Accuracy vs. Validity, Consistency vs. Reliability, and Fairness vs. Absence of Bias: A Call for Quality

Paper Presented at the Annual Meeting of the American Association of
Colleges of Teacher Education (AACTE)
New Orleans, LA.
February 2008

W. Steve Lang, Ph.D.
Professor, Educational Measurement and Research
University of South Florida St. Petersburg
Email: ws.lang@knology.net

Judy R. Wilkerson, Ph.D.
Associate Professor, Research and Assessment
Florida Gulf Coast University
email: jwilkers@fgcu.edu

Abstract

The National Council for Accreditation of Teacher Education (NCATE, 2002) requires teacher education units to develop assessment systems and evaluate both the success of candidates and unit operations. Because of a stated, but misguided, fear of statistics, NCATE fails to use accepted terminology to assure the quality of institutional evaluative decisions with regard to the relevant standard (#2). Instead of "validity" and "reliability," NCATE substitutes "accuracy" and "consistency." NCATE uses the accepted terms of "fairness" and "avoidance of bias" but confuses them with each other and with validity and reliability. It is not surprising, therefore, that this Standard is the most problematic standard in accreditation decisions. This paper seeks to clarify the terms, using scholarly work and measurement standards as a basis for differentiating and explaining the terms. The paper also provides examples to demonstrate how units can seek evidence of validity, reliability, and fairness with either statistical or non-statistical methodologies, disproving the NCATE assertion that statistical methods provide the only sources of evidence. The lack of adherence to professional assessment standards and the knowledge base of the educational profession in both the rubric and web-based resource materials for this standard are discussed. From a policy perspective, such lack of clarity, incorrect use of terminology, and general misunderstanding of high quality assessment must lead to confused decision making at both the institutional and agency levels. This paper advocates for a return to the use of scholarship and standards in revising accreditation policy to end the confusion.

Introduction

The rubric for NCATE Standard 2, Assessment System and Unit Evaluation, is fuzzy. This fuzzy rubric leads to fuzzy decisions about the Standard at both the accreditation agency and university levels. Deans and directors of education are left in a

state of confusion about how to operate within this fuzzy context. So are members of the Board of Examiners who recommend the accreditation decisions.

This paper is written with the minimum goal of bringing clarity to this aspect of the NCATE accreditation process – the assessment of assessment systems. Increased clarity can help college of education deans and directors make policy level decisions about assessment system design and technology support based on a scholarly understanding of the issues at hand. A more ambitious goal is to serve as a call for change, assisting NCATE, and its constituent organizations, in fixing the problem at its root – the rubric for the standard itself.

The fundamental problem addressed in this paper is that NCATE has chosen to use terminology in the Standard 2 rubric that is non-standard – or not commonly accepted. In an effort to make a more user-friendly process (what these authors assert to be an inappropriate policy decision for an agency entrusted with facilitating public accountability), the agency uses the terms "accuracy" and "consistency" as substitutes for "validity" and reliability." It also provides non-standard and confounded definitions for the professionally accepted terms "fairness" and "avoidance of bias." Here an attempt is made to untangle the language, showing its sources, and where things went off-track.

is that institutions are left groping for solutions to solve ill-defined targets (i.e., accuracy vs. validity) with their fear of measurement professionals reinforced and endorsed by NCATE explanations. Software producers that have become proficient in calculating descriptive statistics are now moving incorrectly into the world of inferential statistics, filling the void for statistical help with easy but badly applied math.

Sources of Information

NCATE

Several documents from the NCATE web site have been reviewed and are used. The main NCATE reference is, of course, the current version of the NCATE Standards, Professional Standards for the Accreditation of Schools, Colleges, and Departments of Education. Standard 2, Assessment System and Unit Evaluation, which reads as follows:

> The unit has an assessment system that collects and analyzes data on applicant qualifications, candidate and graduate performance, and unit operations to evaluate and improve the unit and its programs.

The Standard has three elements: (1) Assessment System; (2) Data Collection, Analysis, and Evaluation; and (3) Use of Data for Program Improvement. The focus here is on the first element, and the rubric for that element is presented in Figure 3, with the main words of interest for this paper in boldfaced font.

Unacceptable	Acceptable	Target
The unit has not involved its professional community in the	The unit has developed an assessment system with its	The unit, with the involvement of its professional community, is

development of an assessment system. The unit's system does not include a comprehensive and integrated set of evaluation measures to provide information for use in monitoring candidate performance and managing and improving operations and programs. The assessment system does not reflect professional, state, and institutional standards. Decisions about continuation in and completion of programs are not based on multiple assessments. The assessments used are not related to candidate success. The unit has not taken effect steps to examine or eliminate sources of bias in its performance assessments, or has made no effort to establish fairness, accuracy, and consistency of its assessment procedures.	professional community that reflects the conceptual framework(s) and professional and state standards. The unit's system includes a comprehensive and integrated set of evaluation measures that are used to monitor candidate performance and manage and improve operations and programs. Decisions about candidate performance are based on multiple assessments made at admission into programs, at appropriate transition points, and at program completion. Assessments used to determine admission, continuation in, and completion of programs are **predictors** of candidate success. The unit takes effective steps to **eliminate sources of bias** in performance assessments and works to establish the **fairness, accuracy, and consistency** of its assessment procedures.	implementing an assessment system that reflects the conceptual framework(s) and incorporates candidate proficiencies outlined in professional and state standards. The unit continuously examines the **validity and utility** of the data produced through assessments and makes modifications to keep abreast of changes in assessment technology and in professional standards. Decisions about candidate performance are based on multiple assessments made at multiple points before program completion. Data show the strong relationship of performance assessments to candidate success. The unit conducts thorough studies to establish **fairness, accuracy, and consistency** of its performance assessment procedures. It also makes changes in its practices consistent with the results of these studies.

