Understanding Validity and Reliability in Classroom, School-Wide, or District-Wide Assessments to be used in Teacher/Principal Evaluations

Warren Shillingburg, PhD
January 2016

Introduction

As expectations have risen and requirements for student growth expectations have increased across the country, more and more school districts are being asked to develop local assessments and to use them as part of their teacher evaluation process. As districts explore this process, many are concerned about how they can develop local assessments that provide information to help students learn, provide evidence of a teacher's contribution to student growth, and create reliable and valid assessments. These terms, **validity** and **reliability**, can be very complex and difficult for many educators to understand. If you are in a district that has access to appropriate software or the luxury of hiring a statistician to work through formulas, you are in the minority. Most districts cannot afford these expenses and struggle just to find the resources and time to create any assessment, giving very little time and attention to the concepts of validity and reliability.

In light of this reality, some districts are choosing to simply not go this route and instead follow the premise that everyone contributes in some way to a child's reading and math learning, and using the state or national assessments they currently have in these areas to apply to all teachers. Although there can be some validity in this assumption, with the state expectation that all LEAs shall ensure that multiple measures of student academic progress are used (Arizona State Board of Education, 2016) in a teacher/principal evaluation, districts will need to either find other assessments or develop their own assessments to meet this requirement. Districts will want to address validity and reliability to be sure they are looking at data that will give each teacher a strong source to determine how successful his/her teaching has been in affecting student academic growth. When given the range of choices from creating

assessments and giving no attention to the concepts of validity and reliability to using **statistical software** (e.g., SAS, SPSS) to calculate these **coefficients**, many districts seem to believe their only option is to do the best they can and simply create assessments without concern for these concepts at all. Although this paper will present the proposition that when creating local assessments, you do not have to be as concerned about reliability and validity as a testing company must be when they are creating assessments to generalize to a larger population, it does not mean these concepts should be ignored. The simple practice of giving attention to validity and reliability will help districts create more valid and reliable assessments than they may realize, and districts do not need to be afraid to tackle this challenge for fear of not understanding these concepts.

The goal of this paper is to provide a general understanding for teachers and administrators of the concepts of validity and reliability; thereby, giving them the confidence to develop their own assessments with clarity of these terms. The purpose of this paper is not to provide an in-depth or detailed description of the concepts of validity and reliability; this would require an entire college course. For a teacher, school, or district to create their own assessments, it is not necessary to have such a detailed understanding of validity and reliability. Most of the statistical requirements for assessments have to do with making the assessments generalizable to a larger population, such as is the case with state-wide (AzMERIT) or national (SAT) assessments. When the assessment is going to be used for a classroom teacher to determine content mastery of his/her own students and to try to determine the contribution of the teacher to the child's learning, having a basic understanding of these concepts is sufficient.

The first section of this paper will deal with the concept of validity and its importance to test development. Then the concept of reliability will be covered and how best to create an assessment that gives you an accurate representation of what a child does or does not know in your content area; this will include a brief section on **inter-rater agreement** when dealing with **rubrics** or **performance assessments**. The third section will provide a sample, step-by-step guide on how best to create an assessment to ensure you have the most reliable and valid

measure possible for your classroom, school, or district-wide use. Finally, at the end of this paper you will be provided with a glossary of terms (bold-faced within this document) and a reference list you can refer to if you would like to learn more details about specific concepts shared in this paper.

Validity

> "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests…The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself" (Messick, 1989, p. 9).

The most important factor in test development is to be sure you have created an assessment that allows you to make appropriate **inferences** regarding a child's performance in a content area, and now to feel confident that you can attribute a teacher's contribution to this performance. This requires you to begin with "a clear statement of the proposed interpretations and uses" (Kane, 2006, p.23) of the assessment. "Validity is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose" (AERA, 2008, p.11). Once a thorough attempt has been made to validate the assessment and efforts have failed to determine the assessment is not covering the proposed content, then there can be some degree of trust that the assessment is indeed valid for the intended purpose (Cronback, 1980). In other words, after you have followed the steps to determine your assessment is measuring what was taught, and you are not able to prove differently, you should feel more confident that your conclusions from the assessment are valid.

