Report to

Nebraska Department of Education

# Evaluation of the Six Quality Assessment Criteria used in the Nebraska School-based Teacher-led Assessment and Reporting System (STARS)

Prepared by

Suzanne Lane
Professor of Educational Research Methodology
University of Pittsburgh

September 2006

## Introduction

The Nebraska Department of Education pioneered a unique assessment and accountability system that allows for local control of how districts assess student performance on state content standards (Roschewski, 2004). Districts select their assessments which may include norm-referenced tests, criterion-referenced tests, and/or locally developed classroom assessments that involve teachers in the design, administration and scoring of the assessments. The Nebraska School-based Teacher-led Assessment and Reporting System (STARS) provides the opportunity for assessment to play an integral role in teaching and learning. Moreover, the assessment system provides valuable professional development assessment activities for teachers that emphasize the importance of coherency among content standards, instruction and assessment.

For both mathematics and reading, each district submits to the Nebraska Department of Education a District Assessment Portfolio that documents the technical quality of their assessment system. The Portfolio includes information about each of the assessments used for the reported grade levels and information on how the district assessments meet Six Quality Assessment Criteria that are used in the evaluation of the District Assessment Portfolios. The Six Quality Assessment Criteria are (Nebraska Department of Education, 2005a):

1. The assessments reflect the state/local standards.
2. Students have the opportunity to learn.
3. The assessments are free of bias and insensitive situations.
4. The assessments are at the appropriate level.
5. The assessments are reliably scored.
6. The assessments mastery levels are appropriately set.

The purpose of this report is to provide an evaluation of the Six Quality Assessment Criteria and to provide recommendations to the Nebraska Department of Education regarding the criteria. This evaluation is timely in that the No Child Left Behind (NCLB) Act of 2001 requires states to test students in Grades 3 through 8 in mathematics and reading/language arts starting no later than the 2005-06 school year. NCLB requires each state to adopt challenging content standards and challenging achievement (performance) standards. States must also establish adequate yearly progress (AYP) goals for each year from 2002 to 2014 with all students at or above the proficient achievement standards by 2014. Under NCLB each state develops its own content standards, chooses its own assessments, and sets its own performance standards. A worthy feature of NCLB is its focus on groups of low achieving students. It requires separate reporting of results for economically disadvantaged students, students with disabilities, limited English proficient students, and for groups of different race/ ethnicity. The disaggregated reporting of results allows for monitoring the achievement of different subgroups and monitoring the achievement gap of low achieving students over time.

Typically large-scale assessments serve distinctly different purposes than classroom assessments in that large-scale assessments monitor achievement trends over time and are

used to evaluate educational programs. In Nebraska, however, classroom assessments are used for these purposes. Most large-scale assessments must meet stringent standards for technical quality because of the consequences associated with the purposes they are intended to serve. Therefore, it is imperative to ensure the validity and technical quality of the Nebraska assessment system for these purposes. As Standard 13.2 states in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), "In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose" (p. 145). The Six Quality Assessment Criteria were developed to serve this role for STARS.

Nebraska has developed a portfolio process that helps ensure that local assessments meet the technical standards required by the NCLB mandate. In this process, teachers and administrators are involved in collecting evidence to demonstrate that the procedures used to develop, score and set performance for their assessments are of high technical quality. One study that has examined the quality of local district mathematics assessments used in STARS (Brookhart, 2005) reported that the quality was "generally good" (p. 21) and the mathematics assessments that were evaluated in the study were of "sufficient alignment, clarity, and appropriateness to warrant attention to their results" (p. 20).

## Overview of the Evaluation Process

The Six Quality Assessment Criteria were evaluated in terms of their role in ensuring the validity and technical quality of the Nebraska School-based Teacher-led Assessment and Reporting System within the context of the *No Child Left Behind Act* (NCLB) of 2001. As described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) validity refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Therefore, the validation of an assessment system requires a clear statement of the proposed interpretations and uses, and involves the accumulation of evidence to support the proposed test score interpretations and uses.

In the *School-based Teacher-led Assessment and Reporting System: A Summary* (September 2004) it states that "the two key priorities are to "improve educational opportunities" and "improve learning"" (p. 1). Further, because STARS is used for assessment and accountability purposes with NCLB, an evaluation of the Six Quality Assessment Criteria needs to consider the purposes and requirements of NCLB.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) provides the framework for evaluating the Six Quality Assessment Criteria. Although the *Standards* are geared for large-scale assessments, they are relevant in the evaluation of STARS because STARS is used for high-stakes accountability purposes under the NCLB Act. Further, as indicated by Plake, Impara, and Buckendahl (2004) the Criteria were identified by the Nebraska Department of Education to be congruent with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

The standards that are most relevant to the purpose and uses of STARS were identified and the criteria were evaluated against those identified standards. Other standards developed for assessment and accountability systems were also considered in the evaluation, such as Linn, Baker, and Dunbar's (1991) validation criteria specifically geared to performance assessments and the *Standards for Educational Accountability Systems* (2002) developed by the National Center for Research on Evaluation, Standards, and Student Learning. Two important areas that the *Standards for Educational Accountability Systems* (2002) address is the need for validity evidence for assessments of students with different language backgrounds and students with disabilities.