Figure 3: NCATE Rubric for Element 1 of Standard 2 – Assessment System

Other NCATE sources are as follows:

Assessing the Assessments: Fairness, Accuracy, Consistency, and the Avoidance of Bias in NCATE Standard 2. The authors are not identified and will be referenced merely as "NCATE." This is the primary source document used for this paper, since it begins with the following statement: "Fairness, accuracy, consistency, and the elimination of bias are important concepts in the first element of NCATE Unit Standard 2, Assessment and Unit Operations" (p. 1). In order to cite the entire resource paper within this paper, to avoid confusion, we have shadowed the text so that it stands out from the interpretations provided herein. The paper provides the rubric language and then states its purpose as follows:

> This paper is written to 1.) define the concepts of fairness, accuracy, consistency, and the elimination of bias; and 2.) provide examples of how institutions can ensure that their assessments adequately reflect these concepts. (p. 1)

Specifications for a Performance-Based Assessment System for Teacher Education (Stiggins, 2000). This document appears to be the primary source for the *Assessing the Assessments* paper, although it is not directly quoted other than as a footnote.

> Other NCATE resources, although not directly cited herein, cite the Stiggins' resource, indicating its continuing influence on NCATE policy and procedures. These include:
>
> *Aligning Assessments with NCATE Standards* (Elliott, 2001)
>
> *Student Learning' in NCATE Accreditation* (Elliott, 2005)
>
> *Criteria for Evaluating Performance Assessments* (Beggar and Zornes, 2006)

Professional Measurement Standards

The standards used by the measurement profession are contained in the *Standards of Educational and Psychological Measurement*, published in 1999 by a joint committee of three influential and important organizations: the American Association of Educational Research (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). These *Standards* are used by all professionals who construct tests or other measurement instruments (e.g., performance assessments and portfolios), especially in a high stakes context such as teacher certification. When high stakes decisions are challenged in the courts, these *Standards* serve as the legal standard to which the assessment process and results are held. Because of these two critically important applications, the *Standards* outweigh in importance and credibility all textbooks and scholarly literature in defining the requirements of high quality assessment. They are the standards that set the bar.

Even though the word "accuracy" appears often in the literature, particularly textbooks, it is not a viable substitute for "validity." James Popham, for example, who has written many popular assessment textbooks and is one of the most respected and prolific authors in the measurement field, uses the term "accuracy" for explanatory purposes but them makes the following statement about the necessity to use the real language (i.e., "validity") of measurement in the world of practice:

> Well, as any student of language will tell you, the meanings we attach to words are basically conventions….In a technical field such as education where precision of communication is imperative, it makes sense to rely on widely used conventions. Otherwise, educators will be employing all sorts of aberrant expressions to describe technical phenomena, with the result that confusion, rather than clarity ensues. As a consequence, it will be strategically sensible for us to rely on the terminology conventions most recently sanctified by APA, AERA, and NCME, and this means the terminology endorsed in the *Standards.*" (p. 94)

Scholarly Literature

Scholarly literature adds to our practical understanding of the field in many important ways, so this paper also cites some mainstream references from peer reviewed

journals, textbooks, and encyclopedias in the assessment, evaluation, and measurement field. These references will be used to help clarify the terms and to provide some practical solutions and examples of what BOE members can expect and teacher education units can produce to provide evidence of validity, reliability, and fairness.

NCATE, of course, has clear expectations of colleges and their faculty with regard to scholarship; yet, the argument here is that the agency does not apply its own standards to its own practice, and the references used in their work stand in sharp contrast to that which is proposed herein. For the two primary NCATE resources used in this paper, the first, *Assessing the Assessments* (author unknown) has no list of references at the end. It does footnote the Stiggins' (2000) commissioned paper (mentioned above) as its only source. The Stiggins' paper includes just six references; three are by Stiggins himself, and one is the NCATE *Standards*. The remaining two are from the literature but are somewhat far removed from the requirements of a scholarly version of a conceptual framework. This is particularly disconcerting since this framework provides the specifications for systems focused on measuring the quality of teacher preparation programs that have the well-being of children at their core. (Cited are: Inside the Black Box in *Phi Delta Kappan* and Administrator Certification Requirements in *Applied Measurement in Education.)* These two papers are not representative of the type of classic pieces one would expect to see used in defining a set of national requirements for assessment. This lack of scholarship would not be tolerated in the BOE evaluation of a conceptual framework, which must be "knowledge-based," or the onsite evaluation of faculty scholarship as part of the review for NCATE Standard 5.

It is common practice in the scholarly world to use and cite the literature in the implementation of practice. Failure to use the literature to provide constitutive and operational definitions is "unacceptable." Definitions, then, are the next focus of this paper, beginning with the definition of the term "accuracy," since that is the initial source of much of the confusion. Definitions of reliability, fairness, and avoidance of bias will follow.

Accuracy (vs. Validity)

<u>The NCATE Interpretation from the *Assessing the Assessments* Resource Paper</u>

In *Assessing the Assessments*, NCATE begins its discussion of the term "accuracy" with the following statement:

> Assessments are accurate when they measure what they purport to measure. To this end the assessments should be aligned with the standards and/or learning proficiencies that they are designed to measure. (p. 1)

This first statement is a classic textbook definition of validity that has a long historical use (Cronbach, 1949; Cureton, 1951). Alignment to standards is a practical application of the development of test blueprints, a typical strategy to develop a validity

argument traditionally called content validity, an application used since the earliest modern discussions. The NCATE authors continue with a discussion of three characteristics to determine alignment, taken largely from the resource paper by Stiggins (2000), and providing additional linkages to the form of validity evidence called "content validity." These characteristics and subsequent discussion are:

- the same or consistent categories of content appear in the assessments that are in the standards;
- the assessments are congruent with the complexity, cognitive demands, and skill requirements described in the standards; and
- the level of effort required, or the degree of difficulty is consistent with standards and reasonable for candidates who are ready to teach or take on other professional responsibilities.

> Most institutions already employ several activities that ensure accuracy of assessments. One activity is simply reviewing assessments to ensure alignment and appropriateness, and documenting the review. This can happen once a year at a staff meeting, or in programmatic committees, or could be done by one person and discussed with a larger group. Accuracy can also be supported by documenting the relationship between assessment results and candidate performance on related assessments, grades, and program completion.

