Measuring Teachers' Knowledge &
Application of Classroom Assessment
Concepts:
Development of the
Assessment Literacy Inventory

Craig A. Mertler, Ph.D.
(*mertler@bgnet.bgsu.edu*)
Division of Educational Foundations & Inquiry
School of Educational Leadership & Policy Studies
College of Education & Human Development
Bowling Green State University
Bowling Green, OH 43403

Cynthia Campbell, Ph.D.
(*ccampbell@niu.edu*)
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115

Abstract

Assessing student performance is one of the most critical responsibilities of classroom teachers; yet, many teachers do not feel adequately prepared for this task. Teachers often believe that they need remediation or assistance in applying assessment concepts and techniques, as well as making assessment-related decisions. In an effort to measure teachers' "assessment literacy," an instrument, titled the *Assessment Literacy Inventory* (*ALI*), was developed and its psychometric properties evaluated. The *ALI* was designed to parallel existing *Standards for Teacher Competence in the Educational Assessment of Students*. A two-stage pilot test of the instrument was conducted with 152 preservice teachers in Fall 2003 and 249 preservice teachers in the Spring 2004. Item analyses of the second-stage pilot data revealed an overall instrument reliability (KR20) of .74. Individual item analyses (i.e., item difficulties and item discriminations), as well as other indices, were examined. Recommendations for future research include content and construct validation of the *ALI* (both of which are currently being examined), as well as an investigation of the appropriateness of the *ALI* as a measure of inservice teacher assessment literacy. Finally, the *Assessment Literacy Inventory* provides a practical mechanism for educators to measure assessment literacy. Considering the current state of high-stakes accountability in education, the *ALI* could provide school districts an effective, as well as efficient way to allocate resources for developing or otherwise selecting teacher professional development opportunities on the topic of classroom assessment. (Contains 2 tables, 2 figures, & 1 appendix)

***Measuring Teachers' Knowledge & Application of
Classroom Assessment Concepts:
Development of the Assessment Literacy Inventory***

Introduction

Accurate assessment of student achievement is being more urgently called for at district, state, and national levels. Emphasis on raising standardized achievement scores has resulted in efforts to hold teachers accountable for improving how student assessment is conducted in their own classrooms. However, there exists a paradox in our educational system in that many teacher preparation programs do not require a course in classroom assessment as a requisite of graduation (Roeder, 1972; Schaefer & Lissitz, 1987; Stiggins, 1999; Wise, Lukin, & Roos, 1991). In addition, teachers report feeling inadequately prepared to meet this challenge (Murray, 1991). Consequently, classroom teachers are calling for more training due to their perceived lack of preparedness to assess their students, citing weaknesses in their undergraduate preparation programs (Plake, 1993).

Assessing student performance is one of the most critical responsibilities of classroom teachers. It has been estimated that teachers spend up to 50 percent of their time on assessment-related activities (Plake, 1993, Stiggins, 1999a). Yet, regardless of the amount of time spent, classroom assessment is a vitally important teaching function in that it contributes to every other teacher function (Brookhart, 1998, 1999b). According to Stiggins (1999), "The quality of instruction in any ... classroom turns on the quality of the assessments used there" (p. 20). For these reasons, information garnered from classroom assessments must be meaningful and accurate; i.e., the information must be valid and reliable (Brookhart, 1999a).

In recent years, public and governmental attention has shifted to school achievement as evidenced by performance on standardized achievement tests (Campbell, Murphy, & Holt, 2002). Moreover, there has

been an increase in expectations regarding teachers' assessment
expertise. Teachers have been required to develop classroom assessments
that align curriculum with state standards as a means of improving test
scores (Campbell, Murphy, & Holt, 2002). Research examining the
relationship between classroom assessments and student performance on
standardized tests reveals that improving the quality of classroom
assessments can increase average scores on large-scale assessments as
much as 3/4 of a standard deviation (as much as 4 grade equivalents or
15-20 percentile points), representing a huge potential (Stiggins,
1999). This is important research as it makes an empirical connection
between the quality of teachers' classroom assessments and students'
achievement as measured by standardized tests.

