DOCUMENT RESUME

ED 388 701 TM 024 095

AUTHOR Flaitz, Jim; Perdomo, Toni

TITLE Rethinking the Treatment of Traditional Assessment Topics in Light of a Movement toward Authentic Assessment in the Classroom.

PUB DATE Nov 94

NOTE 18p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Nashville, TN, November 9-11, 1994).

PUB TYPE Information Analyses (070) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Educational Assessment; Educational Innovation; Educational Trends; *Item Analysis; *Measurement Techniques; Methods Courses; *Teacher Education; Test Interpretation; *Test Reliability; Test Use; *Textbook Content

IDENTIFIERS *Authentic Assessment

ABSTRACT

A cursory examination of current measurement texts used in teacher education reveals a treatment of such topics as test reliability, item analysis, and test interpretation based largely upon classical test theory. In the meantime, the landscape of the classroom has been significantly impacted by a growing emphasis on more authentic assessment strategies. Most of the major measurement texts address this topic, often through the inclusion of a separate chapter, or perhaps a sidebar, but the treatment of reliability, item analysis, and test interpretation has typically remained largely intact over the years. This position paper aims to examine the relevant literature regarding the impact of the trend toward more authentic assessment on teacher training and on teacher practices, especially with regard to reliability, item analysis, and test interpretation. Discussion of how a movement toward greater reliance on authentic assessment strategies may, or should, impact teacher practice in other assessment areas is offered. The impact of authentic assessment on preservice teacher training is also considered. The measurement course instructor must give students knowledge and skills to maximize the potential benefits of authentic assessment while minimizing potential harms. (Contains 30 references.) (Author/SLD)

RETHINKING THE TREATMENT OF TRADITIONAL ASSESSMENT TOPICS IN LIGHT OF A MOVEMENT TOWARD AUTHENTIC ASSESSMENT IN THE CLASSROOM

JIM FLAITZ
TONI PERDOMO

THE UNIVERSITY OF SOUTHWESTERN LOUISIANA

USL

A PAPER PRESENTED AT THE
MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION
ANNUAL MEETING
NOVEMBER 9, 1994
NASHVILLE, TENNESSEE

RETHINKING THE TREATMENT OF TRADITIONAL ASSESSMENT TOPICS IN LIGHT OF A MOVEMENT TOWARD AUTHENTIC ASSESSMENT IN THE CLASSROOM

Jim Flaitz
Toni Perdomo
The University of Southwestern Louisiana

ABSTRACT

Most teacher education programs today are addressing the issues and skills associated with classroom assessment, often in courses in tests and measurement. Most of those courses employ a measurement textbook. A cursory examination of the current measurement texts reveals a treatment of such topics as test reliability, item analysis, and test interpretation based largely upon classical test theory. In the meantime, the landscape of the classroom has been significantly impacted by a growing emphasis on more authentic assessment strategies. To their credit, most of the major measurement texts are addressing this topic, often through the inclusion of a separate chapter, or perhaps a sidebar. However, the treatment of the aforementioned topics (reliability, item analysis, test interpretation) has typically remained largely intact over the years.

This position paper aims to examine the relevant literature regarding the impact of the trend toward more authentic assessment on teacher training and on teacher practices, particularly related to the topics of reliability, item analysis, and test interpretation. Discussion of how a movement toward greater reliance upon authentic assessment strategies may (or should) impact teacher practice in these other assessment areas, and of how authentic assessment should impact upon preservice teacher training programs is offered.

Introduction

Every classroom teacher is faced with the responsibility for assessing the achievement of the students in his or her classroom. The assessment practices of teachers, and the skills upon which those practices are based, are the product of a variety of experiences and influences (e.g., assessment practices of their own teachers, practices of the supervising teacher during student teaching, practices of fellow teachers in the schools where they teach). The primary purpose of teacher preparation programs is to develop in preservice teachers a set of skills, specialized knowledge, and values upon which sound and effective teaching practices can be based. One of the important areas of preparation is the area of student assessment (which could be said to subsume, overlap with, or exist as a sub-set of other descriptors such as measurement, evaluation and testing).