In the evaluation of the Criteria, sources of validity evidence that were considered include evidence based on test content (e.g., content sampling including representation and sufficiency; alignment to standards; coherency among tasks and scoring rubrics; item quality; content quality), evidence based on internal structure and response processes, evidence based on relations to other variables, and evidence based on consequences and impacts of testing, including instructional consequences. The Criteria will also be evaluated in terms of addressing the reliability of scores derived from assessments, including score consistency over tasks and raters as well as decision consistency in the assignment of students to performance standards. Issues related to the comparability of scores derived from the assessment portfolios were also addressed. This is of particular importance for the purposes of NCLB.

The extent to which the Criteria address the fairness of assessments was evaluated, including issues related to opportunity to learn and potential differential validity evidence for subgroups of students such as students with disabilities, students with diverse language backgrounds, and students with low socio-economic backgrounds. The Criteria were also evaluated in terms of addressing the quality of procedures for setting achievement levels, quality of scoring, and quality of score reporting. Lastly, the Criteria were evaluated in terms of the applicability to both large-scale assessments and classroom-based assessments.

The Quality Criteria Rating Scale (Nebraska Department of Education, 2005b; Plake, Impara, & Buckendahl, 2004) was also reviewed. Currently, the Rating Scale has 5 levels, Exemplary, Very Good, Good, Needs Improvement, and Unacceptable, and each grade level portfolio from a district receives one of the five ratings. The differential rating system in the Quality Criteria Rating Scale for the Six Quality Assessment Criteria was considered.

## Review of Documents

A number of documents published by the Nebraska Department of Education as well as reports from external evaluators were reviewed. The documents that were consulted for this evaluation include:

- *School-based Teacher-led Assessment and Reporting System: A Summary* (September 2004)
  - *District Assessment Portfolio Rubric For Use in 2003-04, 2004-2005, 2005-06*
  - *Quality Criteria Rating Chart For the 2004-05 District Assessment Portfolio* (Effective until 2006-07)
- *A Guide for Assuring the Technical Quality of Classroom Assessment* (March 2006)
- *District Assessment Portfolio Rubric Effective Beginning 2006-2007*
- *District Portfolio Assessment Rating Form*
- *Guide to Addressing the Quality Components in the District Assessment Plans*
- *Evaluation Report on the Nebraska Sate Department of Education's District Assessment Portfolio Training Process (October 2004)* by Ellen Forte Fast
- *Examining the Potential for Selected NRTs and Locally Developed CRTs to Classify Students into Performance Categories in Reading and Mathematics (May 2004*) by James C. Impara et al.

## Site Visits

To obtain additional evidence of the use and feasibility of the Six Quality Assessment Criteria, the research investigator met with personnel at the Nebraska Department of Education and teachers and administrators at schools in Nebraska as well as attended a District Assessment Portfolio Training Session. These meetings provided additional evidence for evaluating the quality and feasibility of the Criteria and their implementation at the local level.

Two visits were made to Nebraska by the investigator. In March 2006, 3 school districts were visited including Plattsmouth, Nevada City and Lincoln. This provided an opportunity for informal discussions with Nebraska educators and administrators regarding STARS. Discussions centered on the alignment of their curriculum to the Nebraska content standards, incorporating assessments throughout instructional units, ensuring continuity across grades, the use of performance assessments to assess problem solving and reasoning, the use of scoring rubrics for evaluating student work, and the collection of validity and reliability evidence for STARS.

The teachers reported that the focus on classroom assessment informed their instruction and provided them with more accurate information about student understanding and learning. Teachers indicated that students have a better understanding of the criteria by which their work is assessed. The term, *transparency*, has been used by Frederiksen and Collins (1989) to express this idea of providing students the opportunity to understand and internalize the criteria used for evaluating their work. This in turn helps students develop the skills and awareness of what needs to be attended to in order to perform well. Student understanding of the criteria is not just learning the rules of how to get a good grade, but more importantly, "it means learning the discipline itself" (p. 298, Shepard et al., 2005). Nebraska teachers also reported that they spent many more hours on assessment activities since the beginning of STARS. In general, they reported

that the increased time on assessment allowed them to better formulate curriculum goals and to obtain a better understanding of student proficiency. School administrators also indicated that STARS allowed for careful evaluation of individual student achievement and progress, and teachers were gaining valuable experience in assessing their students in a meaningful way. Although most of the comments were very positive, concerns were raised about the amount of work STARS entailed during certain times in the instructional year.

The second visit was for a District Assessment Portfolio Training Session. The training was conducted by Sue Anderson from the Nebraska Department of Education and Gregg Schraw, a Professor at the University of Nevada at Las Vegas. The session involved approximately 42 participants, with approximately 6 national assessment experts and 36 Nebraskan educators and administrators that were involved in the development of their district's portfolios. All of the national assessment experts and a number of the Nebraskan educators and administrators had participated in previous Portfolio Training sessions. Prior to the training, Dr. Pat Roschewski, Director of Statewide Assessment for Nebraska, provided an overview of Nebraska's Assessment and Accountability System. She also discussed the Assessment Quality Review process that focuses on the Six Quality Criteria, Assessing the Assessments, and Strategies for Improvement.

