

Evaluating NCLB's peer review process: A comparison of state compliance decisions

Susan L. Davis

Chad W. Buckendahl

Buros Center for Testing

University of Nebraska - Lincoln

Paper presented at the 2007 annual meeting of the National Council on Measurement in Education, Chicago, IL.

\*\*Draft version – please do not cite without permission from the authors.

## Abstract

Peer review processes are commonly found within professional communities. These may be in the context of professional conferences, journals, books, audits, or accreditation programs that often provide independent, and sometimes anonymous, reviews of quality. The outcomes of these programs may be used for formative and/or summative purposes depending on the intended uses. The peer review process that is part of the federal review of state assessment and accountability programs under No Child Left Behind (U.S. ED, 2004) represents another example of how evaluations of complex systems occur. This primary purpose of this paper is to compare the NCLB peer review process with other peer review processes that occur within the professional community in the context of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). A secondary purpose is to discuss selected letters from USDOE to state departments of education as exemplars of the outcomes of this process. Implications for practitioners who may be responsible for future peer review processes are also included.

Keywords: Peer review, NCLB, Educational Policy

## Evaluating NCLB's peer review process: A comparison of state compliance decisions

The inception of the *No Child Left Behind* (NCLB) federal legislation (Public Law 107-110, 2002) represented a major increase in the stakes associated with K-12 educational assessment. All states from this point forward were required to develop, implement, and clearly document a system by which they could assess the learning and knowledge of students in grades 3-8 and high school in reading and mathematics. The design of the NCLB legislation allows states to maintain autonomy in defining expectations for student learning and how student learning is assessed. This flexibility in approach to design, however, necessitated a federal accountability system. States are required to submit documentation of their assessment system for review by experts in the fields of standards and assessments (U.S. ED., 2004). This is referred to as the NCLB peer review process whereby experts review state assessment programs, as documented in submitted materials, and provide judgments and feedback to the United States Department of Education (U.S. ED) based on pre-defined review criteria.

Both criticisms and support have been voiced many stakeholders regarding many aspects of the NCLB system. The most criticized aspect of NCLB is certainly the expectation that 100% of students need to meet a given state's standard for proficiency by 2014 in reading and mathematics which is considered an unrealistically high goal (e.g., Linn, 2005; Mathis, 2003). Given the importance of the NCLB peer review process within this system, and considering the stakes associated with the decisions based on the outcomes of this process, it is important to take a critical look at the NCLB peer review and determine if this system follows professional expectations for peer review. Although

these processes may be communicated as very clear in design and structure, it may be difficult to implement such a model objectively and systematically across assessment systems. Many components of the federal peer review parallel the peer review processes that are common within the academic community. These models serve as one basis for comparison; however, other strategies have also been observed in practice.

The primary purpose of this paper is to examine the NCLB peer review process in comparison to other models of peer review to evaluate whether the process was designed to be implemented in a systematic fashion across all assessment systems. This review has been conducted in three stages focusing on (1) the design of the review system, (2) the specific criteria used to review the assessment systems, and (3) the results of the review. A secondary purpose is to discuss selected letters from USDOE to state departments of education as exemplars of the outcomes of this process.

### *Peer Review*

The peer review process is a commonly-used practice by which scholarly work can be reviewed and evaluated by experts to ensure it meets the expectations of the field or discipline. There are many variants on the methods by which peer review is executed; however, the basic tenants of the process are present in these various applications. In most cases work, or proposals for work, is submitted to some type of agency, organization, or publisher. The agency or organization solicits reviewers from a pool of experts within a given field and matches the review tasks with reviewers who have expertise in relevant disciplines. Depending on the process, reviewers are provided guidance on how to review the work. The reviewers return their judgments to the agency or organization and this information is used to make a decision about the work (or

proposal). Finally the authors are informed of the decision and any comments/recommendations from the reviewers. This can be a blind process (reviewer identity is not disclosed), a double blind process (reviewer and author identities are not disclosed), or one that contains full disclosure (both author and reviewer identities are revealed).