When investigators have looked at the assessment and grading practices of classroom teachers, the findings have often been a source of disappointment and concern (e.g., Brookhart, 1993; Frary, Cross, & Weber, 1993; Plake, Impara, & Fager, 1993; Stiggins, Frisbie, & Griswold, 1989). What appears to emerge from these studies is a picture of classroom teachers who persistently engage in testing, assessment, and grading practices that are seriously at odds with "recommended practice". Part of this pattern of departure from "best practice" may be attributable to an out-right lack of any formal training in classroom assessment (e.g., Hills, 1991; Schafer & Lissitz, 1987). Even for those teachers who have completed a required course in measurement as part of their professional education program, there remains an underlying suspicion that the prospects for significant departure from the practices emphasized in that formal training continue to be great. Some investigators (e.g., Brookhart, 1993) have examined teacher grading practices and what is revealed may be instructive in understanding why there is a gulf between what measurement specialists advocate and what classroom teachers actual do. As an example, Brookhart found that a large percentage of teachers view grades as the "pay" students earn for their work. If teachers are more inclined to view grading (and by extension other assessment practices) from a social values orientation, as Brookhart suggests, then it is reasonable to expect that measurement instruction, taught from a psychometric perspective, will more often than not fail to alter teacher beliefs or practices. In much the same way that college graduates persist in unscientific explanations for common phenomena (such as why its warmer in the summer than in the winter) even after receiving instruction regarding the phenomena, so teachers may be persisting in beliefs and practices in the area of assessment that run counter to recommended practice even when instructed in those preferred practices.

If the practices of classroom teachers are at odds with traditional recommended assessment practices, what can be expected in an area of assessment such as authentic assessment, which may not be particularly well addressed in teacher preparation programs? The balance of this paper will be given over to an examination of the background of authentic assessment and some of the issues associated with authentic assessment that can make it problematic as a classroom assessment strategy. Attention will be given to the "traditional" measurement curriculum, as it is represented in the major classroom measurement texts, with special attention given to how authentic assessment is being treated in those texts. The paper concludes with some suggestions regarding how preservice teachers may be better prepared to incorporate authentic assessment strategies into their classroom assessment practices.

Some Background on Authentic Assessment

Authentic assessment, also referred to as "performance," "direct," or "alternative" assessments (Fairtest, 1992), can be defined as a task that students must perform rather than selecting an answer from a paper and pencil test (Sweet & Zimmerman, 1992). Authentic assessment has also been defined as the examination of student performance on worthy intellectual tasks (Wiggins, 1990).

In an interview (Kirst, 1991), Lorrie Shepard described the term authentic assessment this way:

> Use of the term authentic assessment is intended to convey that the assessment tasks themselves are real instances of extended criterion performances, rather than proxies or estimators of actual learning goals (p.21).

According to Pierson and Beck, "performance assessment goes beyond what students know to measuring what students can do or apply" (Pierson & Beck, 1993). Performance assessments measure critical thinking skills (Simon & Soileil, 1993), as well as help to prepare learners for obstacle's they will have to face in the real world (Wiggins, 1990). Moreover, proponents for authentic assessment characterize traditional tests as "indirect... simplistic substitutes which we use to make judgments about student achievement" (Wiggins, 1990). Supporters of authentic assessment range from classroom teachers to researchers in various fields of education. Some researchers view performance assessment as an effective means to increase student interest, teach communication of ideas, accommodate different learning styles and develop critical thinking skills (e.g., Simon & Soileil, 1993). Linn & Burton describe performance assessment as reflecting "good instructional activities" and as "better reflections of the criterion performances that are of importance outside the classroom" (Linn & Burton, 1994). Authentic assessment has the advantage of reflecting "real-life challenges," "makes effective use of teacher judgment" (Marzano, 1994), and directly reflects student achievement (Fairtest, 1992).

Because the students must perform an observable task that demonstrates comprehension/achievement, authentic assessments are said to have face validity, which means different things to different people. Some researchers claim that face validity is not "validity in the technical sense" (Mehrens, 1992) or validity that represents accountability (Burger & Burger, 1994). However, other suppc·ers of authentic assessment suggest that face validity is an important characteristic of alternative testing methods (Wiggins, 1991) and should be considered carefully.

Despite the numerous advantages of authentic assessment, there are serious concerns that must be addressed. Much of the research concerning authentic assessment raises questions of validity, reliability, generalizability and accountability (Burger & Burger, 1994; Linn & Burton, 1994; Marzano, 1994; Mehrens, 1992; Miller & Legg, 1993; Willson, 1991). One researcher suggests that rather than center on reliability, the focus should be on validity issues (Bracy, 1993). Researchers have conducted studies to determine the validity of authentic assessments, with the findings showing that assessments having face validity may only "superficially measures what the test intends to measure" (Burger & Burger, 1994). This same study conducted by Burger & Burger shows promising possibilities for authentic assessment, but accountability remains a problem (Burger & Burger, 1994). Another study of the Vermont Portfolio project found "promising effects on instruction," but reliability, validity and rater reliability remained problematic (Koretz, Dtecher, Klein & McCaffrey, 1994).

There are also the issues of time and cost in implementation and grading (Guskey, 1994; Mehrens, 1992; Popham, 1993; Willson, 1991), but supporters of performance assessment counter this argument by purporting that since educators teach to the test, the test may as well be worth teaching to, which performance assessment is, according to these authors (Guskey, 1994; Sweet & Zimmerman, 1992). Another major problem with authentic assessment is the lack of teacher training (Guskey, 1994; Plake, Impara & Fager, 1993; Stiggins, 1991).

