

**OREGON PRIMER ON
EDUCATIONAL ASSESSMENT**

This *Primer* is designed to provide Oregon educators with information on the use of formative and summative assessments. The *Primer* includes information on:

- How to find, review, and select assessments.
- A guide to the use and interpretation of assessment instruments.
- A description of issues in analyzing assessment data including research design issues and recommendations.
- A review of several assessment instruments used in Oregon with information on the assessment, its purpose, use, interpretation, and technical properties.

HOW TO FIND, REVIEW, AND SELECT AN ASSESSMENT

Finding Assessment Instruments. There are several sources of information that are very useful in finding and locating assessment instruments. The Buros Institute of Mental Measurements (see <http://www.unl.edu/buros/bimm/index.html>) provides some of the most extensive resources on testing and assessment. The institute provides professional assistance and information to users of commercially published tests. The Buros Institute also encourages improved test development and use through the critical analysis of measurement instruments. The institute publishes a number of resources for test users including the *Mental Measurements Yearbook* and *Tests in Print*. The reviews provided through the Buros Institute are one of the most authoritative sources of information on tests and assessments.

The Buros publication *Tests in Print (TIP)* is an encyclopedia that provides information on published, commercially available tests in psychology and education. The publication is available in most libraries. For each listed test, there is information on intended population, publication date, author, publisher, and references. TIP provides an extensive index to tests that are in print and available.

The *Mental Measurements Yearbook (MMY)* also provides directory type information on available tests but adds details on prices, norming and sampling, scoring, and reporting and

interpretation. *MMY* also provides technical information on tests including evidence of reliability, validity and test bias. For many listed tests, *MMY* also provides one or more independent reviews of the test and testing materials by qualified professionals in the field. To determine if there is a Buross review for a particular test, go to the following web address:

<http://buross.unl.edu/buross/jsp/search.jsp>.

The Educational Testing Service (ETS) is another source of information on tests and assessments. ETS maintains a library of more than 25,000 tests, measurement devices, and research instruments. Collected from the early 1900s to the present, the *ETS Test Collection* is the largest such compilation in the world and is an excellent source of information on tests designed for research purposes and tests that are not available commercially. The *ETS Test Collection* can be accessed electronically at:

<http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnnextoid=ed462d3631df4010VgnVCM10000022f95190RCRD&vgnnextchannel=85af197a484f4010VgnVCM10000022f95190RCRD>

Another source of testing information is The American Psychological Association (APA). The APA website provides information and a FAQ online about how to find and locate Psychological Tests: <http://www.apa.org/science/faq-findtests.html>. The University of Chicago Library also has a useful test collection available at <http://www.lib.uchicago.edu/e/su/tests/>.

Two other useful resources on tests and assessments are provided by the publisher Pro-Ed, Inc. They produce a publication, *Tests*, that like the Buross Institute's *Tests in Print*, provides an encyclopedia-type listing of many instruments in psychology, education, and business. It briefly describes each test, including test title and author, the intended population, test purpose, information on administration, scoring and reporting, and cost. Another publication produced by

Pro-Ed is *Test Critiques*. *Test Critiques* is an adjunct to *Tests* and provides more technical kinds of information on each test including information on reliability, validity, norm development, and test critiques.

Once a test has been located, users may want to contact the test publisher. Directories of test publishers are included in the testing reference books (*MMY*, *Tests*, *TIP*) described above. The Test Collection at Educational Testing Service (ETS) also has a free pamphlet entitled Major U. S. Publishers of Standardized Tests, which lists the names, addresses, and phone numbers of 28 major test publishers. Call or write to them for your free copy at ETS, Library, Rosedale Road, Princeton, NJ, 08541, (609) 734-5667. Most test publishers also maintain extensive web site to provide information to users. Another source of test information that can usually be requested from a publisher's website is a catalog of available tests or brochures describing particular tests. Publisher catalogs are usually produced on an annual or semiannual basis and often have some of the most recent information on current editions of tests and assessment instruments.

Review and Selection of Assessments. We recommend the development of an explicit evaluation process for reviewing and selecting an assessment. There should be an attempt to include all stakeholders in the development of the evaluation process. One of the most important items to include in the evaluation process is the explicit definition of the purpose of the assessment. While some believe that one test can serve many assessment purposes, a test works best when it is used for the purpose for which it was designed and developed. Users should consider using more than one test or assessment when there are distinct purposes for conducting assessments.