You will see many sources define validity as the test measuring what it claims to measure, but we are actually talking about the ability to make inferences from the data and that these inferences are indeed correct. If we are testing a child's knowledge of 5th-grade social studies concepts, but the child's reading level is at the 1st-grade level, then the resulting score would be

more a reflection of the child's reading ability than the child's knowledge of social studies; therefore, we would not be able to make a valid inference from the test score and the test would not have validity for this child. It would then follow that we cannot make a valid inference about the teacher's contribution to a child's growth in the social studies content if the child's reading level is **convoluting** the final score.

"The controversy over which type of validation evidence is most fundamental has been debated and discussed in the measurement literature for nearly half a century. The type of validation that is most important depends on the inferences to be drawn from the test scores" (Crocker & Algina, 2008, p. 236). The research literature typically breaks down validity into three basic types: **construct validity**, **criterion validity**, and **content validity**. Construct validity "refers to the skills, attitudes, or characteristics of individuals that are not directly observable but are inferred on the basis of their observable effects on behavior" (Martella, Nelson, and Marchand-Martella, 1999, p. 74). This could be something as abstract as intelligence and creativity with less agreement about what demonstrates the construct to mathematical computation where there would be more agreement about what demonstrates the construct; we can find many differing views about what skills constitute creativity, but most would be in agreement about what skills demonstrate mastery of two-digit addition. Criterion validity is "used when test scores can be related to a criterion. The criterion is some behavior that the test scores are used to predict" (Allen & Yen, 2002, p. 97). Criterion validity can also be called **concurrent validity**, where a relationship is found between two measures at the same time. Since this is seldom used in today's testing environment, we will only focus on criterion validity as it deals with the predictability of the scores. The SAT is a good example of a test with predictive validity when the test scores are highly **correlated** with success in college, a future performance. We could also see this used in a test for job performance to determine which candidates were more likely to be successful in specific job tasks. Although teachers should be familiar with the concepts of construct and criterion validity, you will serve your students well if you focus your attention on content validity and making sure the content you are testing is indeed the content that was taught (Popham, 2014).

To determine both construct validity or criterion validity, calculations and examination of **correlations** or other **statistics** are used. On the other hand, to determine content validity no statistical calculations are used (Allen & Yen, 2002). Content validity is where LEAs should focus all of their attention when dealing with the validity of any measurement tool they create. It is most often associated with achievement testing, and it "refers to the representativeness of the sample of items included in measurement devices" (Martella et al., 1999). In other words, does the test content appropriately represent what was taught? "The key ingredient in securing content-related evidence of validity is human judgment" (Popham, 2000, p. 96). The human judgment in any achievement testing would be the teachers who are teaching the content, and to be sure you have stronger validity evidence, you should include several teachers (and content experts when possible) in evaluating how well the test represents the content taught.

"To claim that a proposed interpretation or use is valid is to claim that the interpretive argument is coherent, that its inferences are reasonable, and that its assumptions are plausible" (Kane, 2006, p.23). There is no better resource to make these inferences than the teachers who are teaching the content and who work daily with the students who will be taking the assessments. Especially, since the concept of validity is violated when we use a test designed to demonstrate student growth or achievement as a measure of a teacher's effectiveness, by having teachers design their own tests for this purpose brings us a little closer to a measurement tool that could be more valid for determining a teacher's contribution to a child's achievement. If the teacher is a part of the process to design the test to demonstrate student understanding of concepts taught, there can be more confidence that the appropriate content is being tested. This then can lead to more valid inferences of a teacher's contribution to student achievement.