It is ironic, that despite the increased emphasis placed on
educational testing, assessment, and data-driven decision-making in U.S.
K-12 schools, many colleges of education and state education agencies
still do not require preservice teachers to complete specific coursework
in classroom assessment (Campbell, Murphy, & Holt, 2002; O'Sullivan &
Johnson, 1993). This continues to be an interesting phenomenon as many
inservice teachers report feeling ill-prepared to assess student
learning (Plake, 1993). Furthermore, teachers often claim that their
lack of preparation is largely due to inadequate preservice training in
educational measurement (Plake, 1993). For example, in a statewide
survey asking inservice teachers about their perceived level of
preparedness to assess student learning resulting specifically from
their teacher preparation programs, over 85% of the respondents reported

Yet, research has documented that teachers' assessment skills are
generally weak (Brookhart, 2001; Campbell, Murphy, & Holt, 2002).
Stiggins (2001) is in agreement when he states that we are seeing
unacceptably low levels of assessment literacy among practicing teachers
and administrators in our schools. He continues by stating that this
assessment *il*-literacy has resulted in inaccurate assessment of
students, thereby, preventing them from reaching their full potential.

that they were not well prepared (Mertler, 1999). When asked about their current level of preparedness, slightly more that half indicated that they were well prepared to assess student learning. Mertler (1999) concluded that this potentially implies that teachers tend to develop assessments skills on the job, as opposed to structured environments such as courses or workshops.

Stiggins (1999) has reiterated this implication, stating that many teachers are left unprepared to assess student learning as a result of both preservice and graduate training; they acquire what assessment "expertise" and skills they possess while on the job. Yet, despite beliefs that assessment skills are developed through trial and error in their classrooms, teachers have reported that the greatest influence on their assessment practices is formal coursework in tests and measurement (Wise, Lukin, and Roos, 1991).

When considering teachers' levels of assessment preparation, Plake (1993) found that over 70% of teachers responding to a national survey reported exposure to tests and measurement content (either through a course or inservice training), although for the majority it had been longer than 6 years. Inservice teachers who had previous coursework/training scored significantly higher on a test of assessment literacy than those who hadn't, but the difference was less than one point.

Recognizing the need for teachers to possess knowledge and skills in the area of classroom assessment, a joint effort between the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) was undertaken in 1987 to "develop standards for teacher competence in student assessment out of concern that the potential educational benefits of student assessments be fully realized" (AFT, NCME, & NEA, 1990). The standards were developed to address the problem of inadequate assessment training for teachers (AFT, NCME, & NEA, 1990). *The Standards for Teacher Competence in the Educational Assessment of Students*

specifies that classroom teachers should be skilled in: Choosing and Developing Assessment Methods; Administering, Scoring, and Interpreting Assessment Results; Using Assessment Results for Decision Making and Grading; Communicating Assessment Results; and Recognizing Unethical Assessment Practices.

These *Standards* essentially describe the extent to which an educator is assessment literate. "Assessment literacy" has been defined as follows:

> Assessment literate educators recognize sound assessment, evaluation, communication practices; they
> - understand which assessment methods to use to gather dependable information and student achievement.
> - communicate assessment results effectively, whether using report card grades, test scores, portfolios, or conferences.
> - can use assessment to maximize student motivation and learning by involving students as full partners in assessment, record keeping, and communication (Center for School Improvement and Policy Studies, Boise State University, n.d.).

A similar description is provided by Stiggins (1995), stating that "Assessment literates know the difference between sound and unsound assessment. They are not intimidated by the sometimes mysterious and always daunting technical world of assessment" (p. 240). He notes that assessment-literate educators (regardless of whether they are teachers, administrators, or superintendents) enter the realm of assessment knowing what they are assessing, why they are doing it, how best to assess the skill/knowledge of interest, how to generate good examples of student performance, what can potentially go wrong with the assessment, and how to prevent that from happening. They are also aware of the potential negative consequences of poor, inaccurate assessment (Stiggins, 1995).

Although *The Standards for Teacher Competence in the Educational Assessment of Students* are somewhat dated, they continue to address many of the important facets of classroom assessment knowledge, skills, and competence. However, Stiggins (1999b) asserts that these standards are not nearly comprehensive enough in their coverage to definitively represent how to prepare teachers for the realities they will face in

their classrooms and with their students. Specifically, he lists seven competencies, many of which are covered by *The Standards*. The competencies listed by Stiggins (1999b) are:

- Connecting assessments to clear purposes

- Clarifying achievement expectations

- Applying proper assessment methods

- Developing quality assessment exercises and scoring criteria and sampling appropriately

- Avoiding bias in assessment

- Communicating effectively about student achievement

- Using assessment as an instructional intervention (pp. 25-27)

While there is some debate about the extent to which *The Standards* adequately address those competencies which research shows that teachers need to possess, Table 1 shows that there is a great deal of overlap in the original 1990 *Standards* and the competencies listed by Stiggins (1999b).