In conducting an evaluation of a test, users should also strive to gather evidence that

provides a critical review of the adequacy and quality of the test. Some tests that are commercially available are carefully constructed and serve their purpose well when properly used and interpreted. Other instruments can have much lower quality and technical adequacy. Representations made by a test developer or publisher may not be supported by research evidence or independent reviews. As part of the evaluation process, users should seek out objective information on the quality and adequacy of a test before an adoption decision is made. Two excellent sources for independent review information discussed earlier are Buros reviews and the Pro-Ed publication, *Test Critiques*.

An important consideration in choosing an assessment that usually cannot be determined from published reviews is the degree of alignment of the assessment to local curricular content and standards. While there is a substantial amount of universality in content standards across states, users should carefully examine the match of the assessment's content to the local assessment purpose. This examination should consider the breadth, depth, and match of the content covered as well as whether the level of cognitive activity required by the assessment corresponds to that specified in content benchmarks and standards (e.g., vocabulary versus comprehension; numeracy versus problem-solving),

An important feature of an assessment that has substantial implications for how it can be used and interpreted is the design and methods for reporting scores and assessment results. The more specific and targeted the intended use of the assessment information, the more items are needed on the test and the more important the design of subscale or subscores that can provide specific feedback. For example, many battery-type tests (e.g., ITBS, Stanford, TerraNova, etc.) and many state mandated achievement tests used for NCLB accountability are constructed to provide overall summative indications of performance in reading or mathematics. Only a small

number of items or questions are included for more specific aspects of performance (e.g., multiplying fractions). This means that only overall scores can be reported and used. Users should carefully evaluate whether the design of an assessment's subtests and score reporting meet the purpose and needs for the assessment. Generally speaking, more global score reporting is sufficient for summative purposes but substantial specificity in test construction and score reporting is necessary when the assessment purpose is instructional diagnosis and feedback.

The purpose of assessment should also be considered when examining the number of forms a test has available. When test security is an important consideration as in high stakes, summative situations, it is important to have at least two equated, parallel forms of the test. When a test is to be used for repeated assessments over time as in progress monitoring situations, it is important that the test has many parallel forms so frequent testing can be conducted.

One of the most important areas for review in choosing an assessment is the technical adequacy of the test. Users should carefully examine evidence for the reliability and validity of the assessment. Assessment reliability refers to how consistently an assessment measures across different evaluators, occasions, tasks, or forms of the assessment. Assessment results should not vary substantially across different conditions. If an assessment is not reliable, then the results cannot be trusted in that they will change depending on who gives the test or which day it is given or which test form is administered. Reliability is usually measured using an index that ranges from 0 to 1 where 1 represents perfect consistency. The more important the use of the assessment information, the higher reliability should be. Generally, when assessment information is used to make important decisions, reliability should be .85 or higher.

Validity of an assessment refers to how accurately an assessment measures the specific skill or attribute it is designed to measure and whether the conclusions and inferences drawn from the assessment are accurate. Valid assessment requires not only an assessment that works well but

proper application and interpretation of the test for the intended purpose. The burden of proof for valid test use and interpretation is on the user and the way that assessment information is used. Evidence for valid test use and interpretation can take many forms including demonstrations that test content is relevant and representative, information that test scores are highly correlated with other similar measures, or evidence that test results predict future performance well.

Users should also critically examine evidence that the test developer or publisher has expended effort to obtain independent reviews of the instrument to ensure it is sensitive to all test takers and that it is not biased against protected groups of students. Test bias is the presence of some characteristic of an assessment, a test item, or task in the assessment that results in the differential performance of two individuals of the same ability but from different student subgroups. Test bias can be minimized or prevented through careful test development including clear specification of the construct to be measured and the training of item writers. No matter how careful the test development, however, field-testing and item analysis (e.g., Differential Item Functioning or DIF) must be conducted to evaluate test fairness. Users should check to make sure that test developers have used procedures to ensure test fairness and a lack of bias.