Validity and Authentic Assessment

It is an almost universally accepted belief that the most important quality of a test is its validity (or perhaps more properly, the validity of the interpretations of results of the test). Judging the validity of a performance assessment can be problematic because many of the traditional approaches to examining validity rely upon some form of convergence with other measures of the same trait or characteristic purportedly measured by the assessment. Since this sort of evidence would typically be based upon convergence to alternative paper and pencil assessments of those traits or characteristics, and since it is the very "artificial" nature of paper and pencil measures that authentic assessment has risen up against, clearly convergence with such instruments would be somewhat undesirable. In some quarters of the authentic assessment camp this dilemma has apparently revived the concept of "face validity" (e.g., Wiggins, 1991) which put simply would suggest that if it looks like the behavior you are interested in measuring, then it probably is. The problems with face validity that led to its falling into disfavor in the past won't be chronicled here, but the measurement instructor should be familiar with those problems.

In a review of relevant literature regarding validity in educational measurement, Moss (1992) observed that the emerging consensus among measurement scholars was that construct validity should occupy a central position, with such partial-evidence approaches as content and criterion-related validity taking a subordinate position. She pointed out that measurement textbooks were perpetuating the multiple validity concept:

> To this day, most of the popular textbooks, like the 1985 *Standards*, continue to organize presentations of validity around the three-part traditional framework of construct-, content-, and criterion-related evidence... (p:232).

In her review, Moss summarized three treatments of performance assessment validity. What all three have in common is a much more comprehensive description of the essential steps in assessment design and evaluation criteria than can be found in the typical measurement text. Another facet of validity that seems to be consistently associated with performance assessment (and not treated in measurement texts) is "consequences" (e.g., Messick, 1994), having to do with the impact assessment results will have on how teachers and students will spend their time, and how they will think about the goals of education.

In contrast, most texts provide basic information on the several dimensions of validity (construct, criterion-related, content), but suggest (or concede) that the approach most applicable in the classroom is content validity. Establishing and even ensuring the content validity of a classroom designed test can be achieved through the use of a table of specifications, which is reasonably well suited for use with conventional paper-and-pencil tests, but would seem to be a bit less appropriate as a tool for designing or critiquing a performance assessment. The strategies for developing performance assessments with validity offered by Frederiksen and Collins (1989), Haertel (1991) and by Linn, Baker, and Dunbar

(1991) all describe processes that are at least similar to the "test specifications" approach found in measurement texts, however the major measurement texts don't offer any concrete examples of how a classroom performance assessment might be judged insofar as validity is concerned (that is, how well the assessment matches the intended outcomes of instruction). There may be a special irony here, given that the advocates for authentic assessment seem to base a good part of their argument for authentic assessment on its greater validity. The larger point would seem to be that the measurement texts appear to have little of direct relevance to say about the validity of performance assessment approaches, beyond the general advice to be sure that the "test" measures what it is intended to measure. As with essay tests, the recommended practices for developing performance assessments include making the task as specific as possible, ensuring that the intended skills are represented in the performance of the task, and developing and using a scoring method that is appropriate to the measurement of the task, ideally one that can generate multiple indicators of the competence of the performance.

Reliability and Authentic Assessment

Authentic assessment seems to get its most severe criticism on the issue of reliability. With its dependence upon rater judgments and its limited sampling of learner behaviors, these criticisms would appear to be well-founded. On the other hand, the problem of reliability is one of degree, whatever the approach taken to assessment, and some distinction should be drawn between the high-stakes use of assessments (authentic or traditional) for such purposes as high school graduation, and the use of assessments in the classroom setting, where the stakes are typically lower. It might also be helpful to make the distinction between the approaches to reliability appropriate when the assessment is being interpreted in a norm-referenced fashion or a criterion-referenced fashion. Put simply, the classroom teacher needs to know about reliability in order to create, use and interpret authentic assessment, but the estimation model, and the issue of how reliable must the assessment results be, will be somewhat different for authentic assessment than for traditional paper-and-pencil tests.

Most of the major educational measurement texts continue to treat test reliability, from a relatively traditional perspective- based upon classical test theory. From that traditional perspective a trait, such as achievement, is viewed as an hypothetical construct. The construct is theoretically posited to vary among individuals, and the role of a test is to indicate or reflect the degree to which the trait (achievement) is present in a given individual. The reliability of that measurement would consequently be the degree to which the indicated level of achievement was an accurate reflection of the actual level of achievement.