Insert Table 1 about here

In 1991, a national study was undertaken devise an instrument to measure teachers' assessment literacy (Plake, 1993). *The Standards* were used as a test blueprint for the development of the *Teacher Assessment Literacy Questionnaire* used in the study. A representative sample from around the United States was selected to participate; a total of 98 districts in 45 states surveyed, yielding a total usable sample of 555 respondents (Plake, 1993). The KR20 (r_{KR20}) reliability for the entire test was equal to .54 (Plake, Impara, & Fager, 1993). The researchers concluded that teachers were not adequately prepared to assess student learning, as evidenced by the average score of 23 of 35 items correct (66%).

A similar study, conducted by Campbell et al. (2002), attempted to apply the identical previously described assessment literacy instrument to undergraduate preservice teachers. The renamed *Assessment Literacy Inventory* (*ALI*) was administered to 220 undergraduate students following completion of coursework in tests and measurement. The data from the undergraduate preservice teachers exhibited a higher level of reliability (r_{KR20} = .74) than their inservice counterparts in the Plake et al. study (Campbell, Murphy, & Holt, 2002). The preservice teachers (*M* = 21) averaged two fewer questions answered correctly than did the inservice teachers (*M* = 23).

Mertler (2003b) studied the assessment literacy of both preservice and inservice teachers, and then statistically compared the two groups. Using a slightly modified version of the *Teacher Assessment Literacy Questionnaire*, he obtained similar results to both the Plake et al. (1993) and Campbell et al. (2002) studies. The average score for inservice teachers was equal to 22 items answered correctly—quite similar to the average score of 23 obtained by Plake (1993). Reliability analyses also revealed similar values for internal consistency (r_{KR20} = .54 and .57 for the original study and the study at hand, respectively). The average score for the preservice teachers was equal to 19—also similar to the average score obtained by Campbell et al. (2002). Reliability analyses revealed identical values (r_{KR20} = .74) for internal consistency.

It is interesting to note that both the Campbell et al. (2002) and the Mertler (2003b) study were in essence replications of the Plake (1993) study, in that both used the same original instrument developed by Plake. When the instrument was administered to inservice teachers, it demonstrated consistent, however, poor psychometric qualities [i.e., r_{KR20} = .54 (Plake, 1993), r_{KR20} = .57 (Mertler, 2003b)]. When used with preservice teachers, the instrument demonstrated identical and much improved reliability [i.e., r_{KR20} = .74 (Campbell et al., 2002); r_{KR20} = .74 (Mertler, 2003b)]. Additionally, the original instrument was difficult

to read, extremely lengthy, and contained items that were presented in a decontextualized way. Both Campbell et al. (2002) and Mertler (2003b) recommended a complete revision and/or redevelopment of the assessment literacy instrument.

Purpose of the Study

The purpose of this study was twofold: (1) to develop an instrument that could accurately measure teachers' assessment literacy, and (2) to determine the psychometric qualities of this instrument.

The research questions addressed in the study were:

<u>Research Question 1</u>: What are the psychometric properties of the *Assessment Literacy Inventory* (*ALI*) when used with preservice teachers?

<u>Research Question 2</u>: Could the *ALI* serve as a useful instrument for evaluating preservice competency in classroom assessment?

Methods

During the spring and summer of 2003, the researchers, both with specific expertise in issues of classroom assessment, drafted an instrument titled the *Assessment Literacy Inventory*, hereafter referred to as the *ALI*. The *ALI* consisted of 35 items, embedded within five classroom-based scenarios, featuring teachers who were facing various assessment-related decisions. An example of one of the scenarios, including three of its seven items, is provided in the Appendix in order to give the reader an idea of the contextualized nature of the classroom-based scenarios and related items as they appear on the *ALI*. Each scenario presented a brief classroom situation followed by seven multiple-choice items. Each of the seven items within a single scenario were written to directly align to one of the seven *Standards for Teacher Competence in the Educational Assessment of Students* (AFT, NCME, & NEA, 1990). Following item construction, items were reviewed by the researchers, to check for alignment with the standards, as well as

clarity, readability, and accuracy of keyed answers. Items that raised questions regarding alignment, clarity, wording, or correctness of answer were revised. Judgmental review continued until consensus was reached regarding item appropriateness and quality.