A related issue is test sensitivity. Test developers should also conduct test sensitivity reviews to ensure the appropriateness of test language and design for all test-takers. The goal of sensitivity review is to ensure that the assessment is accessible and respectful of all people and does not unfairly disadvantage or disturb any test-taker. Sensitivity review is intended to eliminate test language or content that is inflammatory, controversial, insulting, slanted, or that unintentionally covers sensitive subjects such as religion, sexuality, or highly political or emotional topics.

The *Standards for Educational and Psychological Testing* (AERA, et al, 1999) provides extensive guidelines for the effective and responsible use of assessments. The *Standards* contain detailed information on best practices in test planning, test design and development, administration, security, and test use and interpretation. Another important component of the *Standards* is its focus on the technical aspects of test development, use, and interpretation. Users should consider a range of criteria in deciding which assessment to use. One of the most important criteria is the match of an instrument to the assessment purpose. Test developers and publishers can sometimes be overly optimistic in describing the breadth of applications of their assessment. However, assessment instruments seldom work well for many purposes. Tests need to be designed and developed in one way for one purpose and in a different way for another assessment purpose.

In summary, there are a host of important considerations for users to evaluate in reviewing and choosing a test for adoption. We encourage users to consult independent reviews and peer reviewed, published research in gathering information. Users may also wish to use consultants who are expert in testing and assessment to provide further support in the decision making process. In choosing an assessment, we encourage an emphasis on critically appraising whether an assessment closely matches the intended purpose and use of the assessment information.

USE AND INTERPRETATION OF ASSESSMENTS

Proper use and interpretation of tests and assessments is a large topic beyond the scope of this *Primer*. We encourage users to consult any of a number of excellent resources on testing and assessment that provide more detailed information on these topics. A selection of such resources is listed at the end of the *Primer*. In the following section, we comment on some general issues that are fundamental in using and interpreting assessment information.

Perhaps one of the most important issues in the use of assessment information is the need

to carefully match inferences and conclusions to the available information provided by the assessment. Users of test information need to be cautious and exact in making sure that the conclusions they reach are truly supported and warranted by the assessment results they are interpreting. Often test interpretations go beyond the available information. Greater certainty in assessment conclusions is bolstered by having multiple test results or sources of evidence as well as making observations of student performance over several occasions. Often, users also ascribe too much confidence to test results when a healthy dose of skepticism is more prudent. Inferences and decisions should be carefully matched to the quality of the assessment information that is available, tempered by the limitations of the assessment methods used, and taking into account other sources of information.

Errors in the use of assessment information are common, both over-interpretations and misinterpretation. Over-interpretations occur when a user makes more of test information than is warranted. Misinterpretation refers to situations in which a user draws an incorrect inference or conclusion from assessment information. For example, a common misinterpretation of grade equivalent scores is that the score indicates the grade level of content that the student can successfully engage. Deeper understanding of how GE scores are computed reveals that the information provided is really just an indication of the relative ranking of a student's performance and is not directly tied to grade-level content. In order to determine if a third grader with a GE score of 5.5 could really manage grade 5 content, one would have to administer a test that contained a representative sample of the grade 5 material. Another misinterpretation that some users make is that a percentile rank on a test indicates how much of the test a student got correct. Percentile rank, however, only indicates the relative ranking of a student's score in comparison to the scores of a group of other students who took the same test. It is entirely

different information than percent correct or percent of material mastered.

Errors in interpretations can be avoided by carefully reading test manuals and interpretative materials supplied by the test maker, by consulting with school or district testing personnel, by ensuring that all staff receives professional development on testing and assessment and by using consultant testing experts to review testing policies and practices.

Almost all assessments require some frame of reference for interpretation. There are few assessments that report results in such a way that results can be interpreted directly. That is, a test score must usually be placed in context or compared to some standard to make the result meaningful. Even a relatively straightforward measure like the total number of words read correctly in an oral reading fluency task (e.g., 58) may not provide much meaningful information without a basis for comparison or interpretation (e.g., How many words is the reading goal for this grade level? What is the average number of words for students in general?).

One of the most common ways to add interpretability to assessment results is through the comparison of a score to other scores obtained by a group of students who took the same assessment. This kind of comparison is usually referred to as a normative comparison. Users should consider the basis for making a normative comparison. What are the characteristics of the normative group? What information is provided by knowing how a student's performance compares to the normative group? Many tests provide normative groups that are created by sampling students to be representative of the national census. If it is important to make a comparison of a student's performance to a nationally representative sample, such comparisons can be useful. Sometimes, however, it might be more meaningful to make comparisons to state or local groups of students taking the same test. Normative comparisons provide relative interpretations; this student's performance is higher or lower than other students. When using

normative interpretations, the user should be sure the comparison group is meaningful and that the desired inference is a relative one of how student performance compares to the performance of others.