Reliability

Reliability is "the desired consistency (or reproducibility) of test scores (Crocker & Algina, 2008, p. 105). We would want to know if the child took the test again, the score would be similar or consistent over multiple testings (Brennan, 2006; McMillan & Schumacher, 1997; Popham,

2014; *Standards for educational and psychological testing, 2008)*. In test design, reliability is referring to the confidence you have that the test score earned is a good representation of a child's actual knowledge of the content, or is a good representation of a child's **true score** if there were no such thing as **measurement error** and we could design a perfect test. Realizing you can never have perfect reliability, due to the fact you can never eliminate all measurement error, and all testing is simply providing an estimate of the child's true score, the goal is to create an assessment that you believe gives you as close to an accurate estimate as possible.

Reliability is more challenging to determine when dealing with national- or state-level assessments where you are trying to generalize to a much larger population. This generalization is not necessary in locally developed assessments for use in a classroom or at the district level. You are not sampling a population, but you are typically including the entire population (your entire class or your entire grade level) in your testing. "Reliability is of more concern on standardized or high stakes testing than they are in classroom assessment. In a classroom, students' knowledge is repeatedly assessed and this allows the teacher to adjust as new insights are acquired" (Moskal & Leydens, 2000, p. 7). Decisions in the classroom, made on the basis of an assessment, can easily be changed if they appear to be wrong. Therefore, reliability is not of the same crucial importance as in large-scale assessments, where there is no turning back (Black, 1998). A well-respected leader in test design, Popham (2014) states, "In general, if you construct your own classroom tests with care, those tests will be sufficiently reliable for the decisions you will base on the test results….you need to be at least knowledgeable about the fundamental meaning of reliability, but I do not suggest you make your own classroom tests pass any sort of reliability muster" (p. 89).

The statistics involved in **reliability coefficients** make most educators uncomfortable. Let's remove this uncomfortable feeling by understanding we do not need the statistics for locally developed assessments. It is important educators understand the concepts involved in reliability, and we will discuss steps to take to create more reliable classroom assessments.

Let's first help you understand what reliability is, and then in the next section, we will help you include steps in your test design that will enhance the reliability of any instrument you create. There are three types of reliability that most state or national assessments may utilize as a means to demonstrate the reliability of their assessment, all of which use statistics to produce a reliability coefficient**: test-retest** (stability), **alternate-form**, and **internal consistency**. If you would like to understand or learn the actual statistical calculations to determine these reliability coefficients, you can read one of the sources listed in the reference list. Basically, the statistics provide a number between 0 and +1.00 that indicates the degree to which the scores are related, with a score near +1.00 showing a strong relationship or high reliability, and a score close to 0 indicating a poor relationship or low reliability. For the purpose of designing your own assessments for internal use, it is not necessary to understand nor to create these reliability coefficients (Popham, 2014), but you should be versed enough in these three types of reliability to help other staff or parents in your district understand their meaning.

First, the **coefficient of stability** (test-retest) is a measurement of how the scores from one test remain fairly similar when a group of students takes the same test twice, not just for the individual student but also in how they perform in relationship to other scores. In the more reliable measures, the student's individual scores remain fairly consistent, as well as the rank order of how well students perform on one test will remain close to the same on the second measure. This is where having a wider spread of scores (a larger **standard deviation**) can help ensure the reliability because the order of performance is less likely to change when the scores are spread farther apart, or when you have a more **heterogeneous group** of students where scores are not clustered in the top or bottom of the range. Second, the **coefficient of equivalence** (alternate form) is a measurement of how the scores on one test remain similar to the scores on a second, or parallel, test given around the same time. The coefficient of equivalence is also dependent on the rank order of how well students perform, having more reliability when the rank order of scores on one measure remains close to the same rank order on the second measure. Finally, the **coefficient of internal consistency** is a measure of how well the items are working together to measure the same concept. This is often computed by

what is called "**split-half correlation**" where there are multiple ways for a test to be divided into two separate scores thus yielding the coefficient of equivalence (Martella et al., 1999; McMillan & Schumacher, 1997).