During fall of 2003, an initial pilot test was conducted with undergraduate preservice teachers enrolled in introductory classroom assessment courses. One hundred fifty-two preservice teachers from the two large Midwestern institutions completed the ALI in an attempt to measure their assessment literacy. It is important to note that the undergraduate introductory assessment courses are a requirement for graduation at both institutions, and that course content, objectives, assignments, and experiences are designed to align with the seven *Standards for Teacher Competence in the Educational Assessment of Students* (AFT, NCME, & NEA, 1990). In addition, students from both institutions take the required assessment course prior to student teaching.

A complete item analysis was conducted on the resulting data using the Test Analysis Program (TAP) (Brooks & Johanson, 2003). Analyses included overall test analysis, individual item analyses, reliability analyses, and options (i.e., distractors) analysis.

Following an examination of the item analysis, the researchers made appropriate revisions to items appearing on the *ALI*. A second phase of data collection occurred in the spring of 2004 with 250 undergraduates following their completion of tests and measurement coursework. Analyses of the data were conducted using SPSS (v. 11) and TAP (v. 5.2.7).

Results

The initial pilot test of the *Assessment Literacy Inventory* with 152 preservice teachers revealed an overall KR20 (r_{KR20}) reliability equal to .75. The mean item difficulty was equal to .64 and the mean item discrimination was equivalent to .32. These values indicate that the *ALI* appeared to function reasonably well, from a psychometric perspective.

Further reliability analyses revealed that only four of the 35 items, when removed from the scale, resulted in an improved overall reliability. Based on this fact, the instrument was slightly revised in an attempt to improve its overall reliability, as well as other psychometric properties.

The second phase of pilot testing with the revised *ALI* was conducted following Spring 2004. To determine the appropriateness of analyzing data from the two institutions together, we established institutional similarity by examining the means, standard deviations, and reliability coefficients, as well as statistically comparing the total scores on the *ALI* across the two groups (see Table 2). After deleting outliers with standardized total scores (i.e., *z*-scores) exceeding +/-3.00 (of which there was only one case), the total scores were compared for the first (*M* = 24.50, *SD* = 4.92) and second (*M* = 22.98, *SD* = 4.05). No significant difference was found between total *ALI* scores for the two institutions, *t*(247) = 2.558, *p*>.01, two-tailed.

Insert Table 2 about here

Examination of the item analysis results from this phase revealed a value similar to that resulting from the first phase for instrument reliability, r_{KR20} = .74. Across the 35 items appearing on the *ALI*, item difficulty values ranged from a low of .212 to a high of .992; the mean item difficulty was equal to .681. The entire distribution of difficulty values is presented in Figure 1.

Insert Figure 1 about here

Across the 35 items appearing, item discrimination values ranged from a low of .014 to a high of .641; the mean item discrimination was

equal to .313. The entire distribution of difficulty values is presented in Figure 2.

Insert Figure 2 about here

Additionally, the analysis showed that only three—Items 17, 21, and 32—of the 35 items, when removed from the scale, resulted in an improved overall reliability. Furthermore, it is important to note that these improvements in overall reliability were extremely small (i.e., +.001, +.003, and +.002, respectively, for the three items).

Discussion

The psychometric qualities of the *ALI* strongly support its use as an acceptable measure of teachers' assessment literacy. The fact that the *ALI* demonstrated an overall reliability coefficient of .74 is consistent with recommendations in the literature regarding measures which result in high or good reliability. For example, Kehoe (1995) recommends that reliability values as low as .50 are satisfactory for short tests (10-15 items), though tests with over 50 items should yield KR-20 values of .80 or higher. Chase (1999) has suggested that for this type of test, reliability coefficients should be no lower than .65, but preferably higher. Similarly, Nitko (2001) advocates for reliability coefficients that range between .70 and 1.00. With its 35 items, the overall reliability demonstrated by the *ALI* in this study place the instrument well within these ranges.