As educational researchers, the process of peer review is very common practice and anyone reading this paper has likely participated in this process in at least one form or another. Peer review is used in journal publication for reviewing manuscripts, in professional conferences for reviewing paper proposals, and in contract and grant work for reviewing project proposals. More specifically, within the testing community, several systems are in place by which tests and/or testing programs can be reviewed. This includes organizations that conduct quality audits of testing programs (e.g., NOCA, BIACO, ANSI, ETS) whereby the processes and procedures used within the testing program are evaluated against criteria set by the auditing organization. In addition to such audits are processes used by organizations that provide independent reviews of commercially available tests to the public such as the Buros Institute for Mental Measurements' *Mental Measurements Yearbook* series (e.g., Spies & Plake, 2005). These reviews are prepared by independent reviewers and made available to potential test users. Finally, even closer to the NCLB peer review process is one component of Nebraska's state assessment model – a system that allows districts and local education agencies to formulate their own assessment systems that are subjected to a peer review by professionals with expertise in assessment and testing (Buckendahl, Plake, & Impara,

2004; Plake, Impara, Buckendahl, 2004). This system, in many ways, parallels the NCLB peer review system.

Many, if not most, of these academic peer review processes for tests or testing programs rely on the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) for guidance of best practice as defined by the testing community along with other professional literature that dictates best practice (e.g., *Educational Measurement 4<sup>th</sup> Ed.*, Brennan, 2006). The *Standards*, and other such published literature, define expectations for testing practices related to validity, reliability, test development, test scales and norms, test administration, scoring, and documentation.

As a research topic, peer review has been studied, applauded, and critiqued (Harris, Gao, Welch, 2002; Jefferson, Wager, & Davidoff, 2002; Kassirer & Campion, 1994; Weber, Katz, Waeckerle, & Callahan, 2002). Rothwell and Martyn (2000) found that in journal publication, the agreement among the two reviewers (on two factors – acceptance and priority of publication) for a series of manuscripts was not better than what was expected by chance. Although rater variability is expected, such differences are important to consider as most editors, conference program chairs, or funding organizations, will likely report relying heavily on peer review recommendations when making decisions regarding publications, presentation proposals, or project funding. Analysis methods such as generalizability theory were developed to respond to how error terms might be interpreted and used to mitigate our confidence in scores or decisions.

Given the importance of the decisions that are made based on the NCLB peer review process, it is important that this review process be considered in light of professional expectations as the other aspects of the NCLB system are. As models, this

paper will consider the peer review systems discussed earlier (e.g., journal, conferences, audit of testing program, BIMM, Nebraska accountability model). There is no “gold standard” for peer review – many of these systems have components or steps that are suited particularly for their intended use. However, one can look to these existing models as an initial basis for comparison to highlight strengths and areas of the NCLB peer review system that could be improved.

#### *NCLB Peer Review Process*

Under NCLB, states are required to compile documentation of their assessment and accountability systems and submit this documentation for evaluation through the peer review process. Such documentation may cover policies related to the assessment system, the processes used in development, administration, and scoring of the assessments, and the reporting of assessment results. Each peer review team is selected by ED and typically consists of a psychometrician, an educator who works with special populations such as English language learners (ELL), or students with disabilities (SWD), and another testing professional with experience in large-scale assessment (Forte, 2006).

The NCLB peer review process includes a framework for evaluating a state assessment system focusing on seven integrated components of the system (U.S. ED, 2004). Specifically, the U.S. ED expects state assessment systems to have:

1. Challenging academic content standards applied to all public schools and local education agencies (LEAs)
2. Challenging academic achievement standards applied to all public schools and LEAs
3. Annual high-quality assessments
4. Assessments with high technical quality
5. Alignment of academic content standards, academic achievement standards, and assessments
6. Inclusion of all students in the assessment system
7. An effective system of assessment reports

Each of these categories includes specific review criteria framed as questions that must be addressed during the review (U.S. ED, 2004). An important feature of these criteria is that they allow reviewers to focus on the process (e.g., how do states estimate reliability for their test scores and how do they deal with low reliability?) and not the product (e.g., what is the reliability of the test?).