Up to this point, the reasoning may be sensible to the preservice teacher, although it may seem an unnecessary exercise. However, when the focus shifts to how reliability is estimated, problems begin to arise. Suddenly we learn that reliability estimates will be "low" if there is little variability among the test scores. *"But what if there actually isn't much variability among the test takers' achievement? After all, isn't one of the intended results of teaching to make the test takers more alike in terms of their level of achievement?"* Only after it has been appreciated that most of the test reliability theory upon which reliability estimate procedures are based came directly out of a norm-referenced, "put the scores in rank-order" field of measurement activity does it begin to be clear why it's so important that scores are spread out. If a commercial test publisher is competing for market share and can promote its tests as "highly reliable," then obviously it's important to do whatever it takes to produce stable differences in test-takers' scores, including exaggerating small differences in achievement.

Once it becomes clear to teachers that the seemingly objective and statistical topic of test reliability is in fact intimately tied to the "purpose" of the test (to compare test performances to each other vs. to compare test performances to some standard), they may come to the conclusion that all this "reliability stuff" is for norm-referenced, commercially published tests, and can be safely disregarded. Interestingly, few textbooks ever actually point out the direct linkage of the reliability estimation methods they present (test-retest, internal consistency, equivalent forms) and the premise that the tests are being used for some norm-referenced purpose.

Unfortunately, this approach, which underlies the estimation models typically included (split-halves, test-retest, internal consistency), is not what classroom teachers intuitively view as the proper role of assessment. Even when their assessment efforts culminate in apparently relative judgments of student achievement (such as grading scales of A, B, C, etc.), they are much more inclined to interpret the results of tests as absolute indicators of achievement than as relative indicators best suited to the task of ranking their students. To the extent that the measurement texts actually address the reliability of criterion-referenced assessments, the treatment tends to consist of a description of a test-retest type of approach (or a form A, form B model).

Even as a classroom effort, many of the factors upon which reliability is dependent continue to be critical when considering authentic assessments. The sufficiency of the task for producing an adequate sample of student behavior would seem to be of concern, given the limitations on time and opportunity associated with the production of an authentic performance of a complex task. Likewise, some concern must be expressed for the issue of relative difficulty of a range of tasks. If large numbers of students are to be assessed in a manner that requires some overt performance of an example of an authentic task, then logistical issues may arise that would require a matrix of tasks, to ensure that a given student doesn't simply mirror the performance of a colleague. [Admittedly, this sort of problem wouldn't arise with a task such as writing an essay, or working a problem on paper, but on the other hand, the main criticism of existing approaches to assessing student achievement is that they are not the actual performance, but a kind of surrogate indicator of the actual skill. As such it only seems appropriate that at least some of the skills to be assessed authentically will have to incorporate some form of public performance].

The central reliability issue for authentic assessment is the matter of inter-rater reliability. Whether the basis for judging proficiency is an absolute standard or a norm-based relative standard, the matter of judgment still ultimately depends on the rater of the performance. Consequently, it is essential that the judgments rendered by any given rater are consistent, either with an established standard or with other raters. The manner in which this degree of consistency is estimated would depend, in part, on the manner in which proficiency is being judged. If the goal of the assessment is to differentiate among learners maximally, to ensure that assigned rankings are stable and consistent, then, to begin with, the rating scale or scales used would need to be capable of producing a significant range among the scores of students. (This step itself might prove to be self-defeating as a means of boosting reliability, since some previous research with the use of rating scales has suggested that having many potential points along the scale doesn't necessarily improve the precision with which ratings are assigned.) Subsequent to the actual eliciting of responses and scoring of those responses, some form of correlation of ratings, or assessment of inter-rater agreement, would be performed. [However, a word of caution might be appropriate here- most rating scales produce, at best, ordinal measures; therefore when selecting an analysis method, this characteristic should be kept in mind].

If the goal of the assessment is primarily to classify the students on a mastery/non-mastery dimension, then the issue of reliability shifts from one of differentiating among the test-takers to one of correctly classifying each test-taker into the appropriate category. Now the reliability of the scoring procedure is reflected in the degree to which raters can successfully discriminate between "products" that are representative of mastery and "products" that are representative of non-mastery. The errors in scoring would constitute the combined percentage of false positive and false negative ratings assigned. Inter-rater agreement approaches would again appear to be one available technique for gauging the reliability of judgments.

However, the real test for how well the rater is performing would be a matter of the degree to which the rater's judgments would agree with some established criterion against which mastery can be judged. That is to say, two raters might achieve a satisfactory degree of agreement regarding when a performance represents mastery and when it does not, but both raters may be operating from a definition of mastery that is at variance from some more officially sanctioned definition. Here, "calibration" training of the raters would seem to be especially crucial. A separate issue in gauging the sufficiency of the scoring rubric employed when judging for mastery/non-mastery is the matter of differential significance of errors. That is, are the consequences of a false positive (judging a student to have mastery when the student is not a master) more, less, or equally problematic as the consequences of a false negative (judging a student to be a non-master when the student is actually a master).