Note that in the first characteristic (a), even the word "content" is used. In fact, both the first and second characteristics are used typically as evidence of content validity. The alignment function is a typical content validity study. The final suggestion of documenting the relationship between assessment results and candidate performance on related assessments, grades, and program completion is most frequently accomplished through correlational studies (predictive or concurrent validity studies) or sometimes regression or discriminant function analysis. In fact, the term "predictor" is used in the actual rubric for the "acceptable" level. Nonetheless, near the end of the discussion (and preceding this last paragraph on "accuracy," NCATE makes the following incorrect statement about validity, attempting to avoid the statistics it has just suggested (italics added for emphasis):

> Accuracy is closely related to the statistical term "validity." *However, establishing validity requires statistical analysis* beyond what is called for in the NCATE standards.

As demonstrated to this point, "validity" is not simply a statistical term, although statistical methodologies are useful in some forms of validity evidence, and even recommended in this NCATE resource piece. The literature on validity provides many opportunities for judgmental studies, most notably content validity –one that is clearly being strongly recommended by NCATE. There is no logical reason to avoid the term, given the above discussion.

<u>The NCATE Interpretation from the Resource Paper by Rick Stiggins</u>

As noted above, the NCATE staff seem to be reliant on the paper commissioned for them by Richard Stiggins (2000). There, Stiggins uses the terms "accurate" or "accuracy" at least 12 times. (See Appendix A for the complete set of citations.) Of most importance in this comparison between the *Assessing the Assessments* resource and the Stiggins' resource is the conclusion written by Stiggins. There he writes about the benefits of building a strong assessment system, citing, among other things, the following (emphasis added):

> The evidence of competence in teaching will be of higher quality. Both formative and summative assessments will be more *valid and reliable* because those developing and implementing them will know what they are doing." (p. 23).

This conclusion is then repeated in the NCATE executive summary of Stiggins' work on page 2 of the paper – a summary which never mentions the word "accuracy" but keeps the words "valid and reliable."

It appears from quotations in the appendix and the above summary, that it was never the intention of Stiggins to substitute the term "accuracy" for "validity," and that earlier versions of NCATE interpretations were appropriate and consistent with Stiggins's work. He, like Popham, seems to understand the difference between a popular term and a professionally acceptable one.

Stiggins writes about accuracy as an all-encompassing term, making clear references to validity and bias throughout his paper as well as references to sources of error and reliability issues in point #6. He uses much of the language of typical discussions of validity, such as ensuring coverage of targets or content, "representative samples" and "relevant domains." These concepts are fundamental aspects of a content validity study.

<u>The Scholarly Interpretation</u>

So what are accuracy and validity, why is validity more important, and why is it so unwise for NCATE to tell institutions that NCATE does not require validity? These questions point to the nationally accepted standards that guide (1) the measurement profession in developing and using assessment devices and (2) the legal profession in defending or prosecuting claims that an illegal result has occurred. These standards are called the *Standards of Educational and Psychological Testing* (1999). In these *Standards,* the term "accuracy" is neither in the index nor in the table of contents. Validity and reliability are discussed in separate chapters. Part II of the three-part *Standards* book is devoted to fairness. The *Standards* define validity as follows:

> Validity is, therefore, the most fundamental consideration in developing and evaluating tests… Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. (p. 8 and 9)

While the definition may sound a bit technical, in essence, it says that validity means that assessors are making justifiable interpretations about their data and good decisions. In a world where teacher preparation institutions are preparing teachers for American's children, what is wrong with that and why should NCATE be so afraid of it?

Tests are later defined to include performance assessments, portfolios, and other methods of assessment. The *Standards* recommend the collection of multiple forms of evidence of validity gathered from a broad range of studies. It is in this form of evidence that concepts such as representativeness and domain sampling are introduced. This advice was well heeded in the two NCATE documents, with these terms used in one or both.

The *Standards* make it clear that there are many forms of evidence that can be collected. Forms of evidence listed include evidence based on four general areas: response processes, internal structure, relations to other variables, on the consequences of testing. The more familiar forms of validity (predictive, concurrent, discriminant, convergent, instructional, and so forth) are embedded within these organizational categories to demonstrate the on-going and unitary nature of the evolution of the term and its application in the field. The important thing, however, is to recognize the criticality of validity as the "sine qua non" of assessment (that without which nothing else matters in the assessment world). Stiggins' conclusion, remembered in his posted NCATE paper, but lost in the current NCATE application, was correct.

It is time now to move to a brief discussion of some of the best of the literature that predates and supports the current version of the *Standards* as published in 1999 but does not support the use of the term "accuracy" as a substitute term. In a definitive book entitled *Test Validity* (Wainer & Braun, 1988) the term "accuracy" is neither in the index nor the table of contents. In *Educational Measurement (3rd): A Handbook* sponsored by the National Council of Measurement in Education and the American Council on Education (Linn, 1989), the same is again true. In *Educational Research, Methodology, and Measurement: An International Handbook* (Keeves, 1988), the term "accuracy" is again absent.

What is "accuracy", then, if it is not listed under reasonable reference sources? "Accuracy,"as a technical term, is the state of being precise or correct according to a traceable reference standard such as those found in the *International System of Units* whose current website lists 76 citations on the keyword "accuracy" or the *National Institute of Standards and Technology* where a site search reveals 5,520 hits. It is always computed as a statistical process based on known and defined units.

Accuracy, then, is decidedly statistical in scholarly literature, but what about validity? Does it have to be statistical? Above the argument was made for use of content validity as a judgmental process. Leading scholars in the previously mentioned works confirm this:

In *Test Validity* (Wainer & Braun, 1988), the first entry on page 4 is "Validation as Evaluation Argument"—a judgmental approach.