The peer review team first reviews the state assessment materials independently and then meets as a team to discuss their findings and come to consensus. During the consensus meeting, a representative from ED is in attendance to clarify any questions regarding the review criteria so that the team may be consistent in how they apply the review criteria across states. In some cases, the ED representative may also clarify particular questions about the state documentation based on knowledge gained through prior conversations with state representatives. From this meeting, the team produces a report for ED on their evaluation of the assessment system based on the seven peer review requirements. The peer review team only provides comments regarding the compliance of the state assessment model with the seven criteria listed above; they do not make recommendations on the final status of the assessment system (Forte, 2006). The peer review team's recommendations are sent to the Assistant Secretary for Elementary and Secondary Education who is responsible for making a decision about the status of the state assessment system (Johnson, 2006a). The designations are:

*Full Approval:* The state system meets all statutory and regulatory requirements.

*Full Approval with Recommendations:* The state system meets all statutory and regulatory requirements but some elements of the system could be improved. A system with this designation is approved but ED makes specific recommendations for improvement

*Approval Expected:* The state system is compliant with all statutory and regulatory requirements; however, some elements may not be complete until the summer of 2006 (and the review occurred before this time). The state must submit all documentation for review prior to administering the 2006-2007 school year assessments.

*Approval Pending:* The state system has one or a few components of the required elements that are missing or do not meet the statutory and regulatory requirements. To achieve this designation (over Non-Approved), a state must be able to administer a fully compliant system during the 2006-2007 school year.

*Non-Approved:* The state system has many components that are missing or do not meet the statutory and regulatory requirements. In addition, this status indicates that a state may not be able to implement a fully compliant system during the 2006-2007 school year.

After receiving feedback from U.S. ED, a state has the right to resubmit materials to address the shortcomings of the assessment system identified by the previous review. These additional materials are then typically reviewed by the same peer review team, although this is not always possible due to scheduling. If a state system did not receive full approval (e.g., designation of *Approval Expected*, *Approval Pending*, or *Non-Approved*) by the 2006-2007 school year, ED reserved the right to impose sanctions including restrictions on grant awards, restrictions on NCLB flexibility, withholding of Title I funds, or creation of specific compliance agreements. However, U.S. ED did consider requests for time extensions and reconsiderations of particular aspects of state assessment systems through peer review (Johnson, 2006b) before imposing such sanctions. In addition, for those states that have received approval of their system, any potential changes to the system must be submitted for review.

### *Methods*

To understand the peer review process, we evaluated documentation of the process provided by the U.S. Department of Education. Our first archival source was the

peer review guidance document made available to states in advance of the process (ED, 2004). This document outlines each of the NCLB review requirements and was reviewed to formulate a comprehensive understanding of each of the seven requirements. We were also able to evaluate the substantive components of the available state decision letters as they provide specific examples of the feedback states are receiving following the peer review process (U.S. ED, n.d.). These state letters served as a secondary data source for our evaluation.

In relation to the NCLB peer review process we also considered other peer review processes that are a part of publication, presentation, and test evaluation systems that exist within educational assessment. In examining each system, we attempted to identify best practices that reflect guidance from professional literature such as the *Standards* (AERA, APA, & NCME, 1999).

The review of this system was divided into three topics. The first of these is the overall design of the peer review process which includes selection and training of raters, review of state assessment systems, and reporting of results. The second topic of review was the specific review criteria; the seven areas of review defined by U.S. ED. The third topic of review is the specific types of results and feedback that is provided to states for the continual improvement of their assessment systems.

### *Results*

Overall, the NCLB peer review process includes some essential components that are characteristic of the typical peer review process. Importantly, there are additional unique components involved in this system that are warranted given the stakes associated

with the decisions based on this particular peer review process. The comparison presented below is organized within the three topics identified above.