After the overview of STARS on the first morning, Sue Anderson and Gregg Schraw conducted the training session. First, a review of the *District Assessment Portfolio Rubric* was given and then the participants were trained on the Six Quality Criteria. Anchor papers were used in the training to help clarify the criteria. The second day of training included independent ratings of portfolios. For each of three portfolios, the portfolio was rated independently. This was followed by a table discussion and consensus of the ratings, and then a large group discussion of the table ratings.

The training on the Six Quality Criteria was shared by Sue Anderson and Gregg Schraw, and their expertise was complementary in that Sue focused on the less technical Criteria (1-4) and Gregg focused on the more technical Criteria (5-6). Further, Sue Anderson's years of involvement working on STARS with Nebraskan educators and administrators and Gregg Schraw's ability to discuss technical issues in a way that was meaningful to individuals with varied technical backgrounds was beneficial in creating an atmosphere of trust and competency. The training provided ample opportunity for discussion and questions from the participants. The discussion and questions indicated that the participants were very conscience in applying their ratings and were knowledgeable about the criteria. Some district educators and administrators, however, expressed some difficulty in working with the last two criteria - assessments are reliably scored and assessment mastery levels are appropriately set.

<u>Recommendations</u>:

1. Continue to have both an individual who is from the state department as a trainer as well as an external measurement specialist. The external measurement specialist is

essential in ensuring that the participants have a good understanding of the technical criteria, while the internal specialist is essential in ensuring that the participants have a good understanding of the STARS system as a whole as well as the less technical criteria.

2. Ensure that enough time is allocated to the two technical Criteria (assessments are reliably scored and assessment mastery levels are appropriately set). Because these two criteria are technical, it is important to allocate enough time on discussing these criteria to ensure accuracy in rating the portfolios. A thorough review of the material related to Criteria 5 and 6 in *A Guide for Assuring the Technical Quality of Classroom Assessment* (March 2006) would be beneficial to the participants of the training session.

## Evaluation of the Quality Criteria and Recommendations

The Six Quality Assessment Criteria are reflective of good practices in educational testing and assessment. The Criteria and accompanying documents are written in a way that school administrators and educators can understand and apply to their district assessment systems. Each of the Quality Criteria is evaluated separately and recommendations are made for each of the Criteria. To satisfy each Quality Criteria, the *District Assessment Portfolio Rubric (Effective Beginning 2006-2007)* identifies a number of requirements. The requirements are provided below for each of the Six Quality Assessment Criteria.

The requirements that are in bold are new additions to the Criteria for the 2006-2007. It should be noted that the new additions in 2006-2007 stipulate that the Criteria apply to all standards and there should be coherency or consistency across Criteria.

### Quality Criteria 1: The assessments reflect the state/local standards

- Qualifications of the independent reviewers are clear and complete.
- Evidence of an independent review for match to standards is clear and complete (reviewers did not write the assessments.)
- The process for matching assessments to standards is clear and complete.
- Results of the matching process are clear and complete.
- Sufficiency process is clear and complete.
- Sufficiency results are clear and complete (sufficiency required for both number of items/performances **and levels of difficulty. Minimum 12 items or equivalent on reading standards 4.1.3, 8.1.1 and 12.1.1 and math standards 4.2.1, 8.2.2, and 12.2.1)**
- **Consistency between Criterion #1 and other criteria is clear.**
  **Matching or sufficiency is provided for all standards.**

Quality Criteria 1 reflects Standard 13.3 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), "When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided" (p. 145), and Standard 1.8, "If the rationale for a test use or score interpretation depends on

premises about the psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided" (p. 19).

Quality Criteria 1 contains 2 important features in evaluating assessments: Alignment and Sufficiency. Alignment refers to the degree to which the assessments match the standards and sufficiency refers to the degree to which the assessments are appropriate for students at different performance levels. Criteria 1 reflects Standard 13.1 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) states, "When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domains should be provided" (p. 145).

The sufficiency aspect of the Criteria addresses the need for an assessment to target students of differing proficiency levels and students across the range within a proficiency level. Items not only need to be of different difficulty levels, but they also need to assess different levels of understanding of the material. As a simple example, an item may be very difficult because it requires students to recall information that was presented early in the instructional unit rather than because it requires a deep understanding of the concept. It is important to have teachers document the extent to which the assessment can tap varying levels of proficiency by examining the cognitive demands of the assessment as well as their difficulty levels. In *A Guide for Assuring the Technical Quality of Classroom Assessment* sufficiency is discussed in terms of whether there are enough items so that students at all levels can demonstrate their skills. Additional guidelines that clearly address how teachers can demonstrate that an assessment has sufficiency would be valuable. Such guidelines may enable teachers to better assess students at various levels and help alleviate a concern expressed by Brookhart (2005). In Brookhart's study, she examined a sample of mathematics assessments used by teachers in STARS and concluded that the assessments were well targeted at the middle of the range of student performance, but were less able to engage students at the ends of the distribution (Brookhart, 2005).