Another method for adding meaning to test scores is to compare an individual assessment result to a standard for performance. This is the method currently used in NCLB accountability assessments. A standard is determined and then student performance is evaluated and interpreted in relation to a standard for performance. In many formative assessment approaches a related alternative approach is to set a learning goal or target and then evaluate student performance in relation to attainment of the learning target. In these approaches interpretation of a test score is made in relation to a standard or target for performance rather than in relation to the performance of other students as in normative assessment. Users should carefully consider which kind of interpretation is supported by the assessment they are using.

The research literature on testing and assessment also provides a number of cautions and suggestions for test use and interpretation. One concern is that many tests emphasize superficial learning and recall and teachers and users may not recognize this limitation in the test. Research suggests that many teachers may focus too much on low-level aims, mainly memorization and recall and may overemphasize grading functions. Better assessment practice emphasizes descriptive feedback to the student and advice for learning, and attention on ways to increase personal performance rather than attention on performance in comparison to other students.

Another difficulty that may be present in current assessment practice is the difficulty of connecting assessment results and information with next steps for learning and student learning goals. Teachers' feedback should be focused carefully on student learning needs and strategies to apply assessment information for student improvement. Teachers may face barriers to

effective use of assessment information including lack of time and limited assessment literacy skills. Even if commercially produced assessments are used, teachers may not know how to interpret results, communicate results to stakeholders (i.e., students and parents), provide the kinds of descriptive feedback necessary for student improvement, diagnose needs for particular intervention strategies, or implement those strategies. As a result, there should be systemic efforts to ensure that teachers have adequate support in the use and interpretation of assessment information including adequate opportunities for professional development.

Effective use and interpretation of assessment results also means that teachers need to explicitly design feedback strategies that connect results with instructional decision-making and planning for intervention. It is also important to clearly identify and communicate learning targets to students and communicate assessment results and expectations to students during the learning process. Finally, test users should make sure that analysis and reporting of assessment results are at a level of specificity that allows clear and direct linkage of results to instructional intervention.

Effective use and interpretation of assessment results for learning can also be enhanced through the involvement of students in the assessment process. Student involvement can enhance student engagement with content and can strengthen student motivation and self esteem. To ensure involvement, teachers should design methods to regularly use assessment results to provide detailed descriptive feedback to students and the feedback should be clearly linked to expectations for learning. Teachers should also plan ways to use student self assessment and self monitoring as additional interventions for instructional improvement. Include test-taking skill practice such as reading comprehension and writing tips as part of instruction for students.

Users should also familiarize themselves with professional standards and guidelines on

how to use and interpret tests and assessments. For additional information on test use, users should consult the following publications that have been developed by professionals in several fields:

- *Standards for educational and psychological testing*. (1999). Washington, DC: American Psychological Association. (800) 374-2721.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register*, 43 (166), 38290-38315.
- Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures (4th ed.)*. College Park, MD. (419) 353-0032.
- Joint Committee on Testing Practices (2004). *Code of fair testing practices in education*. National Council on Measurement in Education. (202) 336-6000.

SELECTED ISSUES IN ANALYZING ASSESSMENT DATA

In this last section of the Primer, we briefly discuss some issues that are relevant in manipulating and analyzing assessment data. Often, educators and school administrators collate and use assessment data to inform a variety of decisions about instruction, curriculum, program effectiveness and other matters. We briefly mention several issues here that are important to consider in analyzing assessment data.

The first issue we raise is that of "numbers". Greater numbers lead to more stability in analyzing data and more certainty in the conclusions that can be reached from the data. In assessment contexts we can interpret this in two ways, greater numbers of items and greater numbers of assessment scores. Generally, more items or tasks lead to a better estimate of what is being measured. Conclusions should never be drawn on the basis of one or a small number of items from a test or assessment. Interpretations should be based on at least 5-10 items or tasks, but the user should review evidence for reliability and validity of the particular assessment being used to determine "how much is enough".

