the assessment was appropriate for the instructional context (Validity Dimension 3). By examining whether items and tasks clearly asked for the learning targeted, students could examine whether assessments were presented in a way that allowed their students to demonstrate learning (Validity Dimension 2; Reliability Dimension 1). By carefully examining the rules for scoring the item/task and how these rules relate to the objective(s) the item/task is intended to measure, students had to think about whether the scores used to represent student performance related to the construct (Validity Dimension 1) and whether their scoring rules would help them be more consistent across students (Reliability Dimension 2). By having to discuss whether all of their students would be able to show their learning through the mode of assessment, our students could begin to explore issues related to bias (Validity Dimension 4). By considering the link between the assessments and the disciplines, students could also begin to grapple with whether assessments were likely to provide appropriate representations of the disciplines for students (Validity Dimension 5). Finally, by creating several assessments in different modes for the same unit and unit objectives, they were able to compare different methods of assessment in terms of their demands for students (Validity Dimension 2).

Feedback. This assignment required students to choose one of their assessments and either try it out with one of their students or mock-up/describe one of their students' responses. They then showed what they would do (either by marking on the paper or by describing a dialogue with their student) to give feedback. Finally, they wrote a rationale for the feedback, including both a discussion about the influence of the feedback on the learner's motivation and self-esteem and a discussion about how the feedback could help their student improve future performance related to the learning target(s). This gave students another opportunity to explore the consequences of assessment interpretation and use (Validity Dimension 5).

Grading Policy. For the grading policy, we had students use the assessment ideas derived from their unit plans and write a grading policy that applied to the entire subject area description. They had to choose an grading philosophy (norm-referenced or criterion-referenced) and explain why they had chosen it. They explained what types of work would contribute to the grades (e.g.,

essays, reports, projects, tests, homework, daily seatwork, etc.) and why this work was important to learning the discipline (Validity Dimension 1), the general strategies they would use to assess various kinds of work (Reliability Dimension 2 [e.g., a generic four point rubric for all homework assignments based on completeness and accuracy of work]), how they would weight the various sources of assessment information, and how they would summarize across assessments to assign a grade. They also had to prepare a sample grade summary for one student using the information from the various assessment sources.

Students had decide how much weight to give to attendance, timeliness, oral participation, and attitude when making judgments about their students' learning of the targeted objectives. By validity standards, some of these variables would be considered sources of irrelevant variance that lead to invalid inferences about student academic learning (Validity Dimension 1). They also had to think about multiple-sources of evidence needed to make reliable decisions about learners (Reliability Dimension 1). Finally, by creating a set of scores for a hypothetical examinee, they were able to look at the impact of various sources of assessment information on overall grades (Validity Dimension 5)

Reflection and Self-evaluation. The final component of the portfolio was the self-evaluation. This component gave students an opportunity to bring closure to the course and to organize their thinking about a few central assessment concepts using the work required in the course as the anchor. In these self-evaluations, they wrote about their understanding of major assessment concepts for the course. They were required to:

1. Discuss their current understanding of the concepts of validity, reliability, bias, and fairness with reference to specific work in the course that helped them understand these concepts and *how* the course work had helped them to understand the concept.
2. Select at least six of assessment course objectives and discuss what they had learned related to each objective, what aspect of the course had helped them to learn it, and how.

The self-evaluations were evaluated for the students' ability to demonstrate their understanding of the assessment concepts using their work as examples. It was not sufficient to provide a text book definition of a term or to explain the impact of assessment in general terms; specific and credible examples were required. In the following discussion, excerpts from student self-evaluations from the Spring 1994 students are used to demonstrate, in their own words, what students thought about as they reflected on their own learning. Selected excerpts represent common thoughts among students.

In the self-evaluations, when students discussed their understandings of validity, most references were made to the unit assessments (Validity Dimensions 1 through 5). Discussions of reliability and fairness usually focused on the use of rubrics and observational checklists and rating scales (Reliability Dimension 1 and Validity Dimension 4). Rarely did students bring up consistency of ratings across students and performances as an element of reliability (Reliability Dimension 2). In discussions of fairness and bias, students often indicated how helpful it had been to use a standardized scoring scheme to evaluate essays or performances in class; how such rules had given them a way to be fair and unbiased in their assessments (Validity Dimension 4). For example:

> "The students in my placement are intentionally given vague criteria. The teacher considers it her right to use her personal judgments of the student's attitude and behavior to influence the grade. If the criteria (are) not spelled out she has the leeway to insert her prejudice. Students realize what is going on and they become cynical and resigned. Few of them try to fight it. This lack of fairness is so widespread that they have come to expect it."

When choosing which component of the portfolio most influenced their learning, each component was selected by someone. For some, the clarification of their disciplines were seen as the most critical element (Validity Dimension 1).

> "The best part of the course for me was the subject area description and goals because it forced me to stop and think about why I want to teach biology. . . . Being a good teacher is a difficult task. The best way to overcome this is going through the process we went through during the development of subject description, goals, objectives, and rationale. . . . It will help me down the road as a teacher."

Some students wrote about the importance of developing a unit of instruction in order to help them conceptualize the role of assessment (Validity Dimension 3).

> "It made me focus on what I really wanted my students to learn, and then I had to find different and appropriate ways to assess whether or not the students learned these things. If one of my unit objectives was to view the American Revolution and its effects from a variety of perspectives, then an assessment that only deals with one perspective is not a valid assessment. It does not tell me if they have learned what . . . I want them to learn."

Many students chose to focus on one or more of the unit assessments, discussing what they had discovered as they developed a given type. A very common observation was about the need for clear directions for performances so that their students would actually provide performances that showed the targeted learning (Validity Dimension 2).

> "Giving the criteria for successful work helps make an assessment valid, as it insures that a student's essay demonstrates the student's conceptual and/or procedural understanding rather than his/her ability to read the teacher's mind."

Another common focus was on the fit between various forms of assessment and either the discipline or the learning objectives as well as what assessments communicate to students about a discipline (Validity Dimensions 1 and 5).

> "Assessments are not neutral!. . . Assessments send messages about a discipline; they communicate to students in a direct, concrete, and powerful way about what is really important to know is this subject."

Students also wrote about grading policies. They typically reflected back on readings about the influences of grading practices on motivation and self-esteem (Covington & Beery, 1976; Canady & Hotchkiss, 1989), discussing the assumptions often made about the motivating power of grades and considering the potential consequences of various ethical and unethical grading practices (Validity Dimension 5). Some students indicated that in being forced to think about the relative weight of each aspect of the grade, they had to look again at the discipline to decide which sources of evidence were best and most important in making judgments about their students' learning (Validity Dimension 1).

These and other comments showed us, as instructors, the power of the work assigned in the course in terms of helping our students understand important assessment concepts. Comments from students suggested that the assignments done for the course as well as the rationales and self-evaluation enhanced their learning.

Comparative Studies of the Traditional Tests and Measurement Assessment

Course and the Portfolio-Based Course

In an effort to evaluate the effectiveness of the revised course, three studies were conducted that compared data available from students who had taken one of the two versions of the course: the portfolio-based course and the traditional tests and measurement course. The classroom assessment component of the original assessment course covered item writing and item analysis techniques (some later sections of this course also covered performance assessment), and statistical procedures for obtaining estimates of validity and reliability of tests. Instructors used a combination of lectures and discussions to teach assessment content. Instructors relied heavily on midterm and final examinations (primarily multiple-choice), which counted for 60 to 70 percent of the final grade (depending on the instructor). Up to 25 percent of the final grade was based on students' development of behavioral objectives (based on Bloom's taxonomy) and tests or sets of items to measure those objectives. Tests or sets of items were independent of any context except that of the behavioral objectives.

Study 1 compared course evaluations across teaching faculty for the two versions of the course. Study 2 compared evaluations of relevant components of an exit survey given to all students graduating from the teacher education program. Study 3 involved analyses of data from follow-up surveys sent to teacher education students in the quarter following their enrollment in the assessment course--the time during which most were engaged in full-time student teaching. In the survey, the pre-service teachers were asked to discuss assessment issues, validity dilemmas, and reliability dilemmas that arose in their teaching. Each of these studies is described more fully below.