When measurement texts address inter-rater reliability they are more often than not relating the approach to the scoring of essays (which it might be argued represent one of the more conventional forms of performance assessment). In the treatment of inter-rater reliability, little if any attention is given to the notion of training of raters, nor is much offered of a concrete nature regarding the actual establishing of inter-rater reliability or the interpretation of any results of such a process.

When one thinks of classroom teachers adopting "authentic assessment" as a major basis for assessing student achievement, the image that is conjured doesn't include a systematic commitment to the training of teachers as raters, nor does it include a routine practice of teachers working in teams to independently rate the work of one another's students to establish the reliability of one another's ratings.

Grading and Authentic Assessment

When the measurement text covers the topic of grading, the typical treatment is that there are various methods of interpreting the scores obtained by students on their test or tests. Some of the methods (achievement judged against ability, achievement judged against effort, before and after comparisons of progress) are addressed mainly to point out the potential drawbacks associated with the approach. The two methods presented as serious contenders are relative standards and absolute standards. Each of these methods also has advantages and disadvantages, and the most typical advice that teachers seem to be offered is to know a lot about both and be prepared to employ whatever approach the school they end up teaching at requires.

Authentic assessment is by its very nature a response to a wide-spread concern that students exit their formal education without ever demonstrating their capacity for actually performing the skills and applying the knowledge that their education was intended to develop in them. If this is a valid expression of the intent behind "authentic assessment" then it would seem that the interpretation of student performances on such assessments would of necessity be criterion-referenced. In so far as the measurement texts and courses are concerned, here again we have to look at what is offered and how it is treated, relevant to criterion-referenced grading practices. While most texts offer some basic information regarding the construction of criterion-referenced tests, and the interpretation of such tests, they don't typically do a good job with some of the other important issues, such as how the criteria get set and who sets them, and what impact such an approach to interpretation of performances has on other practices, such as the very common practice of assigning letter grades. [Isn't this practice at least a little inconsistent with the use of assessment techniques that are aimed at producing evidence that the learner either possesses or does not possess an important skill?]

One of the points routinely made in measurement texts is that those grades that may carry the more serious consequences (end of term or end of year grades, for example) should be based on numerous independent sources of achievement information. This sound advice, however, runs into something of a problem if applied to authentic assessment. By their nature, these assessment strategies are very labor-intensive, both for teacher and for students, and can be reasonably expected to supplant, rather than to co-exist with other assessment strategies. One study has suggested that teachers may be able to incorporate only one performance assessment a month (Marzano, 1994). If teachers opt to abandon more traditional testing formats in favor of performance assessments, the consequence could be that the grades assigned based upon those performance assessments might turn out to be less valid, rather than more valid indicators of student achievement, because they are based on a more limited sample of student behavior.

Item Analysis and Authentic Assessment

Item analysis of test items typically entails the calculation of two indicators- the item difficulty index and the item discrimination index. The calculation of these two indicators of test item quality is usually associated with the items of selected-choice tests, and at a minimum normally require a scoring scheme (for the items) of correct-incorrect. The calculation of the item difficulty ordinarily involves determining the number of test takers who have answered a given test item correctly, and dividing that number by the total number of test takers. A common permutation of this approach is to first rank the test papers by total score (# items correct), then segregate the papers into three groups- an upper group representing those students whose test score would indicate a higher overall level of achievement (and perhaps might represent a "clear mastery" group) a lower group representing students with a distinctively lower overall level of achievement (and might constitute a "clear non-mastery" group), and a middle group, whose performance is in the mid-range of achievement (and whose mastery status might be in some doubt, due among other things to the reliability of the scores produced by the test). Using only the upper and lower groups, the item difficulty would be determined by summing the number of students in the two groups who answered a given item correctly and dividing this number by the total number of students in the two groups.

The second common indicator of test item quality, the item discrimination index, also requires the separation of the test papers into the three groups, as described above, and it is usually as a matter of efficiency that both indices are calculated from the upper group/lower group paper set. In the case of the item discrimination index, the number of students in the lower group answering the item correctly is subtracted from the number of students in the upper group answering correctly, and this difference value is divided by one half of the total number of papers in the two groups (the two groups being comprised of equal numbers of papers). The mechanics of calculation are reasonably straight-forward, and are equally applicable to test item data derived from either norm-referenced tests or criterion-referenced tests.

