Revised
2007

Guidelines & Requirements for Documenting Assessment Quality for STARS

School-based Teacher-led Assessment and Reporting System

Nebraska Department of Education
2007-2008

This revised document replaces
"Assessment Portfolio Instructions and Suggestions" and
"A Guide for Assuring the Technical Quality of Classroom Assessment"

Table of Contents

I. General Guidelines for Documenting Assessment Quality

The following general guidelines will be helpful to school districts in documenting the quality of their assessment processes.

What Constitutes Documentation of Assessment Quality?

Beginning in 2000-2001 public school districts in Nebraska were required to provide written documentation of their local assessment quality including processes, procedures and samples of the assessments themselves. This documentation was submitted to the Nebraska Department of Education in the form of a written portfolio that provided evidence of meeting the Six Assessment Quality Criteria.

Beginning in 2006-2007 the process for documenting assessment quality changed to include an on-site Nebraska-led Peer Review of STARS. This on-site review process provided the opportunity for districts to share the evidence documenting assessment quality including processes, procedures, results, and the assessments themselves. The on-site review allows both written and verbal documentation.

The Six Quality Assessment Criteria

1. The assessments match the standards.
2. Students have an opportunity to learn.
3. The assessments are free of bias and sensitive situations.
4. The assessment levels are at the appropriate level.
5. There is consistency of scoring.
6. The mastery levels are appropriately set.

Using the Assessment Quality Rubric

In assembling the documentation for assessment quality districts should be guided by the Assessment Quality Rubric. (See Attachment A) The rubric identifies how assessments meet the Six Quality Assessment Criteria. Evidence should include:

- **Who** did the process?
- **What** did they do in this process?
- What were the **results** of the process?

Consistency Between and Among the Assessment Quality Criteria

The assessment quality documentation provides a description of the district assessment system overall. Therefore, the processes described for each Quality Criterion should be consistent with each of the other criteria. The following are examples of processes or procedures that are **not** consistent:

- For Criterion #1 (*Assessments reflect the state or local standards*.) the district reported using 16 assessments throughout the year to measure reading achievement. For Criterion #5 (*There is consistency of scoring*.) only 7 of those assessments have been included in the reliability calculations.

- For Criterion #5 (*There is consistency of scoring*.) the district calculated reliability for the assessments by strand (reading, listening, speaking), but assigned scores or mastery levels for Criterion #6 (*The mastery levels are appropriate*.) by standard.

- For Criterion #1 (*Assessments reflect the state or local standards*.) the district reported on both objective and subjective items, but in Criterion #5 (*There is consistency of scoring*.) reliability is reported only for objective items.

- For Criterion #5 (*There is consistency of scoring*.), the district used a decision consistency method for two proficiency levels (met/not met); however, in Criterion #6 (*The mastery levels are appropriate*.) mastery levels for four proficiency levels (beginning, progressing, proficient, and advanced) were reported.

- For Criterion #1 (Assessments reflect the state or local standards) a district revised assessments and replicated the alignment/sufficiency. However, in Criteria #3 and 4 the district did not conduct bias or appropriate level reviews.

Performance Level Descriptors

A key element in the overall assessment process and in the documentation of assessment quality is the establishment of Performance Level Descriptors. These statements describe what student performance looks like for each proficiency level and will guide the work in aligning assessments to standards and in setting the assessment mastery levels. Performance level descriptors should represent a consensus of those who are responsible for setting the mastery levels. Because of new assessment legislation in 2007, level descriptors will be determined at the state level for implementation in 2009-10. This process will be completed with groups of educators across the state.

Acceptable vs. Appropriate Methods for Documenting Assessment Quality

Districts may use a variety of methods for determining how their assessments meet the Six Quality Criteria. While many methods may be acceptable under certain conditions, not all methods are appropriate for all situations. Therefore, it is very important to distinguish the difference

between using an acceptable method versus using an appropriate method. Appropriate methods match both the type of assessment and the student population.

The following examples are intended to illustrate the difference between "acceptable" and "appropriate."

Example One: Criterion #4 *(Assessments are at appropriate levels for students.)*

It may be **acceptable** to complete a readability analysis for determining appropriate level of a reading passage in an assessment. However, it is not appropriate to conduct a readability analysis when the assumption for using the readability method has not been met (e.g., sufficient number of words, assessment directions, assessment items, tasks, rubric.)

A readability analysis would be **appropriate** only when a sufficient number of words can be analyzed.

A more **appropriate** method for documenting Criterion #4, in the case of an assessment with a limited number of words, would be a method of professional judgment by qualified teachers whose professional judgment regarding appropriate level has been validated in a team process.

Example Two: Criterion # 5 *(There is consistency in scoring.)*

It may be an **acceptable** method to calculate consistency or reliability of scoring on a multiple choice test by using an internal consistency method such as a KR-20 or KR-21, but if the group of students consists of fewer than 30, an internal consistency method may not be the most **appropriate** method. A more **appropriate** method would be to use a decision consistency method such as calculating agreement between two comparable independent judgments.

Using appropriate, rather than merely acceptable methods are expected to be consistent across the Six Criteria and allow the assessment system to function coherently.

Decision or Business Rules: What Are They and Why Are They Important?