The designs of the three studies reflect the natural development of curricular revision, rather than the carefully-controlled world of laboratory studies or field experiments. The research opportunity was presented by the decision to redesign the course. Thus, comparisons of the two versions of the course presented in Studies 1 and 2 depended on existing institutional data. The data for Study 3 were collected as part of an evaluation of the course revision, but the decision of one instructor to revert to the traditional format for two sections provided an opportunity for comparison on the follow-up measure.

Study 1: Course Evaluations

Data Source. The university's Office of Educational Assessment provided course evaluation results for each quarter from the summer quarter of 1988 through the spring quarter of 1994. Course evaluations are required for every course for assistant professors and at least once a year for senior faculty. Student participation is voluntary, however, most students complete the form. Results of the course evaluation are not given to the instructor until after grades are submitted.

Data representing 12 quarters of the traditional tests and measurement version of the course and 12 quarters of the revised course were available. The number of respondents from the traditional tests and measurement course ranged from 15 to 55 across different sections with a mean of 32.25. The number of respondents from the revised course ranged from 17 to 74 with a mean of 32.58. Because responses were anonymous, it was not possible to determine the exact number of males and females in the sections nor the number of students who were to be certified in elementary, secondary, or music education. Academic ranks for the instructors in the traditional tests and measurement course ranged from graduate student instructor to full professor. Academic ranks for the instructors in the revised course ranged from graduate student instructor to assistant professor. There were 8 different instructors for the traditional tests and measurement course and 3 different instructors for the revised course.

Only those items common to evaluation forms used in all sections of the course were

included in the analysis. Items common to all forms are given in Appendix A. Each item was rated on a 6 level scale. "Excellent" (5), "very good" (4), "good" (3), "fair" (2), "poor" (1), and "very poor" (0). Four items from this common set assessed students' ratings of the content and relevance of the course.

Results. Mean item scores were averaged across classes for each version of the course. Only those items specifically related to the content of the course and the relevance of the course were included in the analyses. Two analyses were performed on a selected set of the items. In the first analysis, data from four items from the course evaluation forms were used: (a) course as a whole, (b) course content, (c) amount you learned in the course, and (d) relevance and usefulness of course content.. These items were summed to obtain an overall score for the general content of the course; the mean score for the traditional tests and measurement course was 12.09 (SD = 2.04), and for the revised course was 16.48 (SD = 1.62). In the second analysis, relevance and usefulness was analyzed alone, with means for the traditional tests and measurement and revised course 2.92 (SD = .57) and 4.29 (SD = .38), respectively.

T-tests were performed to compare mean ratings for these data. There were significant differences between students perceptions of the general content of the course ($t(22) = 5.85$, $p < .001$) and the relevance and usefulness of the course ($t(22) = 7.00$, $p < .001$). Students in the revised course saw the course as more relevant to their needs and rated the content of the course between "very good" and "excellent." Students in the traditional tests and measurement course rated the course as "good" on both general content and relevance and usefulness.

These differences might have been due to differences in the effectiveness of individual instructors. However, even instructors of the traditional tests and measurement course who received high ratings for instructor's effectiveness received lower ratings on relevance and usefulness of course content. and course content. Two instructors from the traditional tests and measurement course had high ratings for instructor's effectiveness (mean ratings of 4.38 and 4.25), comparable to the average ratings for the two revised course instructors with the highest effectiveness ratings (mean ratings of 4.20 and 4.54). When only these four instructors are compared, the mean ratings for the for relevance and usefulness were 3.52 and 3.83 for the traditional tests and measurement course and 4.81 and 4.71 for the revised course. The mean ratings for course content were 3.90 and 3.64 for the traditional tests and measurement course and 4.71 and 4.54 for the revised course. This suggests that whether students saw the content of the assessment course as relevant to their needs was not merely a function of their perceptions of the effectiveness of an instructor.

Study 2: Teacher Education Program Exit Surveys

Subjects. As part of the ongoing evaluation process of the teacher education program, exit surveys were administered in the last quarter of the program to all students. We obtained 153 of these surveys from three years just prior to the change in the assessment course (1989-91) and 145 from two years after the change (1992 and 1994). In the summer of 1992 an outside instructor taught a traditional tests and measurement course. Since it was not possible to tell which 1993 exit surveys came from students who had taken the revised course, data from that year were not used. All responses were anonymous; therefore, the demographic characteristics of the respondents were unavailable.

Instruments. Exit surveys were general program review instruments and asked a variety of questions about students' experiences in the teacher education program, including both course work and field work. There were several items which provided some information about students' perceptions of assessment course effectiveness. First, a set of items asked students to rate how well the program as a whole had prepared them in a number of areas corresponding to the state requirements for teacher education programs. One of these items was "How well has this

program prepared you to evaluate student work," which students rated on a scale from 1 ("not at all prepared") to 5 ("thoroughly prepared").

A set of open-ended questions asked students to comment on various program aspects. Three of these questions were coded for comments related to the assessment course.

The first of the open-ended questions asked for comments about any of the courses in the program. Coding schemes for this item were as follows:

1. Comments specifically directed at the assessment course, and related to value or worth of the course or its content were coded (0) if they suggested eliminating the course altogether; (1) if they stated the course was worthless, not valuable, not useful for teachers; and (2) if they stated the course was valuable, applicable or useful.
2. General comments (not referring to value) were coded (1) negative or (2) positive.

A second item asked students to list aspects of the teacher education program that were particularly valuable or worthwhile. Raters counted the number of students listing the assessment course here.

A third item asked what important material was left out or not sufficiently covered. Raters counted any mention of an assessment-related topic (e.g., setting up grade books, portfolios, informal observation). Finally, negative comments regarding the work load related to the assessment course mentioned anywhere in the survey were counted.

All coding was completed by the authors and one graduate student who was unfamiliar with the purpose of the research. There was a 98% agreement among the three raters. Final counts for each code assigned to each response were based on absolute agreement among the raters.

Results. Ratings of how well students thought the program prepared them to do assessment were compared across courses using a one-way ANOVA. Students who took the revised course rated the teacher education program as preparing them more thoroughly to do assessment (Mean = 4.07, SD = 0.87) than did students who took the traditional tests and measurement course (Mean = 3.22, SD = 1.04; $F(1, 296)= 58.36$, $p < .001$).

Frequency of responses for each open-ended item appear in Table 2. In general, the comments were more positive for the revised course, though not uniformly so. Typical comments for the traditional tests and measurement course included "[The assessment course] was a useless class. Testing and evaluation are essential, but I learned almost nothing in this class" and "Did not relate to the real world." Typical comments for the revised course included "[The assessment course] provided me with the information that I considered most valuable in my field experience" and "[The assessment course] was the most valuable class overall for my teaching." Eight students in the revised course (5.2%) stated that the work load in the revised course was excessive, while none of the students taking the traditional tests and measurement course did so.

Table 2
Frequency of responses to each item for the traditional tests and measurement course and the revised course

Comments (Value)				
Course	**N of Cases**	Valuable	Not Valuable	Eliminate Course
Revised Course	145	19	2	0

Traditional Course	153	1	17	9
Comments (General)				
Course	**N of Cases**	Positive	Negative	Negative Work load
Revised Course	145	22	4	8
Traditional Course	153	0	9	0
What aspects of the program were...				
Course	**N of Cases**	Particularly Valuable		Not Sufficiently Covered
Revised Course	154	28		3
Traditional Course	129	1		11

Each comment was coded into only one category, but some students mentioned the assessment course in more than one way. Therefore a new variable was created by counting the number of students in each group who had responded in some way that the assessment course was valuable and the number of students who had indicated that the course was not valuable. Students who had taken the revised course were much more likely to mention it as a valuable part of the program (31%) than to say it was not (2%), while those taking the traditional tests and measurement course were more likely to see the course as not valuable (17%) than as valuable (1%) (chi-square(1) = 61.8, $p < .001$).

Study 3: Follow-up Survey

Study 3 aimed to assess the impact of the assessment courses on pre-service teachers' work in their field placement classrooms. We were primarily concerned with their ability to describe assessment issues they faced in teaching, and in their understanding of validity and reliability. We were also interested in the extent to which they could use the assessments (and other components of their work for the course) in their field placement classrooms.