You should also understand the concept of **error of measurement**. Typically, measurement error is reported as a plus/minus number that is then added to your observed score to give the range of possible values of the student's true score. For example, if a student scores a 75% with a measurement error of +/-3, this tells you that 68% (one standard deviation from the mean in **a normal distribution**) of the time the child's true score would range between 72% and 78%, and 95% (2 standard deviations from the mean in a normal distribution) of the time the child's true score would range between 69% and 81%. As has been shared with reliability coefficients, it is not necessary to calculate the measurement error when creating your own classroom assessments. You simply need to be aware of ways to reduce this error as you design your assessments.

A teacher-developed test can be made more reliable and have less measurement error by making sure you have written clear directions so students know exactly what is expected of them from the assessment; that you have appropriate questions that clearly measure the content taught and are not confusing students with the wording or biased to any subgroup of students; and that you have gathered appropriate feedback from colleagues and students who have read through your assessment or taken the assessment to provide feedback to its clarity (Brennan, 2006; Nhouyvanisvong, 2015; Popham, 2014).

Interrater Agreement

Another area of focus when working to enhance reliability is when you are working with performance assessments or rubrics. With rubrics or performance assessments, you will typically be concerned with the internal consistency expressed as a coefficient, most often called **interrater reliability**. This form of reliability is to make sure observers are measuring the same **variables** and are consistent in how they rate each of the variables (McMillan &

Schumacher, 1997; Stemler, 2004). It is not common, however, for local districts to determine the reliability coefficient unless you have access to available software to perform this function; the typical measurement is to use **interrater agreement**, which is expressed as a **percentage of exact agreement** among observers, and the goal is to reach at least the 70% agreement threshold (Jonsson & Svingby, 2007). This is a simple calculation determined by counting the number of exact agreements on rating variables on any part or whole of an observation tool and then dividing this by the number of total observation variables (Among 10 observation ratings, observer 1 and observer 2 agreed on 8 of them, this would then give you 80% agreement among the two observers).

To illustrate the difference between interrater reliability and interrater agreement, we can look at two sets of ratings of four different students on an overall rating variable with a 1-5 rating scale:

	Rater 1	**Rater 2**
Student A	1	2
Student B	2	3
Student C	3	4
Student D	4	5
AGREEMENT	0.0 (low)	
RELIABILITY	1.0 (high)	

(Adapted from Graham, Milanowski, and Miller, 2012)

Since the interrater reliability coefficient is calculated on consistency of ratings and on the relative standing of performance, regardless of absolute value of each rating, we would see high interrater reliability between these two raters, because the rank order they placed each variable is exactly the same. We would, however, have zero interrater agreement, since each of the variables was scored differently between the two raters (Tinsley & Weiss, 2000). For local school-assessment purposes, we are more concerned with the raters agreeing with the level of performance than we are with simply having them score consistently. This would be the same idea behind providing models of each performance level (e.g., excellent, above average, average, poor), working to get each observer to agree with what is considered the appropriate rating for each level.

To take this one step further, you can also calculate the **percent of adjacent agreement** where you would count the total of exact agreements and include the total of agreements that were only one performance level of one another. This calculation should be 90% or higher. When dealing with four or fewer performance levels, you should focus on percentage of exact agreement. If you are using a rubric or performance scale that has more than five levels, then calculating the percent of adjacent agreement may be a more realistic measure to use (Graham, Milanowski, & Miller, 2012).

Finally, not using statistics to determine a reliability coefficient does not allow you to account for **chance** or **random error** (Watkins & Pacheco, 2000), but most would agree for district-level purposes, looking at interrater agreement will give you what you need. Again, we are talking about what is practical and reasonable to expect in a school district, and what will give you confidence with the process and products you develop for your local use.