The requirement that independent reviewers (i.e., not the assessment writers or developers) are needed to examine the alignment between the state content standards (i.e., Nebraska L.E.A.R.N.S. - Leading Educational Achievement through Rigorous Nebraska Standards) and the local assessments is noteworthy. As Standard 1.7 indicates "When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be full described" (p. 19, AREA, APA, & NCME, 1999). It is clearly stated in *A Guide for Assuring the Technical Quality of Classroom Assessment* that the reviewers of the sufficiency criteria need to be independent, but it is not clear in the *District Assessment Portfolio Rubric* whether the reviewers of the sufficiency criteria need to be independent. Consistency across the two documents would help ensure that the districts follow the correct procedures.

Recommendations:

1. Quality Criteria 1 could be divided into two criteria, one that focuses on the alignment of the assessments to state standards and another that focuses on sufficiency, that is, the extent to which the assessment can adequately measure students at different performance levels. Additional guidelines that clearly address how teachers can demonstrate that an assessment has sufficiency would be valuable. Such guidelines would address both the difficulty level of the items as well as the cognitive demands of the items.

2. It should be made clear in both *A Guide for Assuring the Technical Quality of Classroom Assessment* and the *District Assessment Portfolio Rubric* that alignment and sufficiency should be met with respect to the State content standards and if districts have local standards, these local standards should be mapped onto the state standards.

3. To be consistent with *A Guide for Assuring the Technical Quality of Classroom Assessment* clearly indicate in the *District Assessment Portfolio Rubric* that the reviewers of the sufficiency criteria need to be independent of those that wrote or developed the assessments.

**Quality Criteria 2: Students have an opportunity to learn.**

- Qualifications of the opportunity to learn reviewers are clear and complete.
- The process of opportunity to learn is described and is clear and complete (both curriculum alignment and timing of assessment/instruction).
- The results of the process for alignment of standards with local curriculum are clear and complete.
- Dates are provided when standards are taught and they are clear and complete.
- Dates are provided when standards are assessed and are clear and complete (80% of instruction should take place prior to assessment.)
- **Consistency between Criterion #2 and other criteria is clear and complete. Opportunity to learn information provided for all standards.**

The accumulation of validity evidence for assessments needs to consider whether "teachers and schools have the capability and do provide all students with the opportunity to learn what is assessed" (Herman & Klein, 1996, p. 246). In describing the fairness of assessments, the *Standards* (AERA, APA, NCME, 1999) indicate that one view of fairness is opportunity to learn. Multiple modes of assessment throughout the instructional process help ensure equal opportunity to learn, and STARS encourages multiple modes of assessment. Nebraska's effort in ensuring that students have an opportunity to learn the standards is noteworthy. It is one of the most conscience state efforts in collecting opportunity to learn data. Quality Criteria 2 addresses one of the major purposes of STARS, to "improve educational opportunities" (p. 1), as stated in the *School-based Teacher-led Assessment and Reporting System: A Summary* (September 2004).

In *A Guide for Assuring the Technical Quality of Classroom Assessment* it states that Quality Criteria 2 ensures that the standards are present in the curriculum and that 80% or more of the content is taught to students prior to being assessed on it. Teachers who teach the local curriculum form the panel that is responsible for examining the local curriculum material and identifying which standards are taught and at what time during the year. More importantly, this panel of teachers are responsible for developing a plan and a timeline for addressing any needed changes in opportunity to learn. This level of teacher involvement in evaluating the extent to which students have the opportunity to learn the standards provides a much needed link between standards, instruction and assessment.

Lastly, there is a clear alignment between *A Guide for Assuring the Technical Quality of Classroom Assessment* and *District Assessment Portfolio Rubric* for Quality Criteria 2, Opportunity to Learn.

Recommendation:

1. It would be useful to have districts provide an example of how a standard met the opportunity to learn criteria. This may include a description of the curriculum and instruction that was implemented for a particular standard and its relation to the assessment tasks that are used to measure the standard.

**Quality Criteria 3: The assessments are free of bias and sensitive situations.**

- Qualifications of the bias reviewers are clear and complete.
- The description of the bias orientation process is clear and complete.
- The process for bias review of assessment items is clear and complete.
- Results of a bias review are clear and compete.
- **Consistency between criterion #3 and other criteria is clear and complete.**
- **Bias information provided for all standards.**

Assessments need to be responsive to differences in students' experiences and culture (AERA, APA, NCME, 1999). Criteria 3 addresses the need to ensure that aspects of test design, content and format that may lead to bias are not present in the assessment. As Standard 7.4 indicates "Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups…" (p. 82, AERA, APA, NCME, 1999).

Nebraska's effort in collecting this type of information is noteworthy. *A Guide for Assuring the Technical Quality of Classroom Assessment* indicates that qualified leaders should conduct training in assessment bias for those who will be reviewing for bias and it describes the type of training that should be conducted for the bias review. Nebraska's efforts to collect data on the quality of the training process and the results of the bias review help ensure that district educators and administrators are sensitive to ensuring that all students are assessed fairly.

Recommendations:

1. To be consistent with the *District Assessment Portfolio Rubric, A Guide for Assuring the Technical Quality of Classroom Assessment* should indicate that the qualifications of the bias review panel need to be documented. The *Guide* may also describe the qualifications that would be desired.

2. Additional evidence can be collected for Quality Criteria 3 as outlined in a later section in the document, Recommendations for Additional Criteria. The collection of some of this evidence will be dependent on sample size.