Subjects. Students from six different quarters were asked to be part of an anonymous mail survey during quarter following the one in which they took the assessment course. Most of the students were engaged in full-time student teaching. Two classes of students (N = 112) who had taken the traditional tests and measurement course during the summer of 1992 were surveyed. Twenty-one percent (n = 23) of these students completed and returned the surveys. Five classes of students (N = 195) who had taken the revised version of the course between the summer of 1991 and the autumn of 1992 were surveyed. Twenty-five percent (n = 50) of those enrolled completed and returned the surveys.

Results. The follow-up questionnaire addressed a number of assessment and programmatic issues. A complete list of items is shown in Appendix B. There were few differences between groups on the assessment methods used in their field placement, the proportion of planning time spent on assessment, or the amount of time they reported thinking about assessment. Students in the revised course reported spending slightly more time planning assessments (7% of planning time, SD = 4%) than traditional tests and measurement course students (3%, SD = 4%), $t(65) = 9.54$, $p < .01$).

Ninety-two percent of the students in the revised course reported using all or part of the work developed in the course, while only 8% of students in the traditional tests and measurement course reported using any of the work developed in their course (chi-square(1)=9.03, $p < .01$). Students who reported using materials developed in the course rated the process of planning helpful on a 5-point scale from 1 ("not at all helpful" to 5 "very helpful"), with a mean of 4.17 (SD = .81).

Three items provided information on students' post-course understanding of assessment issues, validity, and reliability. Responses to items 4, 6, and 7 (the influence of assessment, validity issues, and reliability issues) were independently coded by three full professors with strong measurement and statistics backgrounds who had previously taught classroom assessment courses. They were not aware of the purposes of the study or the type of course in which students were enrolled. Coding was based on the degree to which the students' responses showed understanding of general assessment concepts. Table 3 provides the scheme used to code student responses.

Table 3

Coding scheme for relevant items of the post-course survey

4 Influence of course on teaching
Code 1: 1 = yes 2 = no
Code 2

- 2 = shows clear, unambiguous understanding of appropriate uses of assessment
- 1 = - shows partial understanding of appropriate uses of assessment
  - describes delivery of instruction; may have assessment links
  - uses assessment terms without examples
- 0 = shows little or no understanding of appropriate uses of assessment in instruction
- NS = not scorable (off task or omitted)

6 Validity issues

- 2 = gives good example of validity issue
- 1 = - possible example of validity issue, somewhat unclear
  - may confuse validity with reliability
- 0 = gives example that is neither reliability nor validity
- NS = not scorable (off task or omitted)

7 Reliability issues

- 2 = gives good example of reliability issue
- 1 = - possible example of reliability issue, somewhat unclear
  - may confuse validity with reliability
- 0 = gives example that is neither reliability nor validity
- NS = not scorable (off task or omitted)

The final code assigned to each item for each examinee was based on a majority agreement among the raters. For students from the traditional tests and measurement group, 35% indicated that the course had no effect on their teaching. For the students in the revised course, 2% indicated that the course had no effect on their teaching.

For influence of assessment course, 70 percent of students from the revised course showed a clear understanding of the appropriate uses of assessment, as judged by the raters, as compared to 44 percent of the students in the traditional tests and measurement course (chi-square(3) = 9.96, $p < .02$). For validity issues, 70 percent of students from the revised course gave good examples of validity issues as compared to 22 percent of the students from the traditional tests and measurement course (chi-square(3) = 15.01, $p < .001$). For reliability issues, 22 percent of students from the revised course gave good examples of reliability issues as compared to 13 percent of the students from the traditional tests and measurement course (chi-square(3) = 8.74, $p < .03$), however, a fairly large proportion of both groups gave no examples at all (61% of the traditional tests and measurement course students and 42% of the revised course students). In addition, a fairly large percent of the students in the revised course (32%) received a score of 1 for this item, indicating that while the students in the revised course may have been better prepared to address issues related to reliability than were the students in the traditional tests and measurement version of the course, they were still not sufficiently prepared.

Discussion

The results of these three studies suggest that the revision of the assessment course was beneficial to preservice teachers. Students taking the revised course were more likely to see the course as useful and relevant to their own work as teachers than students in the traditional tests and measurement course, both at the end of the quarter in which they took the course, and at the end of their teacher education program (following full-time student teaching). Students taking the revised course felt better prepared to deal with classroom assessment than similar students in the traditional tests and measurement course by nearly a full standard deviation. Nearly a third of those responding listed the revised course as one of the most useful parts of the teacher education program; only 1% of students listed the traditional tests and measurement course.

Although student ratings are valuable, they do not bear on whether students actually learned central concepts in assessment and could use those concepts in their own classrooms. The results of the follow-up survey (Study 3) suggest that students in the revised course were indeed able to use the notion of validity generatively. The concept of reliability; however, was not as clearly understood by the majority of students in either version of the course.

Post-course questionnaires showed that, while students in the revised version of the course had a better understanding of reliability (as it applied to the assessments used in their field placements) than did students in the traditional tests and measurement course, their understanding of reliability was still inadequate. This could be due to the intense focus on a broad understanding of validity and inadequate attention to reliability issues. Many of the examples of reliability issues given by students who had taken the revised version of the course were actually validity issues. These comments, while inaccurate representations of the concept of reliability, did show an understanding of the difference between appropriate and inappropriate assessment practices. On the other hand, survey comments from students who had taken the traditional tests and measurement class indicated that they were very confused about meanings of reliability and validity. Several of these students responded to the questionnaire items about reliability and validity with:

" I don't understand the concept. I only memorized it for class."

It appears from these data that the revised assessment course was effective in helping students understand appropriate assessment practices in the context of the classroom and in helping them develop a generalizable understanding of the concept of validity. What was lacking was a deep understanding of reliability and how it transferred to the world of teaching. Subsequent to these analyses, the course was revised in order to help students focus more carefully on the dimensions of reliability. Follow-up studies are planned to determine whether these adjustments accomplished the course goals.

Conclusion

The assessment course outlined here has been designed to engage students in tasks relevant to their own work as preservice teachers and demand that they consider assessment in the context of disciplinary structures and instructional practices. Each component of the portfolio gave students an opportunity to address one or more of the dimensions of validity and reliability highlighted in this paper. The focus on validity guided student learning from the initial subject area description and concomitant goals and objectives (which helped students develop clearer definitions of their disciplines for themselves), to the unit assessments (which helped students explore all five dimensions of validity), to the grading policy (which helped them address issues of multiple sources of evidence, appropriateness of evidence, and potential consequences of assessment interpretations and use).

One powerful aspect of this course may have been that it was a model of the concepts students were learning. In contrast to a course in which teachers act as impartial observers of students' learning, the instructors were engaged as participant observers--using feedback and guidance to help ensure learning for as many students as possible. Multiple sources of information were used to determine whether students were learning the concepts and skills of the course, from the components of the portfolio to the reflections and self-evaluations at the end of the course.. Students had more than one opportunity to return to their work and revise based on feedback from the instructor and later learning. As such, the instructors had multiple opportunities to observe students' growth over time. Public criteria were used to communicate the expectations of performances and scoring rules were consistently applied across students' work and across similar performances.

Another powerful aspect of the course was that it was carefully focused on tasks and reflective writing designed to help students grapple with each of the dimensions of reliability and validity described in this paper. The learning that resulted from the course--in terms of students' transfer of ideas from the course to their own teaching as judged by three full professors with substantial knowledge of assessment concepts--suggests that the validity framework used to organize the work of the students is one that teachers can internalize and understand. A stronger focus on the sufficiency of assessment information and ways of ensuring scoring consistency in students' work was needed if students were to better understand the concept of reliability.

The success of this course in reaching teachers has implications not only for the preparation of teachers, but for the ways in which we present measurement theory in textbooks and instruction and for how classroom assessments are used in large scale assessment programs. While there may be a place for external assessments that provide accountability data to taxpayers, legislators, and state boards of education, the measurement model developed for these external tests does not fit the rich and complex environment in which learning takes place.

If we are to adequately prepare teachers in the area of assessment, clearer thinking is needed about the assessment concepts, types of textbooks and the methods of teaching that are used. Measurement professionals often lament the wide-spread lack of understanding about measurement concepts. Quite possibly we have created this problem ourselves. The problems seen may be due to the fact that the philosophical foundations of test theory don't fit the

classroom context well. Although text book authors may be trying, in their individual ways, to construct texts book that will force a fit where one does not exist, we may need to admit that a test theory that fits the modernist notions of the impartial observer is not appropriate for the context in which the teaching and learning occur.