Steps in Designing a Local Assessment

Progressing through these steps to create local assessments that are more reliable and valid can give you confidence the scores are truly a measure of how well each student understands the tested content, and that the teacher did indeed teach this content and therefore contributed to the child's assessed growth. Following the steps, a brief explanation of each will be provided, but you are encouraged to read several of the sources to gain a deeper understanding of how to ensure each step is completed appropriately (Alias, 2005; Brennan, 2006; Crocker & Algina, 2008; Nhouyvanisvong, 2015; Popham, 2014). The key is to remember that test design is a fluid process, and you will want to revise and update your assessment after each administration of a test, constantly working toward a more valid and reliable assessment:

1. Identify the test purpose or test objective.
2. Construct a Table of Specifications/determine item format.
3. Construct initial pool of items.
4. Review test items with colleagues and students; then revise as necessary.
5. Pilot test your assessment; then revise as necessary.
6. Administer and score your assessment; review and revise as necessary.

1. Identify the Test Purpose or Test Objectives

At this stage of test development, it is important to be specific when identifying the test purpose or **test objectives** to enhance your test validity. It is from these objectives that the rest of the test will be developed. If your test objective is to measure students' general knowledge of algebra I or students' understanding of basic dance movements, then you need to be certain that each question you develop is measuring these skills, and especially that these are the skills that were taught. Validity is stronger when we are confident we are making appropriate inferences by measuring what we are claiming to measure. We do not want to teach students only pre-algebra skills and then test them on algebra skills. That would not be a valid measure and would not give us any data to show student understanding or the effectiveness of instruction. You should be as clear and specific as possible when stating the test objectives and then make sure each question is measuring these stated objectives.

2. Construct a Table of Specifications/Determine Item Format

This is an important step in the test design process, because it ensures your test is covering the course objectives and at the appropriate level. It will also help guide you on the type of question format to use that will best demonstrate your objectives. The **table of specifications** can be very basic, simply listing the skills to assess in one column, the number of questions to test this skill in the second column, the level of difficulty you intend to use for each question in the third and fourth columns (refer to chart on page 12), and the number of total points for each skill in the fifth column. By determining the number of questions to cover each skill and the point value assigned to each skill, you are deciding on the relative importance of each skill. This can also assist with validity by making sure you are adequately covering the tested content. Then by determining the level of difficulty of each skill, you will be deciding on the type of question format appropriate to test each skill.

Also, reliability can be enhanced by having a longer assessment, meaning you have several questions assessing each skill to help avoid the problem with guessing or careless errors in

determining how well a child has mastered a given skill. You don't want the test to be too long, because this can have an effect on validity as students will tend to lose focus. There are no solid guidelines on the appropriate length of your assessments; this you will need to determine through trial and error with your actual population of students.

Typically, when determining the level of difficulty of a question, you are using **Bloom's Taxonomy** or **Webb's Depth of Knowledge**. For our purposes, I will use Bloom as an example. You can also be as detailed as you want with breaking the levels down among Bloom's different levels, but my personal recommendation is to simply decide if you are looking at lower-level skills (factual recall or comprehension) or higher-level skills (application or above). I have taught many classes over the years where we could spend hours debating which level of Bloom was being used. In the end, it is not as important to have the exact level as it is to understand whether you are looking for lower-level or higher-level skills. Below is one example of what your Table of Specifications could look like.

Skill	**Number of Questions**	**Lower-Level Diff.**	**Higher-Level Diff.**	**Total Points**
Skill 1	2	1	1	15
Skill 2	3	1	2	25
Skill 3	1		1	10
Skill 4	4		4	40
Skill 5	2	2		10
Totals	12	4	8	100

Once you have completed your Table of Specifications, you can review it to be sure you have a balance between the levels of difficulty, the appropriate number of questions for the test overall, the appropriate number of questions for each skill you wish to assess, and the appropriate number of points assigned to your test overall. You should make any adjustments now if you have concerns before moving to the next step of designing your test items.

3. Construct Initial Pool of Test Items

From your Table of Specifications, you will now determine the type of questions you will write in order to gain the appropriate understanding regarding student learning. Typically, if you are looking at low-level skills, you will use **selected-response items** (multiple choice, true/false, matching), and if you are looking at high-level skills, you will use **constructed-response items** (short answer, essays, performance assessments with rubrics).