**Quality Criteria 4: The assessments are at the appropriate level.**

- Qualifications of the reviewers for appropriate level are clear and complete.
- Process for appropriate level review is clear and complete.
- Results of the appropriate level review are clear and complete.
- **Consistency between Criterion #4 and other criteria is clear and complete. Appropriate level information is provided for all students.**

Quality Criteria 4 helps ensure that the cognitive level of the assessment is appropriate for the grade level being assessed. Quality Criteria 4 as well as Quality Criteria 1 reflect Standard 1.8 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), "If the rationale for a test use or score interpretation depends on premises about the psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided" (p. 19).

*A Guide for Assuring the Technical Quality of Classroom Assessment* indicates that a panel of educators who are familiar with the grade level and content are responsible for reviewing the assessments to determine if they are at the appropriate level. The *Guide* further suggests that the panel should consist of educators at the grade level for which the assessment is targeted in addition to educators from grade levels surrounding the targeted grade level. It may be useful to further indicate that educators serve on panels across grade levels to evaluate the developmental level of assessments across grades. This would help ensure that there is a developmental progression being reflected in the assessments across grades. The recommendation in the *Guide* to include special education teachers and a school psychologist as members of the panel will help ensure that the needs of various subgroups are being considered in the evaluation.

Recommendations:

1. Members of a grade level panel could serve on adjacent grade level panels so that panel members can evaluate the developmental progression reflected in the assessments across grades.

2. If a readability analysis is required as indicated in the *District Assessment Portfolio Rubric*, then *A Guide for Assuring the Technical Quality of Classroom Assessment* should also address the need for a readability analysis.

**Quality Criteria 5: There is consistency of scoring.**

- Qualifications of the reliability process participants are clear and complete
- Appropriate process for reliability is clear and complete.
- Reliability value provided and calculations are at or above the minimum acceptable level (.70).
- Procedure for improving reliability is clear and complete.
- **Consistency between Criterion #5 and other criteria is clear and complete. Reliability is reported for all standards.**

Assessments that contain both selected response items and constructed response items require the examination of both score reliability and rater reliability. Both types of reliability are addressed under Criteria 5 and explained more fully in *A Guide for Assuring the Technical Quality of Classroom Assessment.* Standard 2.1 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) states, "For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measured or test information functions should be reported" (p. 31). The Nebraska Department of Education may also want to consider requiring districts to report the standard error of measurement for their assessments. This would remind district educators and administrators that test scores are not precise estimates of student achievement but contain some degree of error.

The consistency and accuracy with which test scores classify examinees into performance levels is also important to address. The consistency and accuracy of examinee classifications is of critical importance because of the reporting of test performance in performance categories as well as the reporting requirements of *No Child Left Behind.* The *Guide* divides score reliability according to Internal Consistency and Decision Consistency methods, and rater reliability is the third category. This provides a coherent framework for teachers, allowing them to choose the correct reliability procedures for their assessments.

The internal consistency methods that are recommended include KR20, KR21, coefficient alpha, and split half reliability. The *Guide* indicates that these methods are appropriate for objectively scored tests. It may be useful to indicate that KR20 and KR21 are appropriate for items that are dichotomously scored, that is, they are considered either correct (1) or incorrect (0) as in multiple-choice items. In contrast to KR20, coefficient alpha can be used when items have two or more score levels (e.g., 0, 1, 2, and 3). Therefore, coefficient alpha is appropriate for tests that contain constructed response items that have more than two score levels. It may also be useful to provide the equation for KR21 in the *Guide* and explain how to obtain this index given that the KR21 formula uses test statistics only and can be easily computed by classroom teachers.

The decision consistency method proposed in the *Guide* requires two independent decisions about student performance. The percentage of times the decisions agree is used

as the reliability/consistency coefficient. The *Guide* specifies that the percentage of agreement must meet or exceed .70 to be considered acceptable. The decision consistency methods are recommended for small districts.

The *Guide* suggests three types of decision consistency methods: 1) the use of two assessments measuring the same thing at the same level of difficulty (i.e., alternate forms, test-retest), 2) the use of a criterion-referenced test and a norm-referenced test that measure the standards, and 3) the use of teacher judgment and assessment results which is labeled, Teacher Judgment Decision Consistency. The *Guide* outlines clearly the procedures for obtaining the Teacher Judgment Decision Consistency. In this procedure, teachers first need to have a shared understanding of the performance level descriptors and then they make judgments about the performance level they believe each of their students will achieve without knowledge of their students performance on the assessment. These judgments are then compared to the proficiency level the students obtained. This method provides evidence of the level of agreement between the assessment results and teachers' professional judgment of student proficiency, and provides some validity evidence for the performance standards that are set by the districts as discussed later in this document.

The *Guide* outlines clearly the procedures to obtain inter-rater reliability using percent agreement between raters. It also indicates that exact agreement is required for assessments with fewer than 6 score levels.

In summary, this set of reliability procedures allow for districts to choose a procedure that is most appropriate for their assessments and the size of their student population. The *Guide* is well-written and provides enough information so school districts can easily obtain information on the reliability of their assessment results. For some districts, especially for small districts, the information obtained about the reliability of the scores will be limited.