It is likely that two different frames are needed for educational assessment constructs: one for the context of school and one for the context of external norm-referenced tests. Textbooks could acknowledge the differences between these contexts and frame concepts, procedures, and skills as appropriate for each context. Courses could be designed to help teachers internalize and grapple with these differences. Textbooks and teacher educators could regularly bring teachers back to classroom-relevant dimensions of validity and reliability within chapters that address various assessment problems, skills, decision-making issues and processes for the classroom. They could ask students to think deeply about why very different frameworks and methodologies apply to external assessments. As measurement professionals and teacher educators, we could do a better job of preparing good "participant observers," as well as helping teachers understand the paradigm shifts between the two perspectives on assessment. Most importantly, we should frame our preparation of teachers in such a way that they are clear about their own tasks as teachers: to promote students' ongoing learning.

At this point in time, while we have standards for educational and psychological testing (AERA, APA, NCME, 1985), standards for assessment competencies for teachers (AFT, NCME, NEA, 1990), and standards for various professional groups in the interpretation and use of tests (e.g., American Association for Counseling and Development, 1989; APA Committee on Children, Youth and Families, Committee on Testing and Assessment and Committee on Ethnic Minority Affairs, 1992; American Speech-Language-Hearing Association, 1991), we do not have standards for the preparation of teachers related to assessment or for the materials used in that preparation. In addition, as AERA, APA, and NCME revise the testing standards, it is critical that they look carefully at the contexts in which assessments apply as well as the philosophies underlying the use of assessments within those contexts rather than attempting to create omnibus standards that apply to all assessment circumstances.

Related to this, as large scale assessment programs look at the viability of incorporating classroom-based assessments into statewide accountability information, the nature of the classroom context, and the proposed validity and reliability frameworks, should be considered. Some might say that, given the unstandardized and progress-oriented nature of classroom assessments, the information derived from these sources is too unstable to use for large scale assessment purposes. On the other hand, the richness and breadth of the assessment information that arises from classrooms could give us more and better information if we more appropriately develop teachers' assessment skills.

As state and national programs attempt to incorporate classroom assessment information when reporting on students' learning, the focus must be on the validity and reliability frameworks that fit the classroom rather than ones that fit external tests. The dimensions of validity and reliability presented here make sense to teachers because they make sense in a classroom context of teaching and learning. Large scale assessment programs that use classroom-based evidence should consider the dimensions of validity and reliability relevant to the classroom when making decisions about how to incorporate classroom-based information into large-scale programs.

If, in order to obtain assessment information from classrooms, large scale programs create top-down standardized tasks or tests to be administered by teachers, the validity of such assessments for the classroom context is suspect. Given the validity framework presented here, top-down classroom assessments could not provide valid classroom assessment information because they would not follow from instruction (Validity Dimension 3). They would simply be extensions of external, standardized tests. If teachers are admonished to use standardized administration directions that do not allow for the unique needs of students, top-down classroom

assessments should be suspect because they may prevent some students from showing their learning in ways that accommodate their unique needs (Validity Dimension 4). If standardized, top-down tasks are closely circumscribed in order to strengthen reliability, they limit the capacity of the assessments to assess students' understandings of the subject area disciplines (Validity Dimension 1). This would not only limit fit with the content and constructs to be measured, but would rob the classroom of the opportunity to use important assessments to accurately represent the structures of disciplines (Validity Dimension 5). Limiting classroom assessment information to a few, standardized, top-down assessments would also limit the range of evidence and counter-evidence that teachers could present about student learning--a threat to Validity Dimension 2.

If, on the other hand, several generic outlines for tasks, scoring rules, and tests are created, (e.g., Rekase, 1995), and teachers are allowed to configure these assessment outlines to fit their own instructional methods, content focuses, and timelines, classroom assessments could fit all of the dimensions of validity relevant to the classroom context. Guidelines for adaptation of the assessments to instructional contexts, strategies for evaluating the validity of these adapted assessments, and ideas about what would constitute a reasonable range of assessment information for decision-making could help teachers develop useful assessments, first for themselves and their students and secondly for large scale programs. State programs could provide powerful professional development materials to practicing teachers through such materials.

For too long, rules for creating and evaluating external tests have been seen as the ideal for obtaining valid and reliable information about learning in the classroom. This has led to a lack of fit between the needs of teachers and the notions of assessment professionals. With the current awareness of the importance of assessment among teachers, school administrators, and policy-makers, the classroom has the potential to be a much more powerful and complete source of assessment information. To achieve this potential, however, we must begin with frameworks for measurement constructs that fit the classroom context, teach teachers how to use these frameworks to improve the quality of their assessments, and ensure that external uses of classroom assessment information attend to these frameworks when deciding how to incorporate classroom assessments into large scale programs.

References

Airasian, P. (1991). Perspectives on measurement instruction for pre-service teachers, Educational Measurement: Issues and Practice, 10 (1) 13-16, 26.

Airasian, P. (1991). Classroom Assessment, First Edition. New York: McGraw-Hill.

Airasian, P. (1993). Classroom Assessment, Second Edition. New York: McGraw-Hill.

American Association for Counseling and Development. (1989, May 11). The responsibilities of test users. Guidepost, 12, 16, 18, 27.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). Standards for teacher competence in educational assessment of students, Washington, DC: Author.

American Psychological Association Committee on Children, Youth and Families, Committee on Testing and Assessment & Committee on Ethnic Minority Affairs. (1992). Psychological testing of language minority and culturally different children, Washington, DC: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing, Washington, DC: American Psychological Association.

American Speech-Language-Hearing Association. (1991). Code of ethics of the American Speech-Language-Hearing Association. Rockville, MD: Author.

Anderson, L., Blumenfeld, P., Pintrich, P. I., Clark, C., Marx, R., & Peterson, P. (1995). Educational psychology for teachers: Reforming our courses, rethinking our roles. Educational Psychologist, 30, 143-157.

Bloom, B. S., Madaus, G.F., & Hastings, J. T. (1981). Evaluation to Improve Learning. New York: McGraw-Hill Book Company.

Bricklin, B., & Bricklin, P. M. (1967). Bright child--poor grades. New York: Dell.

Butler, R., & Nisan, M. (1986). Effects of no feedback, task related comments, and grade on intrinsic instruction and performance. Journal of Educational Psychology, 78, 210- 216.

Canady, R. L., & Hotchkiss, P. R. (1989). It's a good score! Just a bad grade. Phi Delta Kappan, 71 (1), 68-71.

Cohen, D., & Ball, D. L. (1990). Relations between policy and practice: A commentary. Educational Evaluation and Policy Analysis, 12 (3), 331-338.

Covington, M. V., & Beery, R. G. (1976). Self-worth and school learning. New York: Holt, Rinehart, 42-63, 77-87.

Covington, M. V., & Omelich, C. L. (1984). Task-oriented versus competitive learning structures: Motivational and performance consequences. Journal of Educational Psychology, 76, 1199-1213.

Cronbach, L. J. (1970). Essentials of Psychological Testing, Third Edition. New York: Harper & Row, Publishers.

Deci, E., & Ryan, R. (1985). Intrinsic motivation and self- determination in human behavior. New York: Plenum.

Deci, E., & Ryan, R. (1987). The support of autonomy and the control of behavior. Journal of Personality and Social Psychology, 53, 1024-1037.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4 (4), 289-303.

Galton, F. (1889). Natural inheritance. New York: Macmillan.

Glesne, C. & Peshkin, A. (1992). Being there: Developing understanding through participant observation. Becoming Qualitative Researchers: An Introduction. White Plains, NY: Longman, 39-61.

Grossman, P. L. (1991). Overcoming the apprenticeship of observation in teacher education coursework. Teaching and Teacher Education, 7 (4), 345-357.

Grossman, P. L., Wilson, S. M., & Shulman, L. S. (1989). Teachers of substance: Subject matter knowledge for teaching. In M. C. Reynolds (Ed.), Knowledge base for the beginning teacher. New York: Pergamom.

Gullickson, A. R. (1986). Teacher education and teacher- perceived needs in educational measurement and evaluation. Journal of Educational Measurement, 23 (4), 347-354.