This is where you should look at additional resources to help you design quality questions and where you can help with the reliability of your scores and reduce the measurement error by making sure the question is clearly stated and the response expected is understood. When creating selected-response items, it is the question itself that needs to be clear to enhance reliability. You want to be sure the student understands the question and that it is clear on what it is asking the student to make sure you get reliable responses. This is typically done through piloting and determining that the feedback you are getting is indeed what you intended the question to ask. When creating constructed-response items, it is the scoring that needs to be clear to enhance reliability and to reduce measurement error. By making it clear what is expected in the response and getting that agreement among the observers, you will have more reliable results. This is often done through rubrics and training of observers. It is also enhanced by providing examples of each level of performance that the observers can use as models.

You may also, at this stage, have access to a question bank. If that is the case, at this time you would choose the appropriate questions to match the skills and the difficultly level you are assessing.

4. Review Test Items with Colleagues and Students; Then Revise as Necessary

This review process will help you with validity to make sure your colleagues agree that the questions are indeed testing the content taught, and it can help you with reliability by making

sure the questions solicit the intended responses. You can also reduce measurement error by getting agreement on what the correct answer should be and how many points each skill or subskill is worth. If anything is not clear to the students or your colleagues, this is your chance to revise and improve the questions before administering the final assessment.

Even if you use a test bank, it is still recommended you review the items with colleagues and students. As is often the case, there are good question banks and bad question banks, and the only way to make sure you have good questions is to do a thorough review of each of them.

5. Pilot Test Your Assessment; Then Revise as Necessary

Now you actually give your assessment to a group of students who are similar to the students who will be taking the final assessment. You will be looking for feedback on the clarity of the questions and evaluating the responses to make sure your questions are giving you the responses you need. This will be your last time to revise the questions before they become final, and your last chance to feel confident you will have reliable responses that can be scored with limited error of measurement. Remember, you can never eliminate error entirely, only put practices into place to reduce it as much as possible.

If you are using a test bank and have access to statistical software that allows you to get item **difficulty indices** and **item discrimination indices**, you would look at these data at this time. For our purposes, we have not gone into an explanation of these statistical measures, since there is a cost involved in having access to this type of software. If you have this access, you should read the appropriate sources to understand exactly what these statistics provide and how best to use them to improve your overall test design.

6. Administer and Score Your Assessment; Review and Revise as Necessary

Now you are ready to give your assessment to the intended students. Once you have scored the assessment, you will take one more look at the responses to determine if you are getting reliable data from each question. As the teacher, you will also look at the overall results to determine if the results match other data you have on the performance of your students. This is another reliability check. You can also give students a questionnaire to solicit additional feedback on the questions. You then use this information to revise again your assessment. Then you should be able to give this test each year to future students, continuing to revise as necessary. The longer you give the assessment, the more confidence you will have that the questions are giving reliable responses with limited measurement error, and you are able to make valid inferences from the data as to each student's mastery of the content and to the contribution of the teacher toward this mastery.

Glossary of Terms

Bloom's Taxonomy – The original taxonomy was developed in 1956 with six levels: knowledge, comprehension (low-level skills), application, analysis, synthesis, and evaluation (higher-level skills) to help academics avoid duplicate or redundant efforts in developing different tests to measure the same educational objectives. The goal was to help researchers and educators understand the fundamental ways in which people acquire and develop new knowledge, skills, and understanding. In 2001, the taxonomy was revised using verbs instead of nouns: remembering, understanding (low-level skills), applying, analyzing, evaluating, and creating (higher-order skills).

Chance – The likelihood that something has happened unpredictably without human intervention or observable cause.

Coefficients – An expression of the change or effect produced by the variation in certain variables, or of the ratio between two different quantities, typically expressed in a range of value from +1.00 (perfect positive relationship) to 0.00 (no relationship) to -1.00 (perfect negative relationship).

Concurrent Validity – A type of criterion-related validity that determines the extent to which the measurement device may be used to estimate an individual's present standing on a criterion variable.