<u>Recommendations</u>:

1. *A Guide for Assuring the Technical Quality of Classroom Assessment* should indicate that an internal consistency method (i.e., coefficient alpha) can be used for tests containing constructed response items that have more than two score levels.

2. In addition to requiring interrater reliability for subjectively scored assessments, a measure of score reliability would also strengthen the portfolio.

3. The *District Assessment Portfolio Rubric* includes a number of features that are not provided in *A Guide for Assuring the Technical Quality of Classroom Assessment* such as information on how the scoring rubric was pre-tested and the need for a plan to improve reliability. It would be useful to have these two documents consistent.

**Quality Criteria 6: The mastery levels are appropriately set.**

- Qualifications for mastery level participants are clear or complete.
- Evidence of mastery level process is clear or complete.
- Results of the mastery level process are clear and complete.
- **Consistency between criterion #6 and other criterion is clear an complete. Mastery level information is provided for all standards.**

Standard 4.19 states in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), "When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented" (p. 59), and Standard 4.20 states, "When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria" (p. 60). Quality Criteria 6 helps ensure that that proficiency levels are appropriately set and are in compliance with the *Standards*.

*A Guide for Assuring the Technical Quality of Classroom Assessment* provides step-by-step procedures for establishing performance level definitions using either a student-based method (i.e., modified contrasting group method) or a test-based method (i.e., modified Angoff method or modified analytical judgment). The description of each of these methods is clear and should allow for easy implementation of the methods for school districts.

To help support the appropriateness of the procedures used various types of validity evidence can be collected. Kane (2001) suggests that standard setting procedures should include the evaluation of procedural evidence, internal consistency evidence and external evidence. Procedural evidence helps support the appropriateness of the procedures used and the quality of the implementation of the procedures (e.g., definition of goals for the decision procedure, selection and training of panel members, definition of performance standards, data collection procedures). Internal consistency evidence uses data obtained within the standard-setting study to provide a partial check on the validity of the results (e.g., obtain the standard error of passing score, that is, the extent to which the same passing score would be obtained if the study were repeated, survey the panelists about the standard setting process and their level of confidence). External evidence compares the results of decision made using the passing score to the results of the same kind of decision made in a different way (e.g., Compare the results from two different standard setting procedures using the same test; compare the passing score using the results from a different test; compare the passing score with independent ratings of student accomplishments such as judgments by teachers).

The requirement that school districts to supply procedural evidence for the standard setting method that they use is indicated in the *District Portfolio Assessment Rating Form*. For example, the *Rating Form* suggests the need for a description of the qualifications of the panel, a description of a method that considers the difficulty of the assessment, and the results for each assessment in the district. Currently, internal consistency evidence is not required for the evaluation of the standard setting procedures. It may be useful to ask districts to survey their panelists about the standard setting

process and their level of confidence in the final mastery levels. This would provide some internal consistency evidence for the standard setting process. With respect to external evidence, if districts choose the Teacher Judgment Decision Consistency method under Criteria 5 external evidence is provided. That is, teachers make judgments about the performance level they believe each of their students will achieve and these judgments are then compared to the proficiency level the students obtained. This method provides evidence of the level of agreement between the decisions made using the passing score and teachers' professional judgment of student proficiency.

Lastly, there is a clear alignment between *A Guide for Assuring the Technical Quality of Classroom Assessment* and *District Assessment Portfolio Rubric* for Quality Criteria 6.

Recommendation:

1. Quality Criteria 6 does address procedural evidence but should also address internal evidence and external evidence for the standard setting method used by the school districts.

**Recommendations for Additional Criteria**

The evidence documented through the Six Quality Assessment Criteria provides a wealth of information about the quality of the district assessment system. This section outlines some additional information that can be collected to expand on the information that is already being used in evaluating district assessment systems within STARS.

1. Criteria that address the quality of the assessments

   The Six Quality Assessment Criteria primarily evaluate the procedures used by districts in the design of their assessment system. The quality of the assessments, themselves, should be evaluated periodically by the state department. This may include an evaluation of the content and technical quality of the items; an evaluation of the rationale for choice of item formats; an evaluation of the quality of the scoring rubrics; an evaluation of the coherency between items and scoring rubrics; an evaluation of the quality of the procedures used to score student responses; and an evaluation of the quality of the administration of the assessments in the classroom.

2. Criteria that address coherency of assessments across grade levels

   Criteria could address the coherency of assessments across grade levels with respect to the content and cognitive skills being assessed as well as the setting of performance standards. As indicated in *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001) coherency is essential if an assessment system is to support learning. Therefore, it is important to evaluate the extent to which there is coherency across grades and coherency across subject areas. The evaluation of coherency across grade levels in terms of both the content and cognitive demands of the assessments could be included in Criteria 1 (assessments reflect the state/local standards) and Criteria 4 (assessments are at the appropriate level),

3. Criteria that address accommodations

The Criteria do not address accommodations for the assessments. It should be noted however that the Nebraska Department of Education published a document on accommodations, *Accommodations Guidelines: For the Instruction and Assessment of Students with Disabilities*. Criteria that address strategies that accommodate the needs of students with disabilities would be consistent with the recommendations of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Criteria should address the need to provide evidence for appropriate procedures for modifying presentation format, response format, and timing and test setting. The criteria should also address the need to provide evidence for the suitability of alternate assessments if they are used by districts.