The interpretation of the two indicators, on the other hand, is considerably more problematic, and may be affected by various factors, including the purpose of the test (norm-referenced or criterion-referenced). For example, if the goal of a test is to "spread" the test scores (to provide a more reliable indication of student differences in achievement) then logic and experience argue for judging test items of moderate difficulty and high (positive) discriminating power as "best". If, on the other hand, the test is administered more as a gauge of student mastery of key skills, then the "appropriate" level of item difficulty might be considerably higher (meaning an easier item), especially if the test were administered at a point in time when most students would be expected to have achieved mastery of the relevant content and skills. Likewise, items might be judged entirely appropriate and suitable with modest or even no discriminating power between the two groups of test-takers, on the premise that many of the students, in both comparison groups, might be expected to have achieved mastery of many of the test items.

At another point in this paper the argument was offered that authentic assessments can be more readily viewed as criterion-referenced assessments than as norm-referenced assessments. This characterization seems most appropriate when considering the issues of task difficulty and task discrimination power (the degree to which the task, or elements of the task can discriminate between those students who have achieved mastery of the requisite tasks and those who have not). Interestingly, our review of relevant literature failed to turn up any instances of articles dealing with item analysis of performance assessment tasks or scales, nor any discussion of methods for examining task difficulty and task discrimination power. Much the same proved to be true in our examination of the major measurement texts. Is it possible that these topics simply have no applicability to performance assessments? This seems unlikely. Regardless of approach (paper-and-pencil or performance assessment) the validity of judgments regarding student achievement will be affected by the relative difficulty of the task(s) presented. Likewise, whatever method of assessing student achievement we use, we ought to expect at a minimum the capability for distinguishing between those students who have clearly mastered the requisite skills and/or knowledge, and those who have not (i.e., discriminating power).

There may be an implicit assumption with authentic assessments that in selecting assessment tasks that are "authentic", the issues of difficulty and discriminating power are rendered moot. That is, if it has been determined that this is what the learner should be capable of doing (without the surrogate intervention of some artificial approach to measuring the requisite skill), then the difficulty of the task will be, by definition, appropriate to its intended purpose, and the discriminating power will be self-evident in that if the student cannot perform the task, then the requisite skill is absent and if the student can perform the task, the requisite skill is present. This same line of reasoning could be (and quite possibly is in many instances) applied to any form of classroom assessment, including paper-and-pencil tests comprised of selected-response test items. The weakness of this argument is that it assumes the teacher's ability to develop assessment tasks (whatever form those tasks may take) that are well matched, in terms of difficulty, to the level of skill production that would be within the capacity of the "successful" student, and that are effectively able to discriminate the "successful" student from the "unsuccessful" student, without benefit of any external evidence of test item quality. Investing this degree of faith in the judgment of the teacher may be ill-advised. Experience in the paper-and-pencil test realm, where more empirical evidence of teachers' skills in developing test items high in these qualities is more readily obtained, suggests that even after receiving instruction on developing test items of high technical quality, the items produced by the typical classroom teacher will be at the knowledge level, will vary widely in difficulty level, and will discriminate to a substantially lesser degree than the teacher would have predicted.

Consider also that much of what affects difficulty in a performance assessment task will be the subjective judgment of the rater. Since the rater judges "success" or degree of success by comparison to private standards (or in those cases where the attempt has been made to define the standard in a more objective manner, the judgment as to whether what has been observed is sufficient evidence that the standard has been met), much of what will make a given task "difficult" lies in how rigorous or lenient the rater may be in judging whether the required standard has been met. Nevertheless, task difficulty could be gauged, crudely at least, simply by noting the number of students who were rated as having accomplished the task in comparison to the number who participated in the assessment.

Similarly, for those authentic assessments in which the scoring rubric provides for multiple criteria, corresponding to the several dimensions of the task, it is feasible to examine sub-task difficulty in much the same manner, and presumably gain useful information regarding the relative difficulty encountered by the learners, as a group, on each sub-task. Here, again, the issue of difficulty is in some measure dependent on the manner in which the separate scales are manifested and interpreted. That is, if the scales take the form of checklists, then presumably the scoring alternatives basically break down to "observed" or "not observed" (with possibly a third category- "no opportunity to observe"). If a multi-point rating scale is employed, then it might be necessary to identify a point along the rating scale continuum where the differentiation between "satisfactory" and "unsatisfactory" should be made.

What if multi-point scales are used for either the whole performance or the sub-tasks of the performance, without the a priori identification of a set position that differentiates the satisfactory from the unsatisfactory response, performance, or product (i.e., the task is not explicitly intended for a criterion-referenced interpretation)? Relative task (or sub-task) difficulty could still be examined through an alternative approach, simply by counting the number of students receiving each of the possible rating values. This information could, in turn, be represented in either tabular or graphical form, for ease of comparison. Easier tasks or sub-tasks would be those for which the majority or plurality of students received the most positive ratings, while more difficult tasks would be those for which the majority of ratings were from the lower end of the rating scale. Such group performance information would presumably serve the same useful function as feedback to both students and teacher as does the more traditional item analysis data.