### *Design of the Review System*

The first step in the peer review process is the selection of the peer reviewers. As noted above in the description, these individuals are considered to be experts in the areas of assessment and testing. For each assessment system, U.S. ED compiles a team of three reviewers that includes at least one person who is an expert in working with special populations (students with disabilities, English language learners). Importantly, individuals with a conflict of interest to a particular testing program are excluded from serving on a review team (e.g., employees of testing contractors or individuals who have a relationship with the state under review; Forte, 2006). These selection criteria are similar to other models such as the Nebraska peer review system which also includes testing professionals that are not involved in development of relevant assessments (Buckendahl et al., 2004).

One noteworthy similarity of these two systems (NCLB and Nebraska) is that neither includes content experts (e.g., reading specialist, mathematics curriculum experts) as neither of these reviews includes the academic content standards<sup>1</sup>, assessments, or test items (Plake et al., 2004; U.S. ED, 2004). This is in contrast to systems such as the BIMM review model whereby two reviewers are selected with one primarily having expertise in testing and the other possessing more content-relevant knowledge that is related to the subject matter of the test. This design is used to provide consumers of the

---

<sup>1</sup> Districts in Nebraska are required to use the State content standards or develop their own standards that have been approved as being more rigorous by the State prior to the review of their assessment system (Plake et al., 2004).

BIMM reviews prospective both on the theoretical and psychometric background for the tests (R. Spies, personal communication, March 12, 2007).

This training process for reviewers is framed around the review criteria defined by U.S. ED. Review criteria are a part of most peer review systems; however, the use of such criteria varies substantially across systems. For example, in the journal and conference proposal review processes, the review criteria are often the last part of the process as reviewers provide a thorough evaluation of the work (or proposed work) through comments and suggestions first, and then address the rating scale included as part of the process. Including review criteria in this process is critical so that states may clearly understand what is expected within each state assessment system and how the documentation must be prepared. Including specific criteria is also important to afford for a structured process that allows reviewers to focus on specific aspects of the assessment programs. Similar to the selection of raters, framing the review of assessment systems against compliance to specific criteria is also characteristic of the Nebraska assessment review model which includes six technical quality criteria (Plake et al., 2004).

The actual review is conducted in two stages, an independent review followed by a meeting of the peer review team to come to consensus on their ratings and recommendations. The independent review is characteristic of the peer review processes used in journals and conferences, as well as evaluation programs such as the Nebraska assessment review and the BIMM model. In contrast however, is the consensus process – this is used in the Nebraska assessment model but not in the others. In this particular case, this step in the process is beneficial to providing a cohesive recommendation to the U.S.

ED in terms of the strengths and weaknesses and a straightforward set of recommendations to the state for improvement of their assessment program.

After coming to group consensus, the peer review teams provide their combined ratings to the Assistant Secretary of Education who makes the final decision about the status of the assessment system. This is similar to the model used by conference and journal peer reviews – a conference program chair or journal editor would be the one to make a decision, considering the comments provided by the peer reviewers. On the other hand, the Nebraska review system also includes a decision making process that is based on a model that incorporates the ratings assigned to each of the six quality criteria (Plake et al., 2004). It must be noted though that each of the criteria is not weighted equally, the final decision is more heavily weighted by the ratings for alignment (of the assessment to the content standards) and opportunity to learn (alignment of curriculum to content standards). These two components were identified as being more critical to the validity of the intended uses of test scores as they are foundational sources of validity evidence for educational testing programs.

After a decision is made based on the peer review process, the results are reported to the state through a letter including the status of the state assessment system and any comments to the state regarding each peer review criteria not met. The unique aspect of this process relative to other peer review processes is the public nature of the distribution as these letters are posted on the U.S ED's website. This sharing of the results is likely due to the public nature of the assessment systems. The BIMM reviews are also made available through the *Mental Measurement Yearbook* series so that any potential users of the assessments have access to the results of this review and the final results of the

Nebraska review are published via the web and through local papers. Other reviews, such as conference proposals, journal submissions, independent audits, only provide results to the author or testing program – dissemination beyond these stakeholders is at their discretion.