In *Educational Measurement (3rd) A Handbook*, it states in the chapter on *validity (*Messick, 1989), "By *evidence* is meant both data, or facts, and the rationale or arguments that cement those facts into a justification of test-score inferences," again a qualitative approach (p.15-16).

In the *Educational Research, Methodology, and Measurement: An International Handbook*, Zeller (1988) writes, "…the key question underlying validity-inferring the dimensional nature of the posited theoretical concepts-is not a question that lends itself to a solely statistical solution" (p. 330).

Again, it appears that the best reflective thinking on the definition and use of the term "validity" clearly allows both judgmental and empirical evidence produced in a variety of ways, while NCATE misinterprets the term validity as "statistical" only.

There are many, many acceptable ways to develop evidence of validity, providing colleges of education with many choices in how they can approach the gathering of evidence of validity for accreditation assessments. This paper describes some of these in more detail under "fairness" where they appear embedded incorrectly (by NCATE). There are obviously variations and overlaps among techniques, but the idea here is to illustrate the wide variety of methods that are described and used for obtaining evidence of validity in assessments so that institutional decision-makers can set assessment policies from a position of choice and information.

A number of practical applications are provided by *Author* (2007) and are summarized below:

Content validity (judgmental with frequency counts): Alignment charts showing coverage of standards and indicators across assessment tasks

Content validity (judgmental with percents): Stakeholder survey for criticality and authenticity of performance assessments

Content validity (empirical, formula-based): Computation of the Lawshe (1975) Content Validity Ratio (CVR) based on stakeholder ratings of criticality

Content validity (empirical, IRT – Rasch model): Calibration of items on a logistic ruler showing gaps in coverage

Consequential validity (judgmental with frequency counts): Study of the appropriateness of reasons for teacher failures or program dropouts

Consequential validity (judgmental with percents and potentially a chi square): Study of disparate impact on protected populations

Predictive validity (empirical with correlation coefficient): Correlation of scores on selected assessments or measures during the program (e.g., scores on internship evaluation or GPA) with future measures (e.g. principals' satisfaction ratings or hire/rehire status of graduates)

Concurrent validity (empirical with correlation coefficient): Correlation of scores on selected assessments or measures during or near the end of the program (e.g., scores on internship evaluation or GPA) with other measures during the or near the end of the program (e.g., PRAXIS scores).

Convergent validity (empirical with correlation coefficient): Correlation of scores on selected assessments or measures with other measures offered that provide similar information (e.g., two sets of similar tasks).

Divergent validity (empirical with correlation coefficient): Correlation of scores on selected assessments or measures with other measures offered that provide different information (e.g., two different measures that tap different attributes such as cognitive and affective).

Note that many of the validity studies proposed above are judgmental with nothing more than frequency counts (or possiblY percents) suggested. Descriptive statistics are possible, just as they are in data aggregation techniques currently used in most institutional reports. Most of the current accreditation software producers provide means and standard deviations on virtually everything they report, and there is no widespread uproar asking them to stop this practice.

Many of the other studies require nothing more than some simple correlations – Pearson's product moment, Spearman's rho, or point biserial, depending on the level of the data – interval, ordinal, or nominal. These can be easily prepared in Excel or free statistical shareware. Mastery of SPSS and SAS are not prerequisites to these studies. Data input is simple, and most master's level students know the levels of measurement and that a correlation coefficient of .70 is not bad (except for divergent validity, where one would want it to be much lower.)

Undergraduate (not to mention graduate) students study validity. "Validity" is a term that appears in virtually every assessment, evaluation, research, or measurement book across the country. NCATE needs to accept the terminology, applying the *Standards for Educational and Psychological Testing* and best practices of the assessment world to the standards of the accreditation world.

Consistency (vs. Reliability)

<u>The NCATE Interpretation</u>

In *Assessing the Assessments* the authors state:

> Assessments are consistent when they produce dependable results or results that would remain constant on repeated trials. Essentially, in approaching consistency, the standards are requiring that the assessments and results be trustworthy. In other words, if the same information were assessed on different occasions, with no intervention, and the results were largely the same, then the assessment could be said to be consistent.

The NCATE authors are correct that the term "consistency" is often linked with reliability, perhaps more so than "accuracy" is to "validity." The problem here, though, is that they limit the discussion to a specific form of reliability, test-retest reliability, which is not likely to be useful in the accreditation process. How often are college faculties likely to replicate an assessment task just to see if they obtain consistent results? It is not typically possible or even advisable. Repeated trials are not the norm. Similarly, it is not normal to repeat the measure without intervention. How could we explain to teacher education candidates that they need to take the same assessment a second time just to see if they provide a similar response, or worse, to see if they have not improved since no intervention was provided? The NCATE authors continue:

> Consistency is closely related to the statistical term "reliability." However, NCATE consciously chose not to use the term "reliability" because the concept can be adequately addressed with methods that can be inclusive of, but also other than statistical analysis.
>
> Institutions can document consistency through providing training for raters that promote similar scoring patterns, using multiple raters, conducting simple studies of inter-rater reliability, and/or by comparing results to other internal or to external assessments that measure comparable knowledge, skills and/or dispositions.

Here, the authors of the posted NCATE paper talk in circles, even suggesting studies of inter-rater reliability but avoiding the use of the term "reliability". There is no explanation as to why they failed to rename it "inter-rater consistency. Virtually all of the strategies suggested – determination of similar scoring patterns, inter-rater reliability, and comparison of results to other measures require a number or two despite the stated intent to avoid statistics. The third item in the NCATE list, comparison to other measures, is really much more a measure of concurrent validity than it is a measure of reliability. Again, there is support in the literature for judgmental approaches to ensuring reliability, although it is less common than it is for validity.

<u>The Scholarly Interpretation</u>

The AERA, APA, NCME *Standards* define reliability as "the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups." Here, the word "consistency" is used as a definition, unlike the case of validity and accuracy. The discussion of reliability in these *Standards* and any other reputable

sources involves an assessment of error in scores. The *Standards* discuss sources of errors that constitute reliability problems and those that do not. They note that measurement error (lack of reliability) is generally viewed as random and unpredictable. Typical sources of measurement error are broadly categorized as those which are rooted within the examinee and those that are external to the persons tested.