Gullickson, A. R. (1993). Matching measurement instruction to classroom-based evaluation: Perceived discrepancies, needs, and challenges. In S. L. Wise & J. C. Conoley (Eds.), Teacher training in measurement and assessment skills. Lincoln, NE: Burros Institute of Mental Measurements, University of Nebraska.

Hanna, G. S. (1993). Better Teaching Through Better Measurement. Fort Worth, TX: Harcourt Brace Javanovich College Publishers.

Hieronymus, A. N., & Hoover, H. D. (1987). Iowa Tests of Basic Skills: Writing Supplement Teacher's Guide. Chicago: Riverside Publishing Company.

Houghton Mifflin Company, American Heritage Dictionary of the English Language, Morris, W. (Ed.). Boston: Author, 1098.

Linn, R. L. (1990). Essentials of student assessment: From accountability to instructional aid. Teachers College Record, 91 (3), 422-436.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13 (1), 5- 8, 15.

Linn, R. L., & Gronlund, N. E. (1995). Measurement and Assessment in Teaching, Seventh Edition. Englewood Cliffs, NJ: Merrill, an imprint of Prentice Hall.

Lortie, D. (1975). Schoolteacher: A sociological study. Chicago: University of Chicago Press.

Mehrens, W. A., & Lehmann, I. J. (1991). Measurement and Evaluation in Education and Psychology, Fourth Edition. Fort Worth, TX: Holt, Rinehart, and Winston, Inc.

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

Moss, P. A. (1994). Can there be validity without reliability? Educational Researcher, 23 (2), 5-12.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. Educational Researcher, 25 (1), 20-28.

National Council of Teachers of Mathematics (1991). Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions. Reston, VA: Author.

National Council of Teachers of Mathematics (1995). Assessment Standards for School Mathematics. Reston, VA: Author.

Nicholls, J. G. (1989). The competitive ethos and democratic education. Cambridge, MA: Harvard University Press.

Nitko, A. J. (1996). Educational Assessment of Students. Englewood Cliffs, NJ: Merrill an imprint of Prentice Hall.

Nolen, S. B., & Nicholls, J. G. (1994). A place to begin again in research on student motivation: Teachers' beliefs. Teaching and teacher education, 10, 57-69.

Oosterhof, A. (1996). Developing and Using Classroom Assessments. Englewood Cliffs, NJ: Merrill, an imprint of Prentice Hall.

Rekase, M. (1995, December). Using portfolios for high stakes assessments. Presented at the Washington State Assessment Conference, Seattle, WA.

Resnick, L. B. & Resnick, D. P. (1991). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement, and instruction. Boston: Kluwer.

Salvia, J., & Ysseldyke, J. E. (1995). Assessment, Sixth Edition. Boston: Houghton Mifflin Company.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. Educational Measurement: Issues and Practice, 10 (1), 3-6, 12.

Schafer, W. D., & Lissitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. Journal of Teacher Education, 38 (3), 57-63.

Schon, D. A. (1987). Educating the reflective practitioner: Toward a new design for teaching and learning in the professions. SF: The Jossey-Bass Higher Education Series.

Schwab, J. J. (1978). Science, curriculum, and liberal education. Chicago: University of Chicago Press.

Shavelson, R. J. & Baxter, G. P. (1992, May). What we've learned about assessing hands-on science. Educational Leadership, 20-25.

Smith, M. L. (1991). Meanings of test preparation. American Educational Research Journal, 28 (3), 521-542.

Sommers, N (1982). Responding to student writing. College Composition and Communication, 33 (2), 148-156.

Stiggins, R. J. (1991). Relevant training for teachers in classroom assessment. Educational Measurement: Issues and Practice, 10 (1), 7-12.

Stiggins, R. J. (1994). Student centered classroom assessment.. New York: Merrill, an imprint of Macmillan College Publishing Company.

Stiggins, R. J., & Bridgeford, N. J. (1988). The ecology of classroom assessment. Journal of Educational Measurement, 22 (4), 271-286.

Stiggins, R. J., & Faires-Conklin, N. (1988). Teacher training in assessment. Portland, OR: Northwest Regional Educational Laboratory.

Stiggins, R. J., & Faires-Conklin, N. (1992). In teachers' hands: Investigating the practices of

classroom assessment.. Albany, NY: SUNY Press.

Stiggins, R. J., Faires-Conklin, N., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. Educational Measurement: Issues and Practice, 5 (2), 5-17.

Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment. Journal of Educational Measurement, 26 (3), 233-246.

Stuck, I. (1995, April). Heresies of the new unified notion of test validity. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Taylor, C. S. (in press). Using portfolios to teach teachers about assessment: How to survive. Educational Assessment.

Taylor, C. S. & Nolen, S. B. (1996). A contextualized approach to teaching teachers about classroom-based assessment. Educational Psychologist, 31 (1), 77-88.

Toom, A. (1993). A Russian teacher in America. Journal of Mathematical Behavior, 12, 117-139.

Valencia, S. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. Reading Teacher, 43 (4), 338-340.

Vidich, A. J., & Lyman, S. M. (1994). Qualitative methods: Their history in sociology and anthropology. In Handbook of Qualitative Research, Denzin, N. K. & Lincoln, Y. S. (Eds.). Thousand Oaks, CA: SAGE Publications, Inc., 23- 59.

Whyte, W. F. (1943). Street corner society: The social structure of an Italian slum. Chicago: University of Chicago Press.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. Journal of Teacher Education, 42 (1), 37-42.

Wolf, D. P. (1991). Assessment as an episode of learning. In R. Bennett and W. Ward (Eds.), Construction versus choice in cognitive measurement. Hillsdale, NJ: Lawrence Erlbaum Associates.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). Measurement and Evaluation in the Schools. New York: Longman.

Appendix A
Course evaluation items common across all evaluation forms

Section	Item	Stem
1: General Evaluation	1	Course as a whole
	2	Course content
	3	Instructor's contribution to the course
	4	Instructor's effectiveness in teaching the subject matter
2: Feedback to Instructor	1	Course organization
	3	Explanations by instructor
	4	Instructor's ability to present alternative explanations

	5	Instructor's use of examples and illustrations
	7	Student confidence in instructor's knowledge
	8	Instructor's enthusiasm
	11	Availability of extra help when needed
3: Information to Other Students	1	Use of class time
	2	Instructor's interest in whether students learned
	3	Amount you learned in the course
	4	Relevance and usefulness of course content
	5	Evaluative and grading techniques (tests, papers, projects)
	6	Reasonableness of assigned work
	7	Clarity of student responsibilities

Appendix B

Post-course survey items:

1. Please check the methods of assessment you are using in your field placement (list of 12 types of assessment, including worksheets, lab write-ups, observational records, paper-pencil tests, written reports, portfolios, peer evaluations)
2. Use the pie chart below to estimate the portion of your planning time you use each week to do the following activities (various planning activities, including planning lessons, assessments, units, writing objectives, etc.)
3. For each of the following situations, how often do you think about assessment issues? (3-point scale: frequently, sometimes, rarely); list of ten situations, including teaching, grading, planning instruction, observing other teachers, riding to and from work.
4. Thinking back on (the course) have any ideas or other aspects of the course influenced your teaching? If so, what part of (the course) has influenced your teaching the most? How has this influenced your teaching?
5. Have you had any new thoughts, questions, or understandings about assessment this quarter? If so, what are they?
6. Have you wrestled with any validity issues in your field placement this quarter? If so please describe one such issue.
7. Have you wrestled with any reliability issues in your field placement this quarter? If so please describe one such issue.
8. Have you taught all or part of the unit you designed for EDPSY 308? (For traditional course students: Have you used any of the materials or assessments you developed?)
9. If so, how helpful was the original plan or planning process? (5-point scale)

About the Authors

Catherine S. Taylor

Assistant Professor of Educational Psychology
312 Miller Hall, Box 353600
University of Washington
Seattle, WA 98195-3600

Voice phone: 206-543-1139
FAX: 206-543-8439
E-mail: ctaylor@u.washington.edu

EDUCATION

Ph.D. University of Kansas, 1986: Educational Psychology and Research
M.S.E. University of Kansas, 1978: Counseling Psychology
B.S.E. University of Kansas, 1974: Language Arts Education

EMPLOYMENT

1991- Assistant Professor, University of Washington, Educational Psychology-Research and Measurement
1986-1991 Senior Editor/Senior Project Manager, CTB/McGraw-Hill
1984-1986 Psychometrician, Psychological Corporation

RESEARCH INTERESTS

My main research focuses are large scale assessment development issues, validity theory, test theory, and research in the preparation of teachers. Current projects include studies of different scoring methods for performance-based tests in mathematics, reading, and writing, and a study of the philosophical foundations for and the social consequences of tests and testing practices.