Could the outcomes from an authentic assessment be analyzed for task or sub-task discrimination? The answer should be at least a qualified yes. If the task is comprised of multiple indicators of sub-task skills, and if the sub-task scores can be characterized as correct/incorrect (or even as satisfactory/unsatisfactory), then a procedure roughly similar to that employed with traditional test items could be applied. Presumably it would be of some constructive use to know which sub-tasks appeared to be most efficacious in discriminating those students who succeeded with the overall task or performance from those who did not. One likely modification of this approach, common to other criterion-referenced assessments, might be to use the overall task performance to define the "clear mastery" and "clear non-mastery" groups (rather than to rank papers and divide them into more arbitrary groups of roughly the top third, middle third, and bottom third). Each "item" could then be examined to see how many of the mas:ers and non-mastered performed satisfactorily on that sub-task.

As noted previously, in many instances the scoring rubric employed for the sub-tasks will consist of a rating scale with multiple points along a continuum, and for which no specific value position is identified as the point where mastery and non-mastery of the sub-task skill can be discerned. In such cases, a more involved statistical technique may be called for. For example, if the separate sub-task scales are ratings (on a scale of 1-5 or 1-10) and the overall score is a summation of the separate scales (or simply another rating on a similar scale), then a Spearman correlation coefficient between each sub-task scale score and the overall task scale score would reveal which sub-tasks were "discriminating" in a fashion consistent with the overall scale. If the overall rating was "collapsed" to simply indicate mastery or non-mastery, then an alternative statistical procedure, possibly based on chi square or proportional reduction in error (PRE) might be more appropriate (e.g., Healey, 1993).

Of course, since one of the underlying tenets of authentic assessment is that the task employed for assessment is authentic and consequently appropriate to judge student skills, the use of any of the aforementioned analysis techniques would differ somewhat from the more traditional test item applications in that we presumably are not likely to alter the features of the assessment (revise, replace or delete the "test items") on the strength of the analyses. On the other hand, the application of these procedures would almost certainly yield relevant information regarding other facets of the test data box- instructional effectiveness and student preparation.

There does exist a separate approach to the examination of test item characteristics based upon an entirely different set of assumptions than those associated with classical test score theory. That approach, most often referred to as Item Response Theory, suggests that it is possible to establish the individual test item characteristics (difficulty and discrimination power) without reference to the performance of the group on other test items. This approach has some promise for use with performance assessments, since estimations of task difficulty and discrimination power do not depend upon a self-referencing procedure, as is the case with classical test score theory-based approaches (Wainer, 1989). However, while large-scale, high-stakes test developers may be moving to the adoption of IRT based item analysis procedures, this approach is not, at present a particularly viable candidate for use by classroom teachers, given the relatively complex mathematics involved in item characteristics estimation.

Suggestions for Changing the Way Teacher are Trained

So what's a teacher educator, responsible for the preparing preservice teachers in the area of student assessment, supposed to do?

1. As a general recommendation, incorporate as much hands-on learning as possible. The biggest reason preservice teachers don't do what we say they should do is that they never try it until they get into a classroom, which is probably the worst setting for learning how to do the things we want them to do. It may also be necessary to incorporate meaningful practicum elements into measurement courses, where preservice teachers can be afforded an opportunity to try out their ideas and skills in a realistic setting, preferably under the close guidance and supervision of someone who is reasonably assessment literate.

2. Supplement the treatment given to performance assessment in the text with readings, exercises, and specific emphasis on how authentic assessment will impact or be impacted by the other key topics. Until textbooks treat apparently disparate topics in a more unified fashion, it will be up to measurement instructors to build the linkages and lead their students to the necessary appreciation of what consequences are likely to follow from the assessment choices made by teachers.

3. Recognize that teachers, with or without formal training show a strong tendency to view many assessment issues from a "values" perspective- seeing grades as "pay" and being inclined to reward effort along with achievement with the "coin of the classroom". Such beliefs and attitudes can be changed, where they should be changed, but only after it is appreciated that teachers are seeing assessment in a very different way than do measurement specialists.

4. Embed the topic of authentic assessment within a larger context of concern over the meaning of assessment results. Authentic assessment makes the most sense, and produces the fewest problems, if seen as one form of criterion-referenced assessment, in which the measurement issue is to construct a task that effectively evokes important learner skills and knowledge while the interpretation issue is to reliably judge the performance as either an instance of mastery of the task or non-mastery.

5. Emphasize the philosophy of- *the right tool for the right job*. Not every important learner outcome will be readily measured using a paper-and-pencil test, nor will all important learner outcomes be readily or effectively measured using authentic assessment techniques. Possessing requisite skills in the full range of assessment strategies, understanding the

strengths and weaknesses of each strategy, and having a larger vision of the purpose of education and of assessment within that framework are all critical elements in developing competence in classroom assessment.