Examples of errors rooted within the examinee are fluctuations in motivation, interest level, or attention, and the inconsistent application of skills. Examples of external sources are differences in test site, freedom from distractions, the random effects of score subjectivity, and variation in scoring. (P. 26)

Raters are trained, scores compared, alternative tests provided, retesting administered, all with the ultimate goal of reducing error present in a score. The NCATE use of the term "consistency" does not change that. Using the term consistency does not reduce the use of statistics!

Scannell & Ingersoll (2002) and *Author* (2007) both suggest that the traditional Cohen's Kappa can be computed to estimate error across raters, and the formula is simple to apply. Free calculators are available online. Expert rescoring of instruments and the insertion of artificial cases to check the consistency of new scorers with established exemplars is another easy way to look for reliability.

What is of particular interest, here, though, is the close alignment of "accuracy" and "consistency" in a scholarly treatment of reliability. In the previously mentioned *Estimating the Consistency and Accuracy of Classifications Based on Test Scores* published in the *Journal of Educational Measurement*, the following definition is provided:

> The term consistency refers to the agreement between the classifications based on two non-overlapping, equally difficult forms of the test. (p. 184)

Within modern item response methods there are improved statistical methods of estimating reliability as internal consistency and individual student estimates of error on each item response (Smith, 2004). Also, new statistical methods of the Rasch model have contributed greatly to detecting and correcting rater effects (Myford & Wolfe, 2004). Even though the use of statistics is necessary for some methods, the math is now done with readily available software for the most complex analyses (Winsteps, FACETS, RUMM, 2005).

Again, *Author* (2007) have suggestions and examples of methods to produce evidence of reliability typical of assessment in the accreditation process:

1. **Inter-rater reliability** (judgmental with narrative): Experts rescore submissions to check the extent to which they agree with raters' decisions, or new raters score

artificial or anchor cases to determine if they are interpreting the rubrics correctly. Differences are noted for further analysis.

2. **Inter-rater reliability** (empirical with kappa statistic): Cohen's kappa is used to check the consistency of pairs of raters on multiple items.

3. **Inter-rater reliability** (empirical with rater calibration and score adjustment): Rater effects are estimated using FACETS (Rasch model) so that individual students' scores can be adjusted to a score that compensates for the leniency or harshness of the individual raters, avoiding the need to rescore once a reliability issue is found.

4. **Internal Consistency Reliability** (empirical with alpha): Cronbach's alpha is computed to see if the items on the assessment all tend to measure the same construct.

5. **Confidence Intervals** (empirical with standard error): The standard error is used to determine confidence intervals for student scores, so that the unit can decide if the assessments are adequately precise and accurate (minimal error) to have confidence in the decisions being made.

6. **Separation and Fit Analysis:** (empirical with separation and fit statistics): Separation and fit statistics are calculated using the Rasch model of Item Response Theory. The separation statistic is similar to Cronbach's alpha and is an indicator of reliability. The fit statistic results from a differential person fit analysis, which is conducted to determine if there are measurement problems associated with students' responses to individual items based on factors external to their knowledge and skill related to the construct.

Fairness

<u>The NCATE Interpretation</u>

In *Assessing the Assessments,* the authors treat fairness and bias as two separate sections of their resource. In actuality, bias is a subset of fairness, and this discussion will address bias in more detail in the next section. The two NCATE resources are so entangled and incorrect on these constructs that is difficult to sort out the problems in a coherent way. Again, quotations from the relevant section (fairness) in the *Assessing the Assessments* piece serve as a starting point:

> Assessments are fair when they assess what has been taught. Candidates should be exposed to the knowledge, skills, and dispositions which are measured in the assessments. Without this type of exposure, it is not fair to expect candidates to have mastered the material.

> One example of how institutions can demonstrate fairness in their key assessments is through curriculum mapping (e.g., a chart that shows where in the curriculum candidates have the opportunity to learn and practice what is specified in the standards). Institutions should identify where in the curriculum candidates have had the opportunity to learn and practice the material being assessed.
>
> In addition, fairness also means that candidates understand what is expected of them on the assessments. To this end, instructions and timing of assessments should be clearly stated and shared with candidates. In addition, candidates should be given information on how the assessments are scored and how they count toward completion of programs.

There is much confusion here with content and instructional validity, and it is appropriate to again address it through a review of the literature. Had validity been used appropriately in the NCATE Standard, the need to confound it with fairness would not have occurred. Here, the key elements of fairness from a technical standpoint are missing. In Stiggins' commissioned paper, the word "fairness" is not used at all. He discusses, instead, bias.

The Scholarly Interpretation -- Fairness

As just stated, to "assess what has been taught" is not *fairness,* but it is *instructional validity.* This term was mentioned earlier in the Introduction as the form of validity that surfaced as a result of the Debra P. v. Turlington (1979) case. It is now commonly used in legal (Rebell, 1998) and educational (Hardy, 1984) discussions to mean that the assessment is constructed of material that was adequately taught. For a detailed description of *instructional validity* and the related phrase, *opportunity to learn,* see Jaeger (1989, pp. 500-509). This interpretation of the resource paper is supported by the next paragraph which refers to *curriculum mapping,* where the phrase "opportunity to learn" is actually used. The term "curriculum mapping," however, was defined by Jacobs (1997) to have yet a different meaning. Perhaps the paper meant *curriculum validity* sometimes used as a synonym for *instructional validity* (Jaeger, 1989). The implication is that units need to demonstrate that the assessment material is balanced or covers the standards. This is what is more commonly called *content validity.* Of course, the paper may have been pointing to *curriculum mapping* (Jacobs, 1997), a planning method that has no obvious use as psychometric evidence of *fairness.* Instead, Jacobs conceived of curriculum mapping in terms of lesson planning within a planned scope and sequence of materials? Finally, there is *concept mapping* (Wilson, 2005, p. 6) that is defined as a way to create a "more precise concept than a *construct"* within the context of a *conceptual framework* as a pretext to the measurement process.