### *Review Criteria*

The criteria for the NCLB peer review process are defined in documentation provided by U.S. ED (2004). This guidance includes specific criteria and examples of what is considered acceptable evidence for state assessment systems. Other peer review processes are guided by professional standards (e.g., AERA, APA, & NCME, 1999) and professional literature for assessment processes such as standard setting (e.g., Kane, 1994; 2001). In this section, we compare the review criteria mandated by U.S. ED to professionally accepted practices from the Nebraska assessment model and the *Standards*. When appropriate, relevant *Standards* are noted.

#### *Requirement 1 – Challenging academic content standards*

To meet this first requirement, a state must define foundational expectations for student learning as content standards that are rigorous and encourage the continual development of skills rather than minimum competency. States are expected to submit documentation of the process used to develop the standards, individuals involved in the process (e.g., content experts and/or individuals with experience writing standards), and use of external resources in development of the standards (e.g., nationally recognized content standards). As noted in the previous section, the peer review teams are not expected (as they do not have the contextual background) to evaluate the rigor of the

content standards. Rather they are looking for evidence that the content standards were developed through an acceptable process.

This aspect of the review reflects common practice of determining content standards (in education) or a set of knowledge skills and ability (in licensure/certification) that would be investigated through a psychometric audit. The *Standards* include requirements for the documentation of test frameworks (3.2), which specify the content of the test and the relative weighting of the content. The process by which these are developed (3.3) is a critical issue for validity to ensure that the test content is appropriate when making the intended inferences from the test results. The test content (i.e., content standards) must be clearly defined so the intended breadth and depth of the test will be as clear as possible during the subsequent phases of the test development process.

*Requirement 2 – Challenging Academic Achievement Standards*

As the focus of NCLB is all students meeting the state definition of proficiency, this second requirement is very important to ensure that ‘proficiency’ exudes a reasonable, yet meaningful goal for all students. These achievement standards must specify at least three achievement levels, include content expectations specified for each achievement level (alignment between achievement and content standards), and correspond to specific cut scores for the state assessments. States must provide detailed documentation of the process by which these achievement levels were set and that a diverse group of stakeholders were involved in the process.

This is closely tied with the sixth technical quality criteria from the Nebraska review model (Plake et al., 2004) as districts must provide evidence of setting appropriate mastery levels through a psychometrically sound and appropriate method. With respect to

determining cut scores and achievement standards, the *Standards* note the importance of documentation of the process by which cut scores are determined for any assessments (4.19). Setting cut scores should include selecting a representative and qualified panel of judges to make judgments about expectations for student performance based on their knowledge of the target population (4.21). The *Standards* recommend that external validity evidence be gathered to support the relationship between test performance and an external measure of similar constructs (4.20).

*Requirement 3: A single statewide system of annual high quality assessments*

To meet the third requirement, states are required to provide documentation of a single statewide system of assessments that includes all students, that is coherent across grade levels, and provides information relevant across subsequent years within a content area. To meet this requirement, states must provide documentation of the quality of the state assessments, the comparability across forms, the coherence of assessments across grades, comparability across assessments, coverage of the breadth and depth of the content areas, and inclusion of an alternate assessment.

The connections between this requirement and the *Standards* parallels the links established in other requirements including technical quality, alignment, and inclusion. The unique contribution of this requirement in the peer review process appears to be determining if the state is thinking about assessment as a comprehensive system across grades and across subpopulations and ensuring that each assessment is contributing meaningful information about what students know and can do related to the state content expectations. The *Standards* specify that a test user must assess the comparability across forms when multiple language versions of an assessment are included (9.9) and that when

testing individuals with disabilities (either through alternate forms or accommodations), test users must ensure that the intended constructs are being assessed (10.1). In turn, the unique contributions of this requirement are somewhat unclear. As is noted by the review and the state decision letters, the components within this requirement can be resolved with satisfaction of another requirement.