To be consistent with the *Standards*, the criteria could also stipulate that if there is a large enough sample size, evidence should be provided for "the validity of inferences made from test scores and the reliability of scores on tests administered to individuals with various disabilities" (p. 107, AERA, APA, & NCME, 1999). Clearly, this would not be possible for districts that have a small number of students with disabilities, but it may be reasonable for districts that have a sufficient number of students with disabilities to provide some validity evidence. While the evidence may not be as comprehensive as the evidence provided for assessments administered to students without disabilities, evidence to support the use of accommodations for students with disabilities would enhance the credibility of the assessments.

4. Criteria that address assessment design issues relevant for students with different language backgrounds and other subgroups

Quality Criteria 3 addresses some issues related to the fairness of assessments to various subgroups. To meet NCLB goals, subgroups of students such as minority students, socioeconomically disadvantaged students and English Language Learners (ELLs), need to make continued progress on state assessments. There is a heightened need for assessment developers to be sensitive to a variety of issues in developing assessments for which all students have access and that provide valid score interpretations for student subgroups. As an example, one way to help guard against construct-irrelevant variance is to simplify the language of test items so as to reduce the level of unnecessary linguistic complexity and potential cultural bias (Albedi & Lord, 2001). Such design issues for ELLs are particularly relevant given hat the U.S. Census Bureau estimates that ELL students will constitute as much as 40% of children in school by 2030. Although, the number of ELLs in Nebraska is lower than many other states, there is an increasing number of ELL students attending schools in Nebraska. As another example, districts with large student populations may be able to conduct differential item functioning (DIF) analyses for some of their assessments. Additional assessment design issues that are relevant for ELLs as well as other subgroups of students could be incorporated into the Criteria.

5. Criteria that address reporting of results

Criteria that address the reporting of results could be included in STARS. The criteria could address the need to evaluate the accuracy and comprehensiveness of the reported results as well as to evaluate the extent to which the scores/results are interpreted accurately by stakeholder groups.

6. Criteria that address Consequential Validity Evidence.

While some information is obtained regarding the effect of district assessment systems on curriculum and instruction through Criteria 2 (opportunity to learn), additional information on the consequences or the impact of the assessment system on instruction and student learning would strengthen the validity evidence. This is especially important given that the purposes of STARS as indicated in the *School-based Teacher-led Assessment and Reporting System: A Summary* (September 2004) are to "improve educational opportunities" and "improve learning"" (p. 1).

As Standard 13.1 in the *Standards* (AERA, APA, NCME, 1999) state, "When educational testing programs are mandated by school, district, state or other authorities, the ways in which the test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from uses of the test, both intended and unintended, should also be examined by the test user" (p. 145). Consequential validity evidence could include an evaluation of the impact of the assessment on the quality of instruction and student learning and the relationship between the impact of assessment on instruction and school performance gains on the assessment.

**Additional Recommendations**

This section proposes some additional recommendations that may enhance the Nebraska School-based Teacher-led Assessment and Reporting System. It should be noted that some of these recommendations have been implied earlier in the report.

1. Examine the consistency across the *District Assessment Portfolio Rubric, A Guide for Assuring the Technical Quality of Classroom Assessment (March 2006)*, and the *District Portfolio Assessment Rating Form*. *A Guide for Assuring the Technical Quality of Classroom Assessment* provides useful guidelines for teachers in applying the Six Quality Assessment Criteria to their classroom assessments and provides further elaboration on the Six Criteria.

   The *District Portfolio Assessment Rating Form* provides a set of comments that can be applied to a portfolio in addition to the rating of met (no further comment is necessary), met (some further comment is necessary), needs improvement, and not met. These comments are useful ways of providing feedback to the districts regarding their portfolios. A review of the comments in relation to the *District Assessment Portfolio Rubric* and *A Guide for Assuring the Technical Quality of*

*Classroom Assessment (March 2006)* would be worthwhile to ensure their consistency with these documents.

2. Continue to include national measurement experts in the portfolio review process. A measurement expert is vital in the training of the panel who reviews portfolios. In addition, measurement experts should also serve on the panel who reviews portfolios. This will help ensure the accuracy of the review process as well as lend credibility to the review process.

3. Continue to emphasize the content of *A Guide for Assuring the Technical Quality of Classroom Assessment* and how it relates with the *District Assessment Portfolio Rubric* in the portfolio review process.

4. Continue to provide professional development activities to schools and districts that address the criteria, especially the technical criteria, as described in *A Guide for Assuring the Technical Quality of Classroom Assessment*.

5. The Nebraska Department of Education may also find the document, *Dealing with Flexibility in Assessments for Students with Significant Cognitive Disabilities* (Gong & Marion, 2006), valuable in addressing issues in the design and interpretation of alternate assessments that are based on alternate achievement standards. This document suggests the need for flexibility in curricular goals, the content and skills students are expected to learn during a particular time span, between students at a particular point in time and over time; flexibility in the instruction; flexibility in the content standards to be assessed; flexibility in the methods/items used to assess; flexibility in the scoring; variance in the performance standards; flexibility in the interpretation and reporting; and flexibility in how scores are handled for school accountability. The document also provides guidelines on how to deal with flexibility in assessments in these areas.