Giving this level of possible misunderstanding with similar terms it becomes useful to clarify the technical definition of "fairness". The *Standards for Educational and Psychological Testing* (1999) describe fairness" as (1) lack of bias, (2) equitable treatment in the testing process, (3) equality in outcomes of testing, and (4) opportunity to

learn (remember the related term *instructional validity?).* Having noted that bias is more appropriately a part of fairness, it is time to turn to that term.

Avoidance and Elimination of Bias

The NCATE Interpretation

Again, this discussion of avoidance and elimination of bias starts with the citations from *Assessing the Assessments:*

> Closely related to accuracy is the elimination of bias. To ensure that the results of assessments adequately reflect what candidates know and can do, it is important to remove any contextual distractions and/or problems with the assessment instruments that introduce sources of bias and thus adversely influence candidate performance. Contextual distractions include inappropriate noise, poor lighting, discomfort, and the lack of proper equipment. Problems with assessments include missing or vague instructions, poorly worded questions, and poorly reproduced copies that make reading difficult.
>
> The elimination of bias also means that the assessments are free of racial and ethnic stereotypes, poorly conceived language and task situations, and other forms of cultural insensitivity that might interfere with candidate performance and unintentionally favor some candidates over others. Further, the elimination of bias includes consistent scoring of assessments and vigilant efforts not to discriminate against groups of candidates.[1]
>
> Ultimately, it is important that units evaluate assessments and assessment conditions and eliminate as many sources of bias as possible. While the standards use the term "eliminate," in fact, it is best to *avoid* sources of bias as the assessments are being developed.

The above citation begins with an immediate and obvious error. It was noted that the technical definition of bias is a part of fairness, not accuracy (a.k.a., validity from NCATE's perspective). Contextual distractions are sources of error and therefore are more related to reliability than bias, although there is discussion in the AERA, APA, NCME *Standards* that could be interpreted to mean that such distractions can be systematic and are, therefore, related to construct irrelevant variance – a validity problem. Either way, they are not bias. The footnote is from Stiggins, who is the first source of this error.

The discussion of bias with regard to stereotyping and discrimination is appropriate to this section, although the gathering of evidence of this form of bias, as will be shown later, is typically statistical in operation. Bias can never be "eliminated," because there is always the potential to see differential performance among groups. The only realistic act is to attempt to avoid it.

Accuracy vs. Validity, Consistency vs. Reliability, and Fairness vs. Absence of Bias: A Call for Quality

In the Stiggins paper, sources of bias that can distort results and cause a problem are discussed. He provides a list of common sources of bias in his figure 3 on page 11. He divides them into two categories, both with sub-categories. The first is sources of bias common to all assessment methods. The second provides sources that are unique to each format and are more closely associated with what is typically presented as item writing criteria. Here is a list of the three common sources of bias with selected examples from his list:

> Potential problems that can occur within the student (e.g., reading, language, emotion, health)
>
> Possible problems that can occur within the assessment context (e.g., noise, lighting, cultural insensitivity)
>
> Examples of problems that arise from the assessment itself regardless of method (e.g., poor directions, wording, or reproduction).

The list is informative but inconsistent with other scholarly and professional approaches. As noted above, these sources of bias are more correctly classified as either sources of measurement error (reliability) or construct irrelevant variance (validity).

<u>The Scholarly Interpretation</u>

An examination of the *Standards of Educational and Psychological Testing* (AERA, APA, and NCME, 1999) provide a more credible discussion of fairness, which is the subject of one of three parts to the book and four chapters. Here, we will list section headings only in outline form with a brief synopsis of their meaning:

> **Lack of Bias**: Bias is defined as a technical term reflecting either of two situations in which examinees from protected populations score differently because of deficiencies in the test itself or are offended by an assessment.
>
> **Equitable Treatment**: Members of protected classes are not treated differently and have equal opportunities to succeed in similar settings and with the same materials.
>
> **Equality of Outcomes**: Overall pass rates need to be comparable across groups. The *Standards* defer here to the legal requirements, which are defined as the 80% rule. This is accompanied by the term disparate impact.
>
> **Opportunity to Learn**: Examinees who are presented with inadequate instruction are likely to score lower, and this context needs to be avoided. Curriculum review is a predominant strategy to overcome this issue, but the focus is on determining if specific groups have differential opportunities.

The *Standards* also define fairness in terms of bias associated with test content and response processes. Here, the issue is again the identification of components that result in systematically lower or higher scores for identifiable groups of examinees. These could be due to inappropriate sampling, lack of clarity in instructions, failure to assign partial credit, or other factors. The language and types of evidence overlap significantly with evidence of validity, but the usage is different and the focus is on protected populations or other group differences. In examining test content, one looks for language that could be interpreted differently by members of different groups or for material that could be offensive or emotionally disturbing to some test takers. One might also look for different patterns of response rooted in test instructions, and other aspects of testing. The *Standards* continue with a discussion of fairness in selection and prediction, and again the focus is on group outcomes due to choice of predictors.

The *Standards* consider the rights and responsibilities of test takers, testing individuals with diverse linguistic and cultural backgrounds, and testing individuals with disabilities. In each case, the protected population should not score differently because of their membership in that class. The rights of test takers include predominantly the right to be fully informed about all aspects of the test, and these rights serve as a link to due process expectations.

The different patterns that are possible among groups, however, are the key. It is the thread that binds all issues of fairness together. Group membership could be based on socio-economic status, native language, religious affiliation, parental status (e.g., single parent), or membership in protected population (e.g., females and minorities). Judgmental reviews are often supplemented by statistical procedures for identifying items that function differently across identifiable subgroups of examinees (e.g., differential item functioning or DIF.)