*Requirement 4: A system of assessments with high technical quality*

The focus of the fourth requirement is the technical quality including validity evidence and reliability evidence focusing on the intended uses of test scores and the decisions to be made from the results. The additional requirements of technical quality include ensuring the tests are fair and accessible to all students, appropriate criteria for accommodations, consistency across forms (over time, across modes), appropriate documentation of process for administration, scoring, analyzing test data, and reporting scores (quality control).

The Nebraska assessment model includes several criteria that are related to technical quality including appropriateness of the assessments to the student population (criterion 3), freedom from bias and sensitive situation (criterion 4), and consistency in scoring (criterion 5). The technical quality criterion is directly linked to many aspects of the *Standards*. The peer review guidance document specifically cites the validity and reliability components of the *Standards*. Technical quality is also an important criterion with other review systems such as that used for the *Mental Measurements Yearbook* reviews (BIMM, n.d.) as well as with conference proposal reviews (e.g., NCME). With respect to validity, the connections are apparent in that both documents reference validity in terms of the intended use of scores (1.1, 1.2) and the importance of providing evidence

of technical quality for a given purpose of test scores (13.2). Included within the validity specifications is the suggestion that test users must consider the unintended consequences of using an assessment (13.1). For reliability, the peer review guidance clearly ties into the *Standards* specifications that reliability be reported for each score (2.1), for each sub-population (2.11), with estimates of the standard error of measurement (2.2, 2.14), and inter-rater reliability (2.10). The *Standards* also include specific requirements for proper documentation of test administration (3.3), specifications for scoring (3.14, 3.21, 3.23), and procedures for interpreting test results (3.4).

*Requirement 5: Alignment of content standards, achievement standards, and assessments*

Under the fifth requirement, states must provide documentation that the content standards specified for their state are aligned with their achievement standards, as well as the assessments designed to measure these content standards. To meet this requirement, state are required to include evidence of a coherent approach to ensuring alignment, evidence that the standards and assessments are aligned (i.e. match to multiple dimensions including content and cognitive challenge), the assessment scores reflect a full range of achievement standards, the assessment results are expressed in terms of the achievement standards, and the state has in place ongoing processes to maintain and improve alignment.

The first requirement of the Nebraska assessment model specifies that a district must provide evidence of alignment of the assessment to the content standards. As noted earlier, this is one of the criteria that are weighted heavily in determining the overall rating of the assessment program given the relative importance to validity. In addition,

the Nebraska model includes a criterion assessing opportunity to learn (criterion 2) or alignment of the assessment and standards to the curriculum. Although the *Standards* do not specifically reference the term alignment, the intention of this requirement is very apparent in the overarching theme of validity noted throughout the *Standards* (1.6). Specifically, in the validity section, the *Standards* note “Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (p. 13). This requirement references a link between the assessment, the construct of interest (content standards), and the intended use of scores (achievement standards).

*Requirement 6: Inclusion of all students in the assessment system*

The sixth requirement mandates that a state assessment program includes all students through participation in the regular assessment, participation in the regular assessment with accommodations, or participation in an alternate assessment for either Students with Disabilities (SWD) or English Language Learners (ELL). This is a necessary requirement given the expectation that a state demonstrate 100% proficiency by the year 2014.

The *Standards* state the importance of considering one’s language abilities when administering a test designed for students with a particular language background. When determining proficiency in a construct other than language ability (e.g., mathematics) one must evaluate how the language of the test may influence a student’s ability estimate as this is a critical issue to validity (9.1). Similarly, when students with disabilities are tested, it is important to ensure that the test results reflect the intended construct and are not influenced by construct-irrelevant variance that results from a student’s disability

(10.1). If an accommodation is to be offered and used, those involved in making the decision about the use of accommodations must rely on the research about the relationship between the disability and test performance and how particular accommodations can mitigate that relationship (10.2).

*Requirement 7: An effective system of assessment reports*

The seventh requirement, an effective system of assessment reports, requires states to provide documentation of their comprehensive reporting system. Specifically, states must provide reports of assessment results that include interpretive guides and meaningful sub-domain scores.