## Evaluation of the Quality Criteria Rating Chart and Recommendations

The *Quality Criteria Rating Chart* uses the ratings provided by the *District Assessment Portfolio Rubric* to provide one overall rating of the quality of the assessment for each grade level portfolio. The overall ratings are Exemplary, Very Good, Good, Needs Improvement, and Unacceptable. As an example, if a portfolio receives a "Met" for each of the Six Quality Criteria, the portfolio would be awarded an Exemplary rating. In contrast, if a portfolio receives a "Met" for Criteria 1 and 2, and a "Not Met" for Criteria 3, 4, 5, and 6, the portfolio would be awarded a Needs Improvement rating; and if a portfolio receives a "Met" for either Criteria 1 or 2, and a "Not Met" for Criteria 3, 4, 5, and 6, the portfolio would be awarded an Unacceptable rating. In order to be acceptable, Criteria 1 (assessments match the standards) and 2 (students have an opportunity to learn) must receive a "Met", indicating that the *Quality Criteria Rating Chart* weighs these two criteria more heavily than the other four criteria. It is clear that these two Criteria are the foundations for any quality assessment system; however, now that district educators and administrators have had a number of years working with

STARS, it would be reasonable to consider increasing the standards for the Quality Rating Chart. It would be useful to have a formal panel convene to evaluate the appropriateness of the way in which the ratings are awarded in the *Rating Chart*, and to determine if any modifications are warranted.

Recommendation:

1. Validity evidence should be obtained to help determine the appropriateness of the criteria specified in the *Quality Criteria Rating Chart* for the assignment of the ratings. A panel composed of measurement experts, national content experts and Nebraska educators and administrators could review the *Quality Criteria Rating Chart* and the criteria specified for the Six Quality Criteria for assigning the 5 overall ratings (Exemplary, Very Good, Good, Needs Improvement and Unacceptable). This panel could review targeted portfolios that range in quality with respect to the Six Quality Criteria and classify the portfolios into the 5 ratings ranging from Exemplary to Unacceptable. The process could be iterative in that discussions could follow the independent assignments and then panelists would have the opportunity to make changes in their assignments.

## Conclusion

The Nebraska School-based Teacher-led Assessment and Reporting System (STARS) provides the opportunity for a state assessment system to have an integral role in teaching and learning at the classroom level. Conversations with district educators and administrators suggest that STARS serves as a valuable tool in the educational system, allowing for the meaningful assessment of student achievement and progress. Professional development efforts in Nebraska appear to have played a key role in ensuring the quality of district assessment systems. As an example, professional development efforts have included support materials, workshops, conferences, training sessions, interactive data bases, and professional development activities supported by Educational Service Units. In addition, a Trainer of Trainers model was also established in 1999 to ensure more teachers within the districts become assessment literate. Further, the website for STARS has valuable documents that support the implementation of STARS at the district level.

The Nebraska Department of Education has been diligent in trying to address the technical quality of the district assessment systems. The Six Quality Assessment Criteria are consistent with many Standards in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), providing valuable support for the use of STARS. While the *Standards* was written for large-scale assessments, not classroom-based assessments, it is relevant in evaluating the quality of STARS given its use as an accountability system under NCLB. The *Standards* point to a number of areas that could be incorporated in Criteria used to evaluate the district assessment systems, providing additional evidence of the technical quality of the assessment systems. This report

provides a number of recommendations that will provide additional validity evidence in support of using STARS as a state assessment and accountability system.

## References


Albedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219-234.

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.

Baker, E. L., Linn, R. L., Herman, J.L., & Koretz, D. (2002). *Standards for Educational Accountability Systems*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.

Brookhart, S. M. (2005). The quality of local district assessments used in Nebraska's School-based Teacher-led Assessment and Reporting System (STARS). *Educational Measurement: Issues and Practice, 24*(2), 14-21.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32.

Gong, B., & Marion, S. (June 2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities* (Synthesis Report 60). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Herman, J.L., & Klein, D.C.D. (1996). Evaluating equity in alternative assessment: An illustration of opportunity-to-learn issues. *The Journal of Educational Research, 89*(4), 246-256.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 15-21.

Nebraska Department of Education (2005a). A Guide for the Technical Quality of Classroom Assessment. Lincoln, NE: Author.

Nebraska Department of Education (2005b). *STARS: School-based Teacher-led Assessment and Reporting System – Update #18.* Linclon, Nebraska: Author.

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment* (National Research Council). Washington, DC: National Academy Press.

Plake, B.S., Impara, J.C., & Buckendahl, C.W. (2004). Technical Quality Criteria for Evaluating District Assessment Portfolios Used in the Nebraska STARS. *Educational Measurement: Issues and Practice, 23*(2), 12-16.

Roschewski, P. (2004) History and background of Nebraska's School-based Teacher-led Assessment and Reporting System. *Educational Measurement: Issues and Practice, 23*(2), 9-11.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F. Snowden, J.B., Gordon, E. & Gutierrex, C. & Pacheco, A. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do.* San Francisco: Jossey-Bass.