Author (2007) suggest several studies of fairness, again both judgmental and empirical:

> **Non-Biased Materials and Processes or Offensiveness** (judgmental): Analysis of items by a representative set of teachers and assessors (including all protected populations and other groups) for language or context that might offend any sub-group.
>
> **Equal Opportunity and Non-Discriminatory Practices or Bias** (judgmental): Analysis of reasons for non-completion and sufficiency of remediation efforts compared for different membership groups.
>
> **Equal Opportunity and Non-Discriminatory Practices or Bia**s: (empirical with percents and possibly a chi square): Disparate impact analysis to ensure that at least 80% of the proportionate percent of each protected population succeeds.

Equal Opportunity and Non-Discriminatory Practices or Bias: (empirical with Differential Item and Person Functioning – Rasch IRT Model): Analysis to determine if differences in group performance are statistically significant.

Conclusions and Policy Recommendations

This paper has recreated key components of NCATE source documents related to NCATE Standard 2, Assessment System and Unit Evaluation, and juxtaposed them against the standards of the measurement profession, The *Standards of Educational and Psychological Testing* (AERA, APA, NCME, 1999). These *Standards* are the professional and legal standards to which all professional educators are held accountable when they conduct any assessments (e.g., multiple choice tests, performance tasks, or portfolios) to make high stakes decisions (e.g., graduation and certification).

This paper asserts that NCATE Standard 2 contains language that is inappropriate. In its resource materials, that are almost reference-free, NCATE has substituted the word "accuracy" for "validity" and the word "consistency" for "reliability" with a notable and stated bias against statistics and a clear misperception that judgmental approaches to gathering evidence of validity and reliability are nonexistent. NCATE has confused the term "fairness" with "validity" and the term "bias" with "reliability." While there are many other errors in their work, this paragraph summarizing the big issues is one that should send shivers up the spines of all professional educators.

No graduate student would present a paper without references attached. Every master's level and doctoral student in the country takes at least one course on educational research that incorporates the level of statistics required for an acceptable empirical analysis of validity and reliability. More complicated things are possible and sometimes warranted, but for the most part, as has been demonstrated herein, Microsoft Excel and/or shareware will do the job adequately.

From a policy perspective, the failure to use commonly accepted terminology, such as "validity" and "reliability," substituting non-standard terms, such as "accuracy" and "consistency," compounded by informal, non-scholarly, and even incorrect definitions and examples of these two terms as well as the term "fairness", weakens the ability of Board of Examiners members to render valid, reliable, and fair (or accurate, consistent, and fair) professional judgments on the quality of teacher education programs. The BOE reviewers arrive on-site without the tools of their trade – professionally acceptable terms that are supported by scholarly definitions, ones they learned in graduate school themselves. The institutions, in turn, work hard to develop the policies and procedures needed for success but without the adequate tools and examples to do so. Given a license for assessment sloppiness, they rely, instead, on software solutions that sometimes play out the worst fears of educational researchers about statistical misuse.

At a policy level, both NCATE and teacher education units need to determine the extent to which they want to continue to use informal terms, accompanied by unscholarly and blatantly incorrect definitions and explanations. The alternative decision calls for the

use of scholarship and the wisdom of practice – a choice to practice what is preached. This is a difficult choice for both. We summarize the issues for both groups below with a suggestion to both to model best practice.

For NCATE

For NCATE, policy makers need to determine if they want to rewrite a standard and its accompanying rubric and explanations that are flawed. If they continue with the current version, the agency may face the slings and arrows of a federal Department of Education that is nipping at the heels of the accreditors, constituting committees that intend to overhaul them. This paper may add fuel to that fire. Specifically, recommendations for NCATE are as follows:

Conduct and use a scholarly review of the literature on how to assess assessments, so that the terms are well understood in-house for the purposes of BOE and institutional training and external communication.

Revise the standard and all associated materials, correcting the discussion of fairness and avoidance of bias, so that the terms are not all confounded with each other. Use commonly accepted terminology, specifically "validity" and "reliability" instead of "accuracy" and "consistency."

For Institutions

Institutions need to decide if they should use the NCATE jargon because they think it is politically correct and safer to comply. If they do, they may face a Board of Examiners team that attempts to apply incorrect definitions, or, worse, they may face a member or two that knows the difference. If they move forward with research and practice-based terminology and methodologies, using current writings on the subject (e.g., *Author*, 2007), they may be able to regain control of the process through scholarship. Specifically, institutions should:

Review the literature to acquire a professional level of understanding of validity, reliability, and fairness that goes beyond the materials on the NCATE website. Institutions should not be afraid to use the correct terms and should be prepared to educate BOE teams that attempt to apply fuzzy interpretations of fuzzy terms that may yield decisions that are not valid, not reliable and/or not fair.

Remind program faculty that they teach students to make valid, reliable, and fair decisions about student learning, and that they should model that practice in examining their own decisions.

Use measurement and evaluation professionals to help develop evidence, on an as needed basis. Administrators should what they want. They should know how to say "no" to complex statistical procedures and "yes" to an appropriate blend of judgmental (qualitative) and empirical (quantitative) evidence.

References

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education (1999). *Standards for educational and psychological testing.*

Beggar, H. & Zornes, D. (2006). *Criteria for evaluating performance assessments*: *2006 NCATE clinic, "Task Force Discussions."* Downloaded June 22, 2007 from http://www.ncate.org/documents/clinics/2006/CriteriaEvaluatingPerformanceAssessments.doc .

Boldt, R. F. (1983). *Review for perceived bias on ASVAB forms 11, 12, and 13* No. ETS-RM-83-4)Educational Testing Service, Research Publications, R-116, Princeton, NJ 08541.

Bond, L. (1998). Disparate impact and teacher certification. *Journal of Personnel Evaluation in Education, 12*(2), 211-220.

Breland, H. M. (1978). *Sex and ethnic comparisons of test validity using A blind-scored essay as A criterion*

Brown, C. R., & Moore, F. L. (1994). Construct validity and context dependency of the assessment of practical skills in an advanced level biology examination. *Research in Science and Technological Education, 12*(1), 53-61.