Throughout the *Standards* there are specific requirements for reporting results of assessments. Related to the focus of validity, the *Standards* specify the need to provide information for proper interpretation of score results in any score report (5.10). Score reports must also include cautionary notes about interpretation of sub-groups scores when they are provided (7.8). Similar to the peer review guidelines, the *Standards* specify a testing program must maintain confidentiality of individual level data (8.5, 8.6).

*Results of the Review process*

The state decision letters posted on the U.S. ED website were used as evidence of the types of feedback provided to states (U.S. ED, n.d.). The letters to states include the final overall evaluation of the assessment model in addition to feedback pertaining to any of the seven criteria not met by the state assessment program. To date, two states have received 4 successive letters, 15 states have received three letters, 17 states have received two letters, and 18 states have received one letter. Currently, 17 states have received *Full Approval* of their state assessment systems (nine with recommendations), 4 states have

received an evaluation of *Approval Expected*, and 29 are categorized as *Approval Pending*.

The feedback provided to states suggests similarities between this model and those used for conference submissions and journal publications. These models include an overall decision that is based (largely) on the feedback from the reviewers and includes specific feedback and recommendations from the reviewers. Some letters do occasionally reference other communication with state education agencies that occurred prior to the drafting of the letter. This is in contrast to the other assessment review models except for possibly a psychometric audit where an accrediting agency may contact those responsible for a testing program to clarify any unknowns.

As noted in the first part of the results section, the ultimate decision about a state assessment program is at the discretion of the Assistant Secretary of Education. Although similar to the conference and journal review model, this is in contrast to the Nebraska review model whereby the rating of each quality criteria is used to determine the overall rating. From these models, it is unclear exactly how the overall rating is determined. From the previously listed definitions, *Approval Expected* indicates that all requirements are met but certain processes may not be complete, *Approval Pending* indicates that one or a few components are missing or do not meet the requirements, and *Non-Approved* has many that are missing or do not meet the requirements. From a review of the decision letters, it is unclear how the distinction is made between *Approval Pending* and *Approval Expected* as one can identify states with either classification that still must propose solutions to problems and substantial information missing from a few (e.g., 2) or many (e.g., 6) requirements that are still not met. Although there may be decision rules for

determining such decisions, it is not apparent in our review of the publicly-available documentation of the process. Clear delineation of the expectations for these different levels of performance would improve the transparency of the process.

### *Discussion*

Through this review, we were able to compare the NCLB peer review process as is described through documentation provided by U.S. ED. Several unique aspects of the NCLB peer review process were revealed that provide strength to this very important system of review. In addition, one can learn unique aspects of other peer review systems that are important aspects to support the validity of each of these programs.

Within the design of the system, a key strength of this process is the framework provided by the seven review criteria. This allows reviewers to focus their review and address specific questions about each assessment program and standardizes the process across states. In addition, it allows states to prepare their materials in an organized fashion that will help facilitate the review. A second unique aspect is the use of a peer review team to create a consensus set of recommendations which are likely very helpful to U.S. ED in making their final decision and provides coherent guidance to states in taking the appropriate next steps to meeting the NCLB criteria.

Looking at the specific review criteria, there was substantial overlap noted between the NCLB peer review process and that used with the Nebraska assessment review. The NCLB review also includes unique aspects such as reporting, inclusion of all students, and determination of challenging standards. The Nebraska assessment model also includes a unique review component – that of opportunity to learn or documenting

the alignment between the curriculum and the content standards. This is reiterated in the *Standards* which emphasize the importance of students having the opportunity to learn in the context of educational testing (13.5). Although many states would consider the content standards (requirement 1) to be reflective of the state or local curriculum (as these may have been developed at the same time), to be fair to students one must show that all students have the opportunity to obtain the knowledge and learn the skills they are to be held accountable for. In some states, curriculum experts work independently from assessment experts. Although there may be some cooperation between these two areas, their expectations for each grade level may not parallel one another.

Finally, within the reporting of the results, the reporting format paralleled that used for journal reviews and conference proposals in terms of an overall decision and supporting evidence from reviewers. In addition, the NCLB model parallels these in terms of a somewhat subjective decision model that is support (but not solely based on) specific review requirements.