Constructed-Response Items – Items on an assessment that require a response to be generated by the student (e.g., short answer, essays, performance).

Construct Validity – The extent to which a measurement device can be shown to measure a hypothetical construct. Examples of constructs are intelligence, creativity, reading comprehension capabilities, or mathematical competence.

Content validity – The degree to which a measurement actually reflects the variable it has been designed to measure.

Convoluting – To make complex and difficult to follow.

Correlated – Having a mutual relationship or connection, in which one thing affects or depends on the other.

Criterion Validity – The extent to which an individual's score on a measurement device is used to predict his or her score on another measurement device.

Difficulty Indices – The proportion or percentage of students who answered the item correctly. Item difficulty can range from 0.0 (none of the students answered the item correctly) to 1.0 (all of the students answered the item correctly).

Heterogeneous Group – A group consisting of students with a wide variety (high to low) of instructional levels and skills.

Inferences - A conclusion reached on the basis of evidence and reasoning.

Interrater Agreement – The degree to which two or more evaluators using the same rating scale give the same rating to an identical observable situation.

Interrater Reliability – The consistency between evaluators in the ordering or relative standing of performance ratings.

Item-Discrimination Indices – The score given to determine how well a test item serves to discriminate between students with higher and lower levels of knowledge on the test. +1.0 would be perfect item discrimination among all high scorers and all low scorers, meaning all high scorers answered the question correctly and all low scorers answered the question incorrectly; 0 would show no discrimination, meaning there was a mix of high and low scorers who answered the question correctly; and -1.0 would show perfect nondiscrimination where no high scorers answered the question correctly and all low scorers answered the question correctly.

Measurement Error (Error of Measurement or Random Error) – Fluctuations in scores because the measurement device does not measure an attribute the same way every time; unknown and unrepeatable causes of variability in task performance over time and context.

Normal Distribution – An arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either end. In education, often referred to as the bell curve.

Percentage of Adjacent Agreement – Calculates the number of times ratings fall within one performance level of one another (including exact agreements), then divides this number by the total number of ratings.

Percentage of Exact Agreement – Calculates the number of times raters agree on a rating, then divides this number by the total number of ratings.

Performance Assessment – A form of testing in which a student is given a task, typically a demanding one, then asked to respond to the task orally, in writing, or by constructing a product.

Reliability – The consistency of results over time.

Alternative Form (Coefficient of Equivalence) – A measure of the magnitude of the relationship between participants' scores on two comparable forms of the measurement device.
Internal Consistency (Coefficient of) – A measure indicating the magnitude of relationship between participants' scores on a single administration of the measurement device usually assessed by comparing two parts of a test (e.g., odd and even items, first half to second half).
Test-Retest (Coefficient of Stability) – A measure indicating the magnitude of relationship determined by administering the measurement device to a sample of individuals and then re-administering the device to the same sample of individuals after some time delay.

Reliability Coefficient – An indicator that reflects the degree to which scores are free of measurement error. The values typically range from 0 to 1.0. A coefficient of 0 means no reliability and 1.0 means perfect reliability. Since all test have some error, reliability coefficients never reach 1.0.

Rubric – A scoring guide employed to evaluate the quality of a student's responses to a performance test, a student's portfolio, or any kind of student-generated response.

Selected-Response Items – Test items that present options from which the student must choose (e.g., multiple-choice, matching).

Split-Half Correlation – A technique of splitting a body of supposedly homogeneous data into two halves and calculating the results separately for each to assess their reliability.

Standard Deviation – A statistic that describes how much the scores are spread out (distributed) around the mean; the larger the standard deviation, the more spread out the scores.

Statistics – The study of the collection, analysis, interpretation, presentation, and organization of data.

Statistical Software – Computer software that will assist with the analysis and interpretation of data.

Table of Specifications – A chart that describes the skills to be covered on a test and the number of items and points that will be given to each skill. The chart can be further broken down into levels of difficulty for each question; therefore, guiding the type of question to be written for each skill.