Susan Bobbitt Nolen

Associate Professor of Educational Psychology
University of Washington
322 Miller Hall, Box 353600
University of Washington
Seattle, WA 98195-3600

Voice phone: 206-543-4011 ('96-'97 only)
206-543-1846
Fax: 206-543-8480 ('96-'97 only)
206-543-8439
sunolen@u.washington.edu

EDUCATION
Ph.D. Purdue University, 1986: Educational Psychology
M.Ed. Lewis & Clark College, 1976: Education of the Hearing-Impaired
B.A. Portland State University, 1975: Speech Pathology & Audiology

EMPLOYMENT

- 1990- Associate Professor, University of Washington Educational Psychology-Human Development & Cognition
- 1986-90 Assistant Professor, Arizona State University West Educational Psychology
- 1978-80 Teacher, Oregon School for the Deaf, Salem, OR High School English and Reading
- 1976-77 Teacher, Lacey Elementary School, Lacey, WA North Thurston Regional Program for the Hearing-Impaired

RESEARCH INTERESTS

My main research focus is the relationship between motivation and learning, and how this relationship develops over time. Current projects include investigations of how motivation develops differently depending on the learner¹s interpretation of their social context for learning. A second interest is in assessment in schools, and the effects of various policies and practices on teacher and student motivation.

EPAA can be accessed either by visiting one of its several archived forms or by subscribing to the LISTSERV known as EPAA at LISTSERV@asu.edu. (To subscribe, send an email letter to LISTSERV@asu.edu whose sole contents are SUB EPAA your-name.) As articles are published by the *Archives*, they are sent immediately to the EPAA subscribers and simultaneously archived in three forms. Articles are archived on *EPAA* as individual files under the name of the author and the Volume and article number. For example, the article by Stephen Kemmis in Volume 1, Number 1 of the *Archives* can be retrieved by sending an e-mail letter to LISTSERV@asu.edu and making the single line in the letter read GET KEMMIS V1N1 F=MAIL. For a table of contents of the entire ARCHIVES, send the following e-mail message to LISTSERV@asu.edu: INDEX EPAA F=MAIL, that is, send an e-mail letter and make its single line read INDEX EPAA F=MAIL.

The World Wide Web address for the *Education Policy Analysis Archives* is http://seamonkey.ed.asu.edu/

Education Policy Analysis Archives are "gophered" in the directory Campus-Wide Information at the gopher server INFO.ASU.EDU.

To receive a publication guide for submitting articles, see the *EPAA* World Wide Web site or send an e-mail letter to LISTSERV@asu.edu and include the single line GET EPAA PUBGUIDE F=MAIL. It will be sent to you by return e-mail. General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, Glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692)

Editorial Board

Dewayne Matthews
dm@wiche.edu

Mary P. McKeown
iadmpm@asuvm.inre.asu.edu

Les McLean
lmclean@oise.on.ca

Susan Bobbitt Nolen
sunolen@u.washington.edu

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
prohugh@ubvms.cc.buffalo.edu

Richard C. Richardson
richard.richardson@asu.edu

Anthony G. Rud Jr.
rud@sage.cc.purdue.edu

Dennis Sayers
dmsayers@ucdavis.edu

Jay Scribner
jayscrib@tenet.edu

Robert Stonehill
rstonehi@inet.ed.gov

Robert T. Stout
stout@asu.edu

EPAAcommentary

Contributed Commentary on
Volume 4 Number 17: Taylor & Nolen *Reframing Assessment Concepts*

19 December 1996

Jonathan A. Plucker
University of Maine

plucker@maine.maine.edu

I read with considerable interest Taylor and Nolen's (1995) recent article on reconceptualizing assessment concepts. As a faculty member who is responsible for teaching undergraduate and graduate education students these concepts and who is very concerned with the shortcomings of traditional methods for teaching such concepts, the description of efforts at the University of Washington was quite helpful and thought-provoking. However, the theoretical underpinnings of the course disappointed me -- I find the end agreeable but not the means.

A few caveats are worth mentioning before I share my comments. First, I hesitated to respond to the original article because I am afraid that any commentary will be perceived as a blind defense of traditional psychometrics -- a perception with which I am not comfortable. But I believe that educators and psychologists criticize standardized testing much too harshly and promote the advantages of alternative assessments far too enthusiastically. For example, alternatives to traditional psychometric approaches are often fraught with important problems (Plucker & Renzulli, in press; Ruiz-Primo & Shavelson, 1996), students from certain ethnic groups may be as culturally biased in favor of standardized testing as others are biased against it (Plucker, 1996), and both theoretical (Sternberg, 1994) and empirical evidence (e.g., Bridgeman & Morgan, 1996) suggests that individuals with specific learning styles may prefer standardized testing over alternative assessments. However, these reservations do not prevent me from researching the use of alternative assessments or working with my students to develop nontraditional assessments. Indeed, in two of my major areas of interest -- creativity and gifted education -- teacher checklists, performance-based assessments, and teacher/parent/peer nominations have been used for decades by educators and researchers. My views are heavily influenced by my work in both of these areas, and it is through this lens that my comments should be viewed.

The following five points are meant to be a springboard for future discussion, since the ideas raised by Taylor and Nolen (1996) are certainly timely and very important. First, is the growth of alternative assessment due to "a growing belief that the teacher can be one of the best sources of information about student learning" or is it due to a lack of satisfaction with traditional (i.e., standardized) assessment? Research on teacher accuracy in the assessment of students calls the opening statement into question (Guskin, Peng, & Simon, 1992; Hocevar & Bachelor, 1989; Holland, 1959; Pegnato & Birch, 1959; Plucker, Callahan, & Tomchin, 1996). This is a minor point, but historically an important one. For example, if the purpose of alternative assessment adoption is to create assessments that are less biased toward students from specific ethnic groups, then the bias inherent in alternative assessments becomes a stumbling block and focus of future

development efforts.

Second, the overarching issue may be that the techniques we use to teach measurement concepts -- not the content -- need to be improved. Taylor and Nolen note that teachers do not "perceive the information learned in traditional tests and measurement courses to be relevant" to classroom contexts, that "few teacher preparation programs provide adequate training for the wide array of assessment strategies used by teachers", and that "teachers do not believe they have the training needed to meet the demands of classroom assessment." They also discuss the ways in which measurement and assessment texts fail to aid the teaching of measurement concepts. Most individuals responsible for preparing future teachers would agree with the authors' summary. But rather than argue that our efforts and the texts fail due to insufficient theoretical foundations, why not argue that the content and text are merely passive objects that are actively manipulated by teachers to create learning experiences for students? The same argument is used by critics of the way we instruct future teachers to foster creativity, apply knowledge of motivation and behavior management, and even construct a realistic lesson plan. All of these areas are marked by a call for greater curricular application and application of principles of situated cognition, but not by a call for a complete revision of content. Why not? Because it may largely be unnecessary.

In the interest of brevity, I will not analyze the authors' reconceptualization of reliability and validity and the resulting description of the courses in detail. But readers should be aware that many of the underlying characteristics of the authors' work are really not any different than those addressed by traditional psychometricians. "Validity Dimension 1" is content validity, Dimension 2 is item or task analysis and construct validity, Dimension 3 is content validity again but from the perspective of the assessment, Dimension 4 is the detection of bias and criterion-related validity, and Dimension 5 is a consideration of the social implications of assessment and score interpretation. All of these concepts are certainly worthwhile (I especially like the emphasis on social implications), and most educators could provide dozens of studies that reinforce the inclusion of each dimension. And while the notion of the objective observer has held too much importance in the past, modern conceptions of reliability and validity (such as intra- and inter-rater agreement and criterion-related validity) tacitly acknowledge the fallibility and bias associated with assessment and evaluation. In most of the examples and discussion provided by Taylor and Nolen, familiar psychometric concepts and ideas are merely recast in postmodern terminology.