Considering characteristics of individual items on the *ALI* also seem to demonstrate the instrument's effectiveness. With respect to item difficulty, Kehoe (1995) states that, on a good test, *most* [emphasis added] items on a test will be answered correctly by 30% to 80% of the examinees. On the *ALI*, 25 of the 35 items fell within this range. Chase (1999) recommends a slightly broader range for effective item difficulties—from .20 to .85; 28 of *ALI*'s items fell within this range.

The seven remaining items showed higher difficulty levels (i.e., they were "easier" items). Mertler (2003a) argues that on a criterion-referenced test—such as the *ALI*—a high difficulty level is a *good* thing as it serves as a clear indicator that examinees have mastered a specific concept.

Finally, with respect to item discrimination, Chase (1999) states that discrimination values of .30 and higher indicate fairly good item quality. Twenty of the 35 items appearing on the *ALI* had discrimination values greater than .30. It should be noted that of the remaining 15 items, 7 had fairly high difficulty levels (> .80). For items on which the vast majority of examinees identify the correct answer (i.e., the difficulty value is high), one could not expect to have good discrimination between the high and low scoring groups; it is not mathematically possible. Since both groups would achieve similar difficulty values, there would be very little discrepancy between respective difficulty values (i.e., the discrimination value would be low).

When examining preservice teachers' overall performance on the *ALI*, it should be noted that their score was far lower than might otherwise be expected given their recent completion of coursework in classroom assessment. Despite explicit efforts to link course content, assignments, and experiences characteristic of educational decisions and practices outlined in *The Standards for Teacher Competence in the Educational Assessment of Students* (AFT, NCME, & NEA, 1990), preservice teachers' mean score was 23.83 out of a possible 35 points, or approximately 68% of items answered correctly. Possible reasons for the observed gap between their *ALI* performance and recent formal training may be related to preservice teachers' limited classroom experience. Perhaps because the *ALI* is specifically designed to measure the real-world applications of assessment concepts and competencies outlined in The Standards, limited familiarity and experience with the day-to-day

realities of the classroom may have precluded preservice teachers from making necessary connections.

The role of teaching experience may be too important to overlook. For example, following a two-week, intensive professional development training course in classroom assessment, a small group of inservice teachers (*n* = 7) completed the *ALI* as a measure of assessment skills and knowledge. Similar to preservice teachers in the present study, the inservice teachers had not had previous coursework in classroom assessment. Although the *ALI* was administered following formal instruction in both groups, average test scores of inservice teachers were much higher (i.e., 28.29 out of a possible 35 points, or approximately 81% of items answered correctly) than scores of this study's preservice teachers tested under similar conditions. While the inservice sample is too limited to claim a link between first-hand teaching experience and assessment competency, it may, however, highlight the importance of experience in providing an important contextual base for transforming theory (with potentially abstract concepts) into real-world practice. Research is needed to examine the extent to which teaching experience influences the development of assessment competency (as measured by the *ALI*) by comparing preservice groups who complete assessment coursework at different points in their teacher preparation program. Examining the assessment skills of students from teacher education programs that vary the placement and sequencing of assessment coursework (e.g., before, during, or following student teaching) may help to disentangle the effects of experience on acquiring assessment skills as applied to educational decision making.

Although when used with preservice teachers the reliability of the Plake and Impara instrument produced similar reliability to the *ALI*, the user friendly format of the *ALI* (seven items relating to a single scenario for a total of 5 scenarios, 35 items) may reduce cognitive overload associated with reading 35 unrelated items, each containing its own unique scenario description. Moreover, the *ALI* may likely be more

appealing to the test taker because of the 35 item's thematic connection to a running story. Consequently, test takers may be more motivated and willing to complete the *ALI*. Still, we recognize that further research using the *ALI* with preservice teachers is needed to identify ways to improve reliability and establish validity evidence. Construct validity evidence is currently being examined through confirmatory factor analysis to identify whether the proposed 7-factor structure corresponding to *The Standards for Teacher Competence in the Educational Assessment of Students* is observed. To determine concurrent validity evidence, the connection between preservice teachers' *ALI* score and their end of semester percentage points earned in assessment coursework is currently being examined.