Test Objectives – Defines the content to be assessed on the test. This will be further broken down into skills, and these skills are where the actual test questions are developed.

True Score – The average of the scores that would be earned by an individual on an unlimited number of perfectly parallel forms of the same test; the error-free value of test-taker proficiency.

Validity – The extent to which a test measures what it claims to measure and the conclusions reached are appropriate based upon the data.

Variables – A factor or condition that is subject to change, especially one that is allowed to change in a scientific experiment to test a hypothesis.

Webb's Depth of Knowledge – Developed in 1997 as a process and criteria for systematically analyzing the alignment between standards and standardized assessments. The model is based upon the assumption that curricular elements may all be categorized based upon the cognitive demands required to produce an acceptable response. Each grouping of tasks reflects a different level of cognitive expectation, or depth of knowledge, required to complete the task. The depth of knowledge (DOK) level describes the kind of thinking required of a task, not whether or not the task is difficult. The DOK levels are 1) recall and reproduction, 2) skills and concepts, 3) short-term strategic thinking, and 4) extended thinking.

References

Alias, M. (2005). Assessment of learning outcomes: validity and reliability of classroom tests. *World transactions on engineering and technology education*. 4 (2). Retrieved January 5, 2016 from http://www.wiete.com.au/journals/WTE&TE/Pages/Vol.4,%20No.2%20(2005)/16-Alias32.pdf.

Allen, M.J. & Yen, W.M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (2008). *Standards for educational and psychological testing*. Washington, DC.

Arizona State Board of Education. (2016). *Arizona framework for measuring educator effectiveness*. Approved by the Arizona State Board of Education in January 2016.

Black, P. (1998). *Testing: Friend or foe?* London: Falmer Press.

Brennan, R.L. (Ed.). (2006). *Educational measurement*. (4th Ed.). Westport, CT: Praeger Publishers.

College Board. Validity Evidence. *Validity handbook*. Retrieved December 24, 2015 from http://research.collegeboard.org/services/aces/validity/handbook/evidence

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.

Cronback, L.J. (1980). Validity on parole: How can we go straight? *New directions for testing and measurement: Measuring achievement over a decade*, 5, 99-108.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. (DOE Publication No. ED-06-CO-0110). Center for Educator Compensation Reform. U.S. Department of Education, Office of Elementary and Secondary Education, Washington, D.C.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational research review.*2 (2), 130-144.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational Measurement*. (4th Ed.). Westport, CT: Praeger Publishers.

Martella, R.C., Nelson, R., & Marchand-Martella, N.E. (1999). *Research methods: Learning to become a critical research consumer*. Needham Heights, MA: Allyn & Bacon.

McMillan, J.H. & Schumacher, S. (1997). *Research in education. A conceptual introduction.* (4th Ed.). New York: Addison-Wesley Educational Publishers, Inc.

Messick, S. (1989). Vailidty. In R.L. Linn (Ed.). *A nation at risk: The imperative for education reform.* Washington, DC: U.S. Government Printing Office.

Moskal, Barbara M. & Leydens, J.A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10). Retrieved December 24, 2015 from http://PAREonline.net/getvn.asp?v=7&n=10

Nhouyvanisvong, A. (2015). *Improving teacher-developed assessments and items*. Retrieved January 5, 2016 from http://www.naiku.net/wp-content/uploads/Improving-Teacher-Developed-Assessments.pdf.

Popham, W.J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. (3rd Ed.). Boston: Allyn and Bacon.

Popham, W.J. (2014). *Classroom assessment: What teachers need to know*. (7th Ed.). Boston: Pearson Education, Inc.

Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved January 4, 2016 from http://PAREonline.net/getvn.asp?v=9&n=4.

Tinsley, H.E.A., & Weiss, D.J. (2000). Interrater reliability and agreement. In H.E.A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling*. 95-124. New York: Academic Press.

Watkins, M.W. & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of behavioral education*. 10(4), 205-212.