Fourth, abstract concepts (e.g., standard error of measurement [SEM]) and traditional standards of psychometric quality still need to be taught. Most, if not all, students take standardized tests as they progress through the educational system, and many of our future teachers will administer these tests and/or advise students who are about to take them. A case in point, and one that I use with my undergraduates, is the importance of standard errors of measurement. The students find this topic to be quite dry and lacking in application when I introduce the topic, but they begin to see the importance of SEM after we discuss several high-stakes applications (including school-by-school test score comparisons, which are known to influence parental decision-making in climates of school choice [Maddaus & Marion, 1995]). The question becomes one less of replacing traditional concepts and more of modifying our coursework and course sequences to include additional concepts. If qualitative inquiry has taught us nothing else, it has shown that multiple perspectives can be taught successfully within the framework of a single course.

Fifth, the most important issue may be the distinction between norm-referenced and criterion-referenced measurement. As Taylor and Nolen note, "classroom teachers are less interested in the consistency of student performance across similar measures than they are in whether students learn what [teachers] are teaching (the targeted constructs)." Alternative assessments used for high stakes (i.e., norm-referenced) purposes should be required to meet traditional standards of reliability and validity. As the authors state, "the meanings of assessment

in the context of the classroom must be considered carefully when large scale assessment programs decide to use classroom assessments for the purposes of district, state, or national accountability." At the same time, classroom- based assessment and evaluation used for primarily criterion- referenced purposes should be held to slightly different standards. Attention has been focused on the type of assessment and not how it is used (which is this most important aspect of measurement and evaluation).

Finally, given the validity concerns surrounding the use of alternative assessments (e.g., Ruiz-Primo, & Shavelson, 1996), educators must avoid the appearance of calling for new conceptions of reliability and validity because they cannot produce high quality alternative assessments as judged by traditional standards. While this was almost certainly not the authors' intent, it is not altogether impossible to understand why critics of alternative assessment infer this logic from our reasoning. Arguing that teachers and future teachers do not learn measurement concepts because of the way in which they are taught is reasonable; arguing that they do not learn measurement concepts because of how they are taught AND because the concepts are not applicable is more of a stretch.

In conclusion, I find much within the Taylor and Nolen article with which to agree. Indeed, if they had simply described their course which was "designed to engage students in tasks relevant to their own work as preservice teachers and demand that they consider assessment in the context of disciplinary structures and instructional practices," I would have filed the article away in a folder that was easily accessible for myself, my colleagues, and my students. But the authors' proposed reconceptualization of psychometric concepts is merely a presentation of the wolf in postmodernism's clothing. Educators need to begin questioning whether we need to replace our conceptualizations of and standards for psychometric quality or expand the conceptualizations and the teaching of them to incorporate fresh perspectives. The latter course is more reasonable and more feasible than the former.

References

Bridgeman, B., & Morgan, R. (1996). Success in college for students with discrepancies between performance on multiple- choice and essay tests. Journal of Educational Psychology, 88, 333-340.

Guskin, S. L., Peng, C.-Y. J., & Simon, M. (1992). Do teachers react to "multiple intelligences"? Effects of teachers' stereotypes on judgments and expectancies for students with diverse patterns of giftedness/talent. Gifted Child Quarterly, 36, 32-37.

Hocevar, D., & Bachelor, P. (1989). A taxonomy and critique of measurements used in the study of creativity. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.), Handbook of creativity (pp. 53-75). New York: Plenum Press.

Holland, J. L. (1959). Some limitations of teacher ratings as predictors of creativity. Journal of Educational Psychology, 50, 219-223.

Maddaus, J., & Marion, S. F. (1995). Do standardized test scores influence parental choice of high school? Journal of Research in Rural Education, 11, 75-83.

Pegnato, C. W., & Birch, J. W. (1959). Locating gifted children in junior high schools: A comparison of methods. Exceptional Children, 25, 300-304.

Plucker, J. A. (1996). Gifted Asian American students: Curricular and counseling concerns.

Journal for the Education of the Gifted, 19, 315-343.

Plucker, J. A., Callahan, C. M., & Tomchin, E. M. (1996). Wherefore art thou, multiple intelligences? Alternative assessments for identifying talent in ethnically diverse and economically disadvantaged students. Gifted Child Quarterly, 40, 81-92.

Plucker, J. A., & Renzulli, J. S. (in press). Psychometric approaches to the study of creativity. In R. J. Sternberg (Ed.), Handbook of human creativity.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. Journal of Research in Science Teaching, 33, 1045-1063.

Sternberg, R. J. (1994, Nov.). Allowing for thinking styles. Educational Leadership, 52(3), 36-40.

Taylor, C. S., & Nolen, S. B. (1995). "What does a psychometrician's classroom look like?": Reframing assessment concepts in the context of learning. Education Policy Analysis Archives, 4(17). (WWW URL: http://olam.ed.asu.edu/epaa)

EPAAcommentary

Contributed Commentary on
Volume 4 Number 17: Taylor & Nolen *Reframing Assessment Concepts*

19 December 1996

Rick Garlikov *demo42@uabdpo.dpo.uab.edu*
Catherine S. Taylor *ctaylor@u.washington.edu*
Susan Bobbit Nolen *sunolen@u.washington.edu*

The following is an e-mail exchange that Susan B. Nolen and Catherine S. Taylor had with Rick Garlikov in response to follow-up questions Garlikov had about their article "What Does the Psychometrician's Classroom Look Like?: Reframing Assessment Concepts in the Context of Learning" (EPAA Volume 4 Number 17)

Garlikov: I have read "through" (more than cursorily, but less than thoroughly) your EPAA paper, and I have some questions.

1) If teachers and pre-service teachers are evaluating their own assessment instruments, how can they see the flaws that they didn't see when they made up the instruments? Isn't it almost impossible to evaluate your own evaluation efforts very accurately? E.g., a friend made up a test one time where her students had to rank the developmental order of some five or more stages of development. But she took off for each one that was not in the correct stage -- i.e., first, second, third, etc. The problem was that if a kid missed the first one and then had the right order of the rest, but each one slid back one notch, they missed them all. But a kid who had no clue as to the order might get two or more in the right slot by just guessing. My point was that my friend didn't see the problem with this test question. She said each part was worth only five points, but actually, each part was worth much more, because any wrong answer COULD throw off the other ones. If SHE were evaluating her test, she would have said it was a good test. Yes, no? How does your course deal with this aspect of judging one's own evaluations of students?

Nolen: This is why the draft-feedback-revision loop is so critical. Our job as instructors is to find the flaws and point them out. Often this entails telling the intern how students might interpret a misleading or unclear question, playing the role of their students. In the course of the term, most students are not able to actually try out items and assessments, though some do. More try them out the following quarter when they are responsible for teaching a unit of this size.

The fact that we require students to construct model responses and scoring rules for their items, as well as write rationales for assessments, scoring, and objectives (and their relationship), provides another way for them to see the problems with their initial work. This comes through pretty clearly in their self-evaluations.

Garlikov: 2) It SEEMS to me, at least today, that if you cover all the things in your

course that you discuss in the paper, that kids will be hard-pressed to get an intuitive understanding of testing or evaluating students -- though they will probably get an intuitive understanding of the problems of testing or evaluating students. It seems to me there is too much specific, technical detail involved for understanding to become likely. It may be that I just tried to read too much in too short a time, but I felt the quantity and nature of the material you discussed made the concept of evaluating more complex than it had to be -- for a student. Your article seems good for someone who already understands general forms and limits of evaluations --gives nice details in a systematic and thorough way-- but I worry about how much a pre-service teacher could absorb of it all. Or do you teach about this in a way very different from the way you constructed this particular article?

Nolen: In the course we try to teach both by presenting some information through readings and lecture-ettes, and by having students construct a cohesive unit plan. Although the presentation part is necessary, they don't really begin to *learn* it until they try to put the theories and methods into action. This learing (we think) continues long after the course is over as they try to assess fairly and informatively in their own classrooms. There is a lot of information in the course, but I think it is made learnable in several important ways:

First, the fact that all of the activities are embedded in their unit plans, rather than merely appearing as unrelated activities. The unit description and goals lead to the learning targets (objectives), which along with their knowledge of the subject-matter disciplines leads to appropriate methods of instruction and assessment. Because the most important mode of instruction for us is also our mode of assessment, we model for them how assessment can be an instructional tool.