Given the importance of the NCLB peer review process to the credibility of the decisions it is critical that independent professionals in the testing community provide critical analysis of this system and the way such reviews are carried out and the results are reported to states. This study provides one examination of this process from the perspective of the academic peer review process. We have attempted to highlight the similarities and areas of difference between the NCLB peer review system and other academic peer review process and the significance of these similarities and differences. Research such as this will hopefully encourage others to take their own critical look at processes such as these. The testing community has a responsibility to ensure such

systems are appropriately evaluating assessment systems for validity evidence that is consistent with our professional standards.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Brennan, R.L.(Ed.) (2006). *Educational Measurement, 4<sup>th</sup> Edition*. Westport, CT: Praeger Publishers.
- Buckendahl, C. W., Plake, B.S., & Impara, J.I. (2004). A strategy for evaluating district developed assessments for state accountability. *Educational Measurement: Issues and Practice, 23(2)*, 17-25.
- Buros Institute of Mental Measurements (n.d.). *Reviewers guide for the mental measurements yearbook*. Retrieved August 7<sup>th</sup>, 2006 from [www.unl.edu/buros/bimm/html/suggestions.html](http://www.unl.edu/buros/bimm/html/suggestions.html).
- Forte, E. (2006). A status report on the NCLB standards and assessment peer reviews. *NCME Newsletter*, June 2006.
- Harris, D., Gao, X., & Welch, C. (2002, April). *An analysis of NCME annual meeting proposal reviewer ratings*. Paper presented at the annual conference of the National Council on Measurement in Education, New Orleans, LA.
- Jefferson, T., Wager, E., & Davidoff, F. (2002). Measuring the quality of editorial peer review. *Journal of the American Medical Association, 287(21)*, 2786-2790.
- Johnson, H. (2006a). Letter to chief state school officers regarding possible outcomes and consequences of the peer review process. Retrieved September 1, 2006 from [www.ed.gov/admins/lead/account/saapr3.doc](http://www.ed.gov/admins/lead/account/saapr3.doc)

- Johnson, H. (2006b) Letter to chief state school officers on department priorities and future actions. Retrieved March 9, 2007 from [www.ed.gov/policy/elsec/guid/stateletter/index/html](http://www.ed.gov/policy/elsec/guid/stateletter/index/html).
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425 – 461.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 53 – 88). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Kassirer, J.P., & Campion, M.D. (1994). *Peer review: Crude and understudied, but indispensable. Journal of the American Medical Association*, 272, 96-97.
- Linn, R.L. (2005). Fixing the NCLB Accountability System. CRESST Policy Brief #8.
- Mathis, W.J. (2003). No child left behind: Costs and Benefits. *Phi Delta Kappan*, 84(9), 679-686.
- Plake, B.S., Impara, J.I., & Buckendahl, C.W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice*, 23(2), 1215.
- Rothwell, P.M., & Martyn, C.N. (2000). Reproducibility of peer review in clinical neuroscience: Is agreement between reviews any greater than would be expected by chance alone? *Brain*, 123, 1964-1969.
- Spies, R., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

U.S. Congress (2002). *Public Law 107-110: No Child Left Behind Act of 2001*. Accessed May 15, 2006 from <http://www.ed.gov/policy/elsec/leg/esea02/index.html>

U.S. ED (2004). *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the Not Child Left Behind Act of 2001*. Accessed April 28, 2006 from <http://www.ed.gov/admins/lead/account/saa.html#examples>

U.S. ED (n.d.). *Decision Letters on Each State's Final Assessment System Under No Child Left Behind (NCLB)*. Accessed April 28, 2006 from <http://www.ed.gov/admins/lead/account/nclbfinalassess/index.html>

Weber, E.J., Katz, P.P., Waeckerle, J.F., & Callahan, M.L. (2002). Author perception of peer review: Impact of review quality and acceptance on satisfaction. *Journal of the American Medical Association*, 287(21), 2790-2793.