6. Train preservice teachers as competent raters/judges of student performance. Nearly every article encountered concerning authentic assessment, whether in support of the practice or casting doubt on the practice, noted the primacy of teacher skills in rating and judging. Whereas in traditional testing the goal is to imbue the test with the qualities of reliability and validity, in authentic assessment *the teacher is the test*, in a sense, and those same qualities must be present in the teacher if the assessment is to yield valid and reliable information. As with most skills, the most effective method of developing the skill of judging is through systematic and recurring practice in the skill, under careful supervision and with insightful feedback. For the classroom teacher the sub-skills of rating will include skill in developing the criteria upon which a performance will be rated or judged, learning how to apply the criteria in a fashion that is relatively free of bias or the influence of irrelevant factors, learning how to interpret the results of such assessments, and learning how to set relevant standards against which performances can be judged. The importance of developing these skills is especially critical because the probability is very low that teachers are going to have the opportunity (at least in the near term) to have their ratings or judgments validated by the judgments of other teachers for the same students and the same tasks.

At least until the major measurement textbooks alter their treatment of many of the topics most relevant to the wise and effective use of authentic assessment, it will be up to the measurement course instructor to design an experience for developing a meaningful perspective on assessment generally, and authentic assessment specifically. The approach taken to topics of validity and reliability is going to have to be revisited, to mirror more closely the emerging views of what assessment validity involves as well as to more appropriately frame the concern with authentic assessment reliability in terms of purpose and consequence. Other topics addressed in this paper, such as grading practices and item analysis should likewise be addressed in the context of authentic assessment, even if the most reasonable conclusion is that the fit between the topic and authentic assessment is poor. For better or worse, authentic assessment is with us, and for the measurement course instructor the only intelligent recourse is going to be to provide his students with knowledge and skills to minimize the potential harm and maximize the potential benefits associated with this practice.

REFERENCES

Bracy, G. (1993). Assessing the new assessments. *Principal, 72* (3), 34-36.

Brookhart, S. (1993). Teacher's grading practices: Meaning and Values. *Journal of Educational Measurement, 30* (2), 123-142.

Burger, S. & Burger, D. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice, 13* (1), 9-14.

Frary, R.B., Cross, L.H., & Weber, L.J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12* (3), 23-30.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18* (9), 27-32.

Guskey, T. (1994). What you assess may not be what you get. *Educational Leadership, 51* (3), 51-54.

Haertel, E.H. (1991). New forms of teacher assessment. *Review of Educational Research, 17*, 3-29.

Healey, J.F. (1993). *Statistics: A Tool for Social Research.* (3rd. Ed.) Belmont, CA: Wadsworth.

Hills, J.R. (1991). Apathy concerning testing and grading. *Phi Delta Kappan, 72* (7), 540-545.

Kirst, M. (1991). Interview on assessment issues with Lorrie Shepard. *Educational Researcher, 20* (2), 21-23, 27.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and Implications. *Educational Measurement: Issues and Practice, 13* (3), 5-16.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 5-21.

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

Marzano, R. (1994). Lessons from the field about outcomes-based perfo/mance assessment. *Educational Leadership, 51* (3), 44-50.

Mehrens, W. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11* (1), 3-9.

Messick, Samuel. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessment. Educational Researcher, 23(2), 13-22.

Miller, D. & Legg, S. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12* (2), 9-15.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62* (3), 229-258.

National Center for Fair and Open Testing (Fairtest). (1992). What Is Authentic Evaluation? (Report no. TM 019 263). Cambridge, MA. (ED 352 373).

Pierson, C. & Beck, S. (1993). Performance Assessment: The Realities That Will Influence the Rewards. Childhood Education, 70, 29-32.

Popham, W.J. (1993). Educational testing in America: What's right, what's wrong? A criterion-referenced perspective. *Educational Measurement: Issues and Practice*, *12* (1), 11-14.

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

Simon, K. & Soileil, G. (1993). Alternative Assessment-Can Real World Skills Be Tested? (Report no. TM 020 649). Charleston, West Virginia: Appalachia Educational Lab. (ED 362 575).

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

Stiggins, R., Frisbie, D., and Griswold, P. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice*, *8* (2), 5-14.

Sweet, D. & Zimmerman, J. (1992). Performance assessment. *Educational Research Consumer Guide*, *2*, 2-4.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26* (2), 191-208.

Wiggins, G. (1990). The Case for Authentic Assessment. (Report no. TM 016 142). Washington, D.C.: American Institute For Research. (ED 328 611).

Wiggins, G. (1991). A response to Cizek. *Phi Delta Kappan*. *72*, 700-703.

Willson, V. (1991). Performance assessment, psychometric theory, and cognitive learning theory: Ships crossing in the night. *Contemporary Education*, *62*, (4), 250-254.