Number is also important when interpreting assessment results for groups of students. Little confidence should be placed on differences in scores for groups unless there are a reasonable number of students in the group. How many is enough depends on the importance of the conclusion being drawn or decision being made; the more important the decision the larger the numbers should be. Statistical procedures like the mean and standard deviation start to become very accurate and efficient when the number of people in a group reaches about 25. Users should also be careful to protect student confidentiality when small numbers of students are used in analyses or reports. Even when there are several students in a group, accompanying information like gender or ethnicity may serve to reveal student identities and violate confidentiality principles.

When test data are analyzed and statistics are used to summarize or describe results (e.g., means, standard deviations, correlations, etc.), users should recognize that all estimates are flawed to some extent. Measurement is never perfect. One of the best ways to take this into account and avoid misinterpretations is to use confidence intervals whenever assessment data and summary statistics are being reported and interpreted. Use of confidence intervals provides a direct indication of the precision of estimation embodied in the reported statistics. The wider the confidence interval, the more uncertainty there is in the estimate. When groups have confidence intervals that overlap, performance of the two groups should not be considered different. Many test reports provide confidence intervals directly or report a standard error of measurement that can be used to compute a confidence interval. Users should consult a text on educational statistics or one of the references listed below for further information on how to compute and interpret a confidence interval.

Some other important issues in analysis surround matching the analysis of assessment

data and interpretations to the kind of scores and scales used on the assessment. We describe here three such issues. First, users should know the kind of information contained in the test score they are using. Percentile ranks (PR) and grade equivalent scores, for example, only contain information on the relative ranking of one student versus another student or in comparison to a normative group. Ranking information does not indicate how far apart student performances are just which is higher or lower (e.g., percentile ranks of 50 and 54 do not represent the same difference in performance as 54 and 58). Second, users should be careful in making comparisons of one type of score to another from content to content. Scores may not be comparable unless the test developer has specifically created a linkage or equating between the scores. For example, a score of 80 on reading may not mean the same thing as a score of 80 on mathematics even if standard scores are used. Third, comparisons from one test form or test to another may not be straightforward unless the forms or tests have been placed on the same scale by the test developer. This process, called test equating, is very important to allow valid comparisons. A percentile rank of 72 on one publisher's mathematics test may not mean the same thing as a percentile rank of 72 on a second publisher's mathematics test. Or, scores on the first form of a progress monitoring tool may not be exactly comparable to scores on the other forms of the progress monitoring tool administered later, unless the forms have been equated to be equal in difficulty. Users should carefully study the kinds of procedures and methods used by test developers to produce test forms and report score scales and then temper conclusions and interpretations accordingly.

Users should also carefully examine how well group results (e.g., mean or median) apply to individuals within a group. Averages and aggregate results can mask important differences among individual students. An intervention or program that appears successful on average may

turn out to be effective for some students but not for others. To ensure that errors of interpretation are not made, do not rely exclusively on aggregate statistics and always examine results for individual students as well.

Assessment data are often used for evaluation. In applied educational settings, it is difficult to draw strong conclusions about the impact of a program or intervention on student performance. There are many other factors that may be working simultaneously to effect or influence student performance beside the intervention or program of interest. The most important factor in determining the strength of conclusions that can be drawn about program effectiveness is the adequacy of the research design used to make evaluations. Research design and program evaluation are topics that cannot be covered well in the scope of this *Primer* and users are encouraged to consult one of the references listed below for information beyond the brief recommendations made here.

People often refer to the randomized experiment as the "gold standard" of research design. In such an experiment, participants are randomly assigned to treatment and control groups and researchers or program evaluators have control over almost all aspects of the evaluation environment and the delivery of the program intervention to participants. When carefully designed experiments are conducted successfully, strong conclusions can be drawn about the effectiveness of a program or treatment intervention because other competing explanations or causes for observed outcomes are so carefully controlled and ruled out. However, in educational research that is conducted within complex classroom and school environments, such research designs are rare. We offer here several suggestions for users to consider in designing evaluations that can strengthen conclusions even when experiments are not possible.