Moreover, it is recommended that further studies also be undertaken to ascertain the appropriate use of the *ALI* as a measure of inservice teacher assessment literacy, as well. Although some preliminary work in this area has been explored, formal studies of the validity, reliability, and appropriateness of the *ALI* as a measure of inservice teacher assessment literacy could potentially lend credibility to its use as a diagnostic instrument, specifically geared toward the identification — and ultimate remediation — of classroom assessment weaknesses or misconceptions.

The day-to-day work of classroom teachers is multifaceted, to say the least. However, none of these daily responsibilities is more important — or more central — to the work of teachers than that of assessing student performance (Mertler, 2003a). Previous studies have reported that teachers feel — and actually *are* — unprepared to adequately assess their students (e.g., Mertler, 1999, 1998; Plake, 1993). They often believe that they have not received sufficient training in their undergraduate preparation programs to feel comfortable with their skills in making assessment decisions.

The *Assessment Literacy Inventory* provides a mechanism for educators to measure assessment literacy (i.e., their knowledge of and

abilities to apply assessment concepts and techniques to inform decision-making and guide practice). Considering the current state of high-stakes accountability in education, the *ALI* could provide school districts an effective, as well as efficient way to allocate resources for developing or otherwise selecting teacher professional development opportunities on the topic of classroom assessment. Because the *ALI* is based entirely on the *Standards for Teacher Competence in the Educational Assessment of Students*, its use could provide educational leaders with a diagnostic tool for identifying areas (i.e., as represented by a given standard) where teachers may be deficient and in need of further remediation and training. In this way, such efforts could provide school districts with a roadmap for facilitating teachers' knowledge and application of assessment concepts and techniques, thereby improving the accuracy of educational decisions contributing to student learning and school improvement.

References

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *The Standards for Competence in the Educational Assessment of Students*. Retrieved July 22, 2003, from http://www.unl.edu/buros/article3.html

Brookhart, S. M. (1998). *Teaching about grading and communicating assessment results*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA. (ERIC Document Reproduction Service No. 419838)

Brookhart, S. M. (1999a). *The art and science of classroom assessment: The missing part of pedagogy*. Washington, D.C.: ERIC Clearinghouse on Higher Education and Office of Educational Research and Improvement.

Brookhart, S. M. (1999b). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice*, *18*(1), 5-13.

Brookhart, S. M. (2001). *The <u>Standards</u> and classroom assessment research*. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, TX. (ERIC Document Reproduction Service No. ED451189)

Brooks, G. P., & Johanson, G. A. (2003). Test Analysis Program. *Applied Psychological Measurement, 27,* 305-306.

Campbell, C., Murphy, J. A., & Holt, J. K. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Center for School Improvement and Policy Studies, Boise State University. (n.d.). What is assessment literacy? In *Assessment literacy*. Retrieved July 24, 2003, from http://csi.boisestate.edu/al/

Chase, C. I. (1999). *Contemporary assessment for educators*. New York: Longman.

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation* , 4(10). Retrieved August 18, 2004 from http://PAREonline.net/getvn.asp?v=4&n=10.

Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education*, *120*(2), 285-296.

Mertler, C. A. (2000). Teacher-centered fallacies of classroom assessment validity and reliability. *Mid-Western Educational Researcher*, *13*(4), 29-35.

Mertler, C. A. (2003a). *Classroom assessment: A practical guide for educators*. Los Angeles, CA: Pyrczak.

Mertler, C. A. (2003b, October). *Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference?* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, Ohio.

Murray, D. R. (1991). *An analysis of the perceptions of teacher education graduates on their preparation for service: A five year review*. Eric Document: ED335349.

Nitko, A. J. (2001). *Educational assessment of students*. Upper Saddle River, NJ: Merrill.

O'Sullivan, R. G. & Johnson, R. L. (1993). *Using performance assessments to measure teachers' competence in classroom assessment*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA. (ERIC Document Reproduction Service No. 358156)

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher*, *6*(1), 21-27.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, *12*(4), 10-12, 39.

Roeder, H. H. (1972). Are today's teachers prepared to use tests?
Peabody Journal of Education, 49(3). 239-240.

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

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta
Kappan*, 77(3), 238-245.

Stiggins, R. J. (1999). Are you assessment literate? *The High School
Journal*, *6*(5), 20-23.

Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment.
Educational Measurement: Issues and Practice, *20*(3), 5-15.

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.

Table 1. Comparison of Stiggins' (1999b) Classroom Assessment Competencies and *The Standards for Teacher Competence in the Educational Assessment of Students* (1990)

Stiggins (1999b) Competencies	The Standards for Teacher Competence in the Educational Assessment of Students (1990)
Competence 1: *Connecting assessments to clear purposes*	Standard 1: *Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.* Standard 2: *Teachers should be skilled in developing assessment methods appropriate for instructional decisions.* Standard 4: *Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.* (Also addressed in section titled *The Scope of a Teacher's Professional Role and Responsibilities for Student Assessment*)
Competence 2: *Clarifying achievement expectations*	Standard 4: *Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.* (Also addressed in section titled *The Scope of a Teacher's Professional Role and Responsibilities for Student Assessment*)
Competence 3: *Applying proper assessment methods*	Standard 1: *Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.* Standard 2: *Teachers should be skilled in developing assessment methods appropriate for instructional decisions.*

Table 1 (continued).   Comparison of Stiggins' (1999b) Classroom
Assessment Competencies and *The Standards for Teacher
Competence in the Educational Assessment of Students* (1990)

Stiggins (1999b) Competencies	*The Standards for Teacher Competence in the Educational Assessment of Students* (1990)
Competence 4: *Developing quality assessment exercises and scoring criteria and sampling appropriately*	Standard 2: *Teachers should be skilled in developing assessment methods appropriate for instructional decisions.* Standard 5: *Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.*
Competence 5: *Avoiding bias in assessment*	Standard 5: *Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.* Standard 7: *Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.*
Competence 6: *Communicating effectively about student achievment*	Standard 6: *Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.*
Competence 7: *Using assessment as an instructional intervention*	Standard 3: *The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.* Standard 7: *Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.* (Also addressed in section titled *The Scope of a Teacher's Professional Role and Responsibilities for Student Assessment*)

Table 2. Descriptive Statistics for the Total *ALI* Scores for the Two
         Institutions Studied

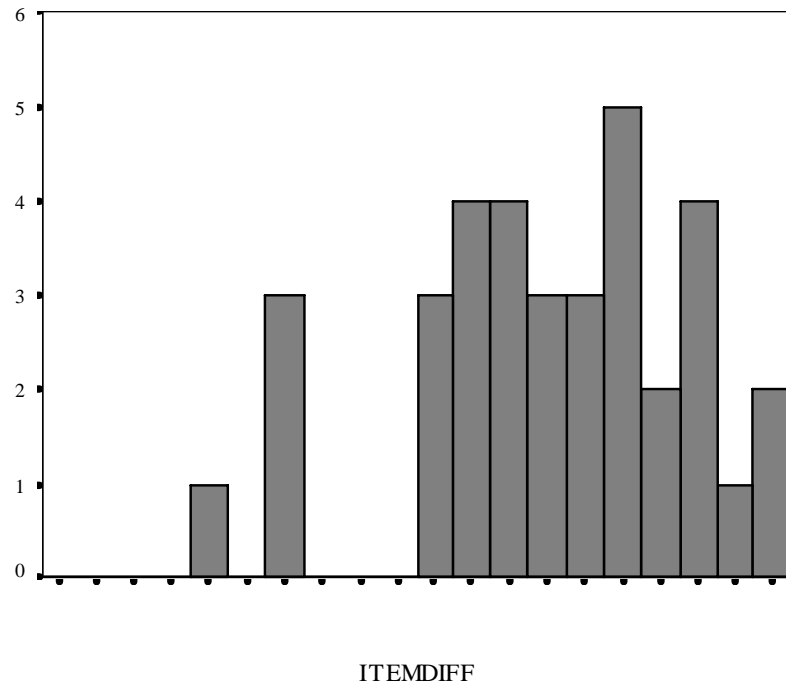| Institution	*n*	Mean	Standard Deviation	r_{KR20}
Institution #1	150	24.50	4.92	.78
Institution #2	99	22.98	4.05	.62
Total	249	23.90	4.64	.74