Second, we consciously draw on what they have studied in their other teacher prep courses, and expect them to use their knowledge from those courses to justify their assessment decisions. Thus the assessment content is seen, in part, as a natural extension of what they have been learning in the program.

Finally, remember that students have been assessed a lot by the time we see them, and most have also seen and/or assisted with assessment in their field experiences. Most are quite concerned about being able to fairly and accurately assess their own students. From our experience, this (along with subject-matter preparation) is sufficient base on which to build their knowledge of assessment. I'm not sure what you mean by intuitive understanding, exactly; I know that several students have told us that they can no longer think about planning instruction without thinking about assessment as part of instruction. It seems, for many, to have become part of their instructional schema.

As you say, the article seems good for those who already know something about assessment: That is to whom it is directed. We don't have our students read the article, we have them learn by doing with feedback.

Taylor: We know that learning in a University context is imperfect and that our students will continue to learn after they take our course. At this time, students take the assessment course during the 2nd or 3rd quarter of a 5 quarter sequence. Our students have an opportunity to use their work in the field after the course is completed and to give us feedback about how things are going. We literally have students tell us "thank you" in the halls during the quarters following our course. They claim that the thinking they learned to do in the course helps them develop strategies to focus their teaching better and to plan their teaching better. One even

said, "My students even thank you." As Susan said above, rather than this course resulting in superficial coverage of assessment concepts and skills disconnected from other elements of teaching, the course is set up so that they grapple with the meanings of assessment concepts in a meaningful context - their own plans for teaching subject matter. I guess I'd have to say that I don't know what "intuitive" means in the sense you are using, but a habit of mind about how to stay clear about the goals of instruction, how instruction helps students reach those goals, and how assessment actually assesses for students' learning of those goals seems to me to be a powerful "intuitive" process.

As a parent, much of what I see in my own children's school experiences is random or text book driven. My children are learning more about how to please the idiosyncracies of teachers than they are substantive conceptual or procedural understandings (or even social understandings). The kinds of assessments they have reinforce a notion that science (or social studies, or English) is a list of facts to be memorized or is a teacher whim. I ask them "what did you learn from that experience" fairly regularly and am dismayed by the responses.

Garlikov: As Susan knows from my EDPOLYAN/EDPOLICY writing, this is one of the things about schools that frustrates me the most. Some of it is due, of course, as you say, to assessment techniques that give the impressions they do, but I also suspect many teachers (and many adults in general) really DO think that these subject matters really are some specific body of facts that need to be learned or memorized. So they may actually be assessing in ways that reinforce what they intend to be teaching. Which, if so, is, of course, disappointing to me.

Taylor: I hope that what we give our students is a way of thinking that helps them, not only be technically better assessors, but better, more focused, more fair teachers who use assessment as ways to assess student learning and to communicate to students about what is important to learn. Because our students' first passion is helping students learn, many can embrace this view of assessment more easily than they can a view that portrays assessments as tools for dispassionate observers and graders. The spin off is also surprising. Today, in reviewing file for a teacher education scholarship, the winner was one whose cooperating and supervising teachers both stated that his teaching was focused, his goals for instruction were clear to students, and his assessments were "eagerly embraced" by students because they "knew what the purpose was."

Garlikov: Thanks for answering my questions. I think you pretty much addressed what I asked, but I have one more simple question, and then a much more important question.

Simple question: Do you think that your students understand that the kind of crucial feedback you give them when you go over their assessment plans, etc. is what they also need to see their students as doing if and when their students complain about particular items or scoring, etc.? That is, can your students generalize from the kind of thing you are doing in this regard, or is it that they pay attention to you because a) you are the teacher so they pay attention to your claims of invalidity (or some other sort of flaw), or b) you can articulate the flaws in their assessments extremely well and cogently. In short, are you able to get them to see that whenever one of their students might complain about the reliability or validity

(or fairness or whatever) of a particular evaluation tool, they need to really listen and try to figure out whether there is any merit to the claim?

Nolen: For some, yes. Probably for most in theory. We talk some about the power dynamics involved in these things, and in both the preceding ed psych course and this course we emphasize listening hard to students. (In fact, their major project for the previous course is to do just that: Listen hard to two students talking about what they learned in two consecutive class periods, and trying to explain why that's what they learned.) We are not the only ones in their program who model this, though not all profs do as much to encourage revision and rethinking because of time required.

Garlikov: Difficult question: *I* don't know how to construct good "formal", individual exams (or paper assignments) about philosophical/conceptual/logical issues that measures what I wanted students to learn. I only get some ideas about what students might have learned by continuing back-and-forth dialogue that further probes answers they give, questions they ask, and comments they make as we go along. This almost never ends up giving me the impression that their initial answers gave me; and often I am left with the vague feeling that if we carried it even further, they would either change or they would give me an even different/better understanding of what they know. So I am never happy with a time-slice assessment except in those few cases where students either seem to have learned nothing about a particular issue or where they seem to express remarkably perceptive, genuinely independently discovered, reflective views. Since much of what I think is important in schools IS of a philosophical or conceptual or logical nature, how can teachers in general design formal assessments that can be said, or shown, to accurately reflect what students really know or understand? And how can one tell such assessments do that.

An example of the problem is almost any discussion on EDPOLICY [an educational listserv forum to which Susan Nolen and Rick Garlikov subscribe], where it takes a number of responses back and forth for everyone to even be clear about what is being asked or responded to; if it occurs even then. The initial tendency is to feel that someone is dead wrong or terribly misguided in some way, perhaps in some cases not even very bright. But it usually turns out that they were making a fairly cogent interesting point that was just difficult to express in some way that everyone could understand it. Yet teachers or assessment "instruments" often don't give students a chance to clarify, discuss, argue, clarify some more, etc. Or do you have a way around that?

Nolen: We just try to model and talk about listening to students, especially trying to understand the responses that seem off the wall. We give examples from our own teaching (both in the u. and in public schools) of times we found great insight lurking in what seemed initially to be a wacky answer. And we try to model this during the (many) discussions we have in our classes. Capturing some part of the essence of a discipline, including the habits of mind or approaches as well as the big ideas and questions, is something our students are encouraged (required) to struggle with in several courses.. It often seems to come to a head in the assessment course, where they have to be very clear about what is important to learn and how they will know when their students have learned it, AND how that learning (and assessment) is what Bruner calls "intellectually honest" or what Schwab calls "true to the discipline." They will continue (we hope) to struggle with this throughout their

careers, as we do.

Garlikov: You both mentioned you were not certain what I meant by "intuitive" understanding when I wrote:

> It SEEMS to me ... that if you cover all the things in your course that you discuss in the paper, that kids will be hard-pressed to get an intuitive understanding of testing or evaluating students -- though they will probably get an intuitive understanding of the problems of testing or evaluating students.

You did answer it in a way that pleases me, however. Catherine's additional comments particularly helped in light of what Susan had already written about the "details". I take it now that, although when your students leave your course they have some particular techniques and methods for assessing students in ways that are valid and instructionally re-inforcing and useful, the most important thing you give them is a sense for how assessment needs to work in classroom teaching, and what the essential pitfalls and problems of assessment are. That is why I distinguished between an intuitive understanding of testing on the one hand, and an intuitive understanding of THE PROBLEMS (and issues) involved in testing. The latter is I think (1) the most important thing you could give your students, and (2) the only ingrained ("intuitive") thing you were likely to be able to give them. I didn't think that in one course you could teach pre-service teachers how to intuitively make up flawless assessments for their students.

I had had the FEELING as I had read the EPAA paper that there was a possibility you were claiming that you taught your students how to make up wonderful or perfect tests with no problem at all. But from your responses, I see that what you do (which is what I had hoped you were doing) is to give your students a really good understanding of what KINDS of things need to be done in evaluating students, and a really good understanding of why those KINDS of things are important. Insofar as you helped them learn how to design some specific assessments, that is good; but it is better that you have helped them understand the concept of assessment, since the specifics may change for them as they get into situations different from any you may have anticipated, or as their instructional ideas and principles change.

Thanks for answering. And 'Good For You Guys' for doing such a good job teaching this sort of thing, and for writing the EPAA paper about it.