First, when comparing an intervention group to a comparison group, choose comparison groups to reflect similar demographic composition. One of the factors that most undermines interpretations of intervention effectiveness is different composition of treatment and comparison groups. By selecting a comparison group that is similarly composed, the evaluator can have greater confidence that observed assessment results are due to the program or intervention and not group differences. For example, in evaluating the effects of a reading program in one school, the evaluator might choose another nonparticipating school with a similar student population rather than making a comparison to district average results. Another strategy to use is, instead of comparing intervention group performance to one "control group", make comparisons to several different groups that have not received the intervention or program. Even though students are not randomly assigned to each group, demonstrating that the intervention group performs differently than several existing comparison groups can be a relatively strong demonstration of an effect. A third suggestion is to use more than one outcome measure to evaluate the effect of intervention. Demonstrations that the intervention effect generalizes to performance or improvements in several arenas adds confidence that the intervention is effective. Another strategy to ensure that intervention effects are not the result of other causes or factors is to deliver the intervention at different times for different participants or groups of participants.

The last strategy we suggest for strengthening conclusions about intervention or program effectiveness is to use longitudinal designs. When participants are tracked over time and their performance is measured repeatedly, much stronger conclusions can be drawn about program and intervention effectiveness. One of the greatest challenges in estimating program or intervention effects when there is no random assignment is separating differences in groups that are the result of the composition or make-up of the groups from differences that are due to the

treatment or intervention. Longitudinal evaluation designs are less susceptible to the influences of student background, intake characteristics, and other confounding factors that make it hard to assess the effects of a program or intervention. In a longitudinal design, students serve as their own controls. As a result, stable characteristics of the child are constant over time and cannot confound estimation of change. Another strength of the longitudinal approach is an inherent focus on fundamental interests of education that change over time like learning and development. When applying longitudinal designs, make sure three or more measurement occasions are used, the assessment uses equated test forms on each occasion, and conditions of administration are made comparable across occasions.

OTHER REFERENCES AND RESOURCES

The ABC's of School Testing (<http://www.apa.org/science/jctpweb.html>)

A videotape developed by the Joint Committee on Testing Practices (JCTP) and a collaboration of several other testing organizations. Designed to help parents understand the many uses of testing in schools today. In addition to the videotape, two publications are also included: *Leader's Guide* and the *Code of Fair Testing Practices*.

AERA Position Statement on High-Stakes Testing in Pre-K – 12 Education:

<http://www.aera.net/policyandprograms/?id=378>

The Assessment Training Institute provides newsletter articles and other publications about classroom and formative assessment as well as videos and training sessions for a fee.

<http://www.assessmentinst.com/>

Black, P. (1998). *Education Assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco: Jossey-Bass.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Research Council.

The Center for Research on Evaluation, Standards, and Student Testing (CRESST) has many useful resources and publications:

CRESST products and resources: <http://www.cse.ucla.edu/products.html>

CRESST newsletters (<http://www.cse.ucla.edu/products/newsletters.asp>) offer full texts of the organization's activities and policy views since Fall 1991

CRESST policy briefs provide guidance to educators and policy makers:

<http://www.cse.ucla.edu/products/policy.html>

CRESST technical reports: <http://www.cse.ucla.edu/products/reports.asp>

Marczyk, G. R., DeMatteo, D., & Festinger, D. (2005). *Essentials of Research Design and Methodology (Essentials of Behavioral Science)*. Wiley.

Marzano, R. J. (2006). *Classroom Assessment & Grading that Work*. Alexandria: VA: Association for Supervision and Curriculum Development.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology*, Wadsworth.

National Council on Measurement in Education (NCME) has a series called ITEMS: The Instructional Topics in Educational Measurement Series. The goal of ITEMS is to improve the understanding of educational measurement principles by providing brief instructional units on

timely topics in the field, modules developed for use by college faculty and students as well as by workshop leaders and participants. <http://www.ncme.org/pubs/items.cfm>

Nitko, A. J., & Brookhart, S. M. (2006). *Educational Assessment of Students* (5th Edition). Prentice Hall.

Northwest Regional Educational Laboratory provides an extensive professional development toolkit on assessment: <http://www.nwrel.org/assessment/toolkit98.php>

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. Lawrence Erlbaum.

Stiggins, R., Arter, J. A., Chappuis, J., & Chappuis, S. (2007). *Classroom Assessment for Student Learning: Doing It Right--Using It Well*. Prentice-Hall.

Thorndike, R. M. (2004). *Measurement and Evaluation in Psychology and Education* (7th Edition). Prentice Hall.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.