ITEMDIFF

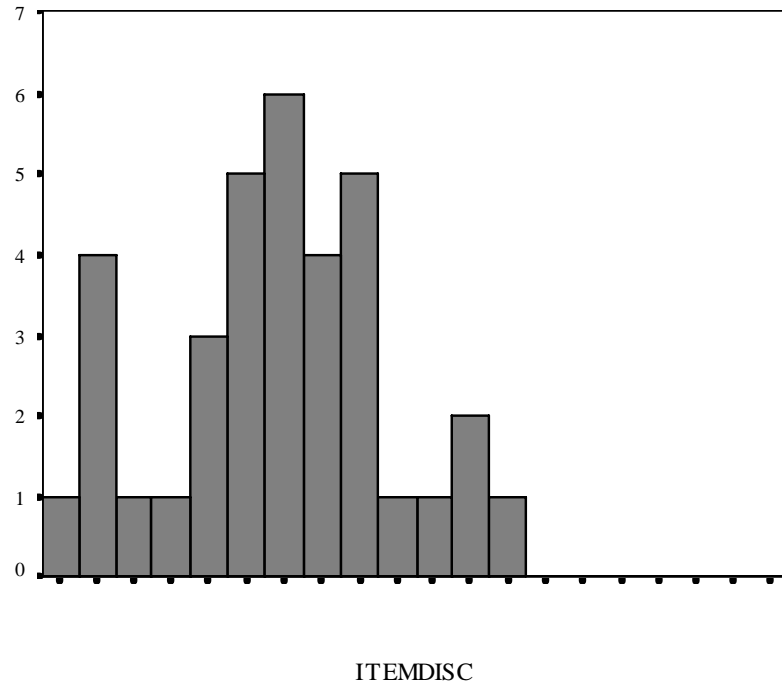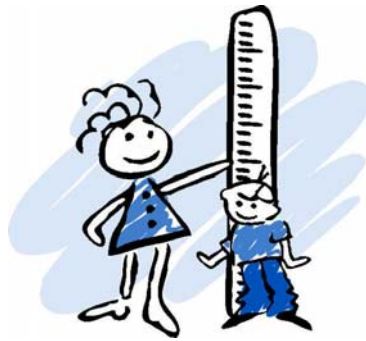*Figure 1*. Distribution of *ALI* Item Difficulty Values.

Figure 2. Distribution of *ALI* Item Discrimination Values.

Appendix

Sample Scenario and Selected Items from

The *Assessment Literacy Inventory* (*ALI*)

*A*ssessment
*L*iteracy
*I*nventory

Cynthia Campbell, Ph.D.
Northern Illinois University

and

Craig A. Mertler, Ph.D.
Bowling Green State University

<u>Description of the *ALI*</u>:

The **Assessment Literacy Inventory** (**ALI**) consists of five scenarios, each followed by seven questions. The items are related to the seven "Standards for Teacher Competence in the Educational Assessment of Students." Some of the items are intended to measure general concepts related to testing and assessment, including the use of assessment activities for assigning student grades and communicating the results of assessments to students and parents; other items are related to knowledge of standardized testing, and the remaining items are related to classroom assessment.

<u>Directions</u>:

Read each scenario followed by each item carefully; select the response you think is the best one and *mark your response on the answer sheet*. Even if you are not sure of your choice, *mark the response you believe to be the best*.

Ms. O'Connor, a math teacher, questions how well her 10th grade students are able to apply what they have learned in class to situations encountered in their everyday lives. Although the teacher's manual contains numerous items to test understanding of mathematical concepts, she is not convinced that giving a paper-and-pencil test is the best method for determining what she wants to know.

1. Based on the above scenario, the type of assessment that would *best* answer Ms. O'Connor's question is called a/an
   A. performance assessment.
   B. authentic assessment.
   C. extended response assessment.
   D. standardized test.

2. In order to grade her students' knowledge accurately and consistently, Ms. O'Connor would be well advised to
   A. identify criteria from the unit objectives and create a scoring rubric.
   B. develop a scoring rubric after getting a feel for what students can do.
   C. consider student performance on similar types of assignments.
   D. consult with experienced colleagues about criteria that has been used in the past.

3. To get a general impression of how well her students perform in mathematics in comparison to other 10th graders, Ms. O'Connor administers a standardized math test. This practice is acceptable *only* if
   A. the reliability of the standardized test does not exceed .60.
   B. the standardized test is administered individually to students.
   C. the content of the standardized test is well known to students.
   D. the comparison group is comprised of grade level peers.