Campbell, J. T., & And Others. (1973). *An investigation of sources of bias in the prediction of job performance: A six-year study. final project report* No. PR-73-37)Educational Testing Service, Rosedale Road, Princeton, New Jersey 08540 ($10.00).

Cronbach, L. J. (1949). Essentials of psychological testing. New York: Harper & Row.

Cureton, E.E. (1950). Validity, reliability, and baloney.*Educational and Psychological Measurement. 10:1,* 83–95.Denzine, G. M., Cooney, J. B., & McKenzie, R. (2005). Confirmatory factor analysis of the teacher efficacy scale for prospective teachers. *British Journal of Educational Psychology, 75*(4), 689-708.

Debra P. v. Turlington, 730 F.2d 1405 (11th Cir. 1984).

Distefano, M. K., Jr., & And Others. (1983). Application of content validity methods to the development of a job-related performance rating criterion. *Personnel Psychology, 36*(3), 621-631.

Elliott, E. (2001). *Aligning assessments with standards: A synthesis of guidelines from current practice adapted for use in teacher education and NCATE accreditation.* Downloaded June 22, 2007 from http://www.ncate.org/documents/aligning%20assessments%20and%20standards.pdf .

Elliott, E. (2005). *Student learning in NCATE accreditation.* Downloaded June 22, 2007 from http://www.ncate.org/documents/papers/STUDENT_LEARNING_4th.pdf .

Fidler, J. R. (1993). An application of practical strategies in assessing the criterion-related validity of credentialing examinations. *Evaluation and the Health Professions, 16*(1), 13-43.

Grill, J. J., & Bartel, N. R. (1977). Language bias in tests: ITPA grammatic closure. *Journal of learning disabilities,*

Gonzalez-Tamayo, E. (1991). *A closer look at test scores, selection and prediction.* ERIC Document Reproduction Service (ED339702).

Hagtvet, K. A. (1997). The function of indicators and errors in construct measures: An application of generalizability theory. *Journal of Vocational Education Research, 22*(4), 247-266.

Ingersoll, G. M., & Scannell, D. P. (2002). Performance-based teacher certification: Creating a comprehensive unit assessment system. Golden, CO: .Fulcrum

Jackson, J. F. L. (2006). Hiring practices of african american males in academic leadership positions at american colleges and universities: An employment trends and disparate impact analysis. *Teachers College Record, 108*(2), 316-338.

Keeves (Ed.) (1988). Educational research, methodology, and measurement: An international handbook.Oxford:Pergamon Press.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4). 563-575.

Linacre, J. M. (2003). *A user's guide to Winsteps and Ministep: Rasch-model computer programs.* Chicago: Winsteps.

Linacre, J. M. (1994). *A user's guide to FACETS.* Chicago, MESA Press.

Linn, R. L. (1989). *Intelligence: Measurement, theory, and public policy.* Urbana: University of Illinois Press.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement, 32,* 179-197.

Littlefield, R. S. (1983). Competitive live discussion: The effective use of nonverbal cues. The Forensic, 69, (2), 14-20.

Markert, R. J., & Shores, J. H. (1980). *Assuring fairness in the medical school admission interview through analysis of rater difficulty and consistency*

McKillip, J., & Cox, C. (1998). Strengthening the criterion-related validity of professional certifications. *Evaluation and program planning, 21*(2), 191-197.

Meijer, R. R. (1994). The number of guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4), 311-314.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Miller-Whitehead, M. (2001). *Inter-rater reliability on performance criteria: Theoretical issues* No. R-01-1018-01)Tennessee Valley Educators for Excellence, P.O. Box 2882, Muscle Shoals, AL 35662. Tel: 256-446-5278; Fax: 256-446-5278; Web site: http://www.dpo.uab.edu/~tnmarie.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the graduate record examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55*(2), 309-316.

Newell, J. A., Dahm, K. D., & Newell, H. L. (2002). Rubric development and inter-rater reliability issues in assessing learning outcomes. *Chemical Engineering Education, 36*(3), 212-215.

National Council for Accreditation of Teacher Education (2002). *Professional Standards for the Accreditation of Schools, Colleges, and Departments of Education.* Washington, D.C.: Author. Retrieved November 4, 2004 from http://www.ncate.org/2000/unit_stnds_2002.pdf Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7*(2), 143-164.

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

Roskam, E. E., & And Others. (1992). Commentary on guttman: The irrelevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27*(2), 205-267.

RUMM (2003). *Rasch Unidimentional Measurement Models.* Duncraig: Western Australia: RUMM Laboratory.

Stiggins, R. (2000). *Specifications for a Performance-Based Assessment System for Teacher Preparation.* National Council for Accreditation of Teacher Education, Washington, D.C. Retrieved June 15, 2004 from http://www.ncate.org/resources/commissioned%20papers/stiggins.pdf

Trochim, William M. (2006) *The Research Methods Knowledge Base, 2nd Edition*. Internet WWW page, at URL: http://trochim.human.cornell.edu/kb/index.htm (version current as of August 10, 2006).Van Rooy, D., L., & Viswesvaran, C. (2004). Emotional intelligence: A meta-analytic investigation of predictive validity and nomological net. *Journal of vocational behavior, 65*(1), 71-95.

Wainer, H., & Braun, H. I. (1988). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

Wright, B. D. (1995). Which standard error? *Rasch Measurement Transactions, 9*, 436-437.

Wightman, L. F., & Muller, D. G. (1990). An analysis of differential validity and differential prediction for black, mexican american, hispanic, and white law school students. LSAC research report series No. LSAC-R-90-03)

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn and equity: New standards examinations for the california mathematics renaissance* No. CSE-TR-484)

Yoshida, R. K. (1973). A guttman scalogram analysis of haptic perception for trainable mentally retarded children. *American Journal of Mental Deficiency,*

Zeller, M.J. (1988) Titles of jobs in human services for students with a bachelor's degree in psychology, In P.J. Woods (Ed.), *Is Psychology for Them?: A Guide to Undergraduate Advising* (pp. 195-196). Washington, D.C., American Psychological Association.