As districts describe their step-by-step process for meeting the Six Quality Assessment Criteria, it is important to explain how decisions were reached by those who conducted the process. These are commonly referred to as "decision rules" or "business rules."

For example, in Criterion #1 *(The assessments match the standards.),* a group of individuals deciding whether an assessment item matches a standard or whether there are sufficient numbers of items at each proficiency level, requires some basis for how their final decision is reached. The following questions provide insight into how decision rules might be established:

1) Did the group reach 100% agreement?
2) Did 5 of 6 members have to agree?
3) Was the item or issue discussed until consensus was reached?
4) How were dissenting opinions handled?

5) How many dissenting opinions prevented the agreement?
6) Did individuals make their judgments independently and then move toward consensus in a group?
7) Was the decision based on the whole group or a small group?

In order for a district to describe their decision rules, a district would not have to answer all of the above questions, but rather describe in general the processes used in their final decisions.

What Constitutes Evidence in the Nebraska-led Peer Review of STARS?

The peer review offers opportunities for dialogue and conversation about the assessment practices and procedures within a district. Although a complete assessment portfolio does not need to be prepared for the on-site peer review, certain evidence needs to be available or electronically displayed during the review.

For each of the six criteria, districts should have the following documents available on-site:

Who did the process?

- ☆ A list of educators involved, qualifications, degrees, years of experience – the criteria making them experts.

What process did the district conduct?

- ☆ Step-by-step what was done, who led it and their qualifications, the business rules, the methods or procedures districts used to come to agreement or decision.
- ☆ Evidence of the process, i.e., complete worksheet, a lesson plan, a calculations or review sheet, depending upon the criterion.

Results of the process?

- ☆ A chart of results for all standards being reported, including changes made and the plan for making those changes in the current year.

Please note: If the following evidence is electronically displayed, a hard copy needs to be available upon request of the peer reviewers.

How Often Should Reliability Be Calculated?

Insuring the fairness of scores and the confidence in the use of those scores for each administration of the assessments is a process that should occur every year.

In the case of the peer review process scheduled before the end of the year, prior to all assessments being administered or reliability being calculated, reliability calculations from those same assessments during the previous year are acceptable as evidence for the on-site peer review as long as the assessments or the method for calculating reliability was appropriate and has not changed.

How Often Should Mastery Levels Be Set?

Mastery levels are either student centered or content based. The general rule is that if the definition of proficiency (the PLD's) or the assessments change, the mastery levels process should be reviewed.

Student centered methods such as the "Modified Contrasting Group" method bases cut scores on teachers' professional judgment about students. These are not appropriate methods to use with small numbers of students. If these methods are used with a large, stable representative group of students, they do not need to be repeated every year.

Content-based methods such as the "Modified Angoff" is content-based, as professional judgment is specific to the items/tasks on the assessment. Therefore, these mastery levels will need to be revisited when the items, tasks, or PLD on the assessment change.

II. Assessment Quality Criteria

Quality Criterion #1:

The assessments match the standards.

- This criterion assures that the assessments match the standards in content, cognitive complexity, and performance demand (validity) and that there are enough opportunities within the assessments to demonstrate skill or knowledge at each proficiency level (sufficiency).
- Evidence of the match of assessments to the standards is based on an independent review of the alignment of assessments to the standards.
- A minimum of 12 objectively-scored items or equivalent on reading standards 4.1.3, 8.1.1, and 12.2.1 and math standards 4.2.1, 8.2.2, and 12.2.1 are required.
- Subjectively-scored tasks must be determined to provide sufficient opportunities for all students to show their knowledge or skill on the standards.
- Documentation for meeting this criterion is required for all standards used for reporting.

In meeting this criterion, districts should provide documentation to answer the following questions:

I. Who did the process?

This panel must be independent reviewers. They may not be the assessment writers or developers.

The independent reviewers should be experienced professional educators who are familiar with the content and grade level being assessed. It is recommended that a span of grades be represented (e.g. 3-6 grade teachers for a 4th grade assessment.) It is helpful to include special education teachers, a school psychologist, and a counselor in the group.

- ✓ The documentation for *Who did the process?* may or may not include the names of the panelists, but should include their years of experience, the grade level(s) they teach, and/or areas of expertise that would qualify them be subject matter experts capable of conducting this process.
- ✓ Information regarding the K-12 composition of the panel and the number of teachers who participated in comparison to numbers of total staff will help describe the adequacy of the representation on the panel.
- ✓ Describe who led the process. Was it a teacher leader, an educational service unit staff developer, an administrator? Include qualifications of the leader(s) of the process.

2. *What did they do in this process?*

Describe how the panel matched assessment items or tasks to the standards.

- ✓ The panelists should examine the standard and the assessment item(s)/task(s) and make an independent decision about whether they match in content, cognitive complexity, and performance demand.

- ✓ Each panelist should make independent decisions about whether or not the assessment task/item(s) capturing the essence or main purpose of the standard.

- ✓ After the independent decisions are made about the match to standards, they should be recorded. Panelists should identify the name of the assessment and the item type (subjectively or objectively scored). Forms such as Attachment B may be used to record the responses.

- ✓ The panelists should decide how consensus will be reached. This will require the panel to determine decision rules for this process (e.g. 100% agreement, discuss until the majority agrees, etc.) Explain the business rules or the process used to come to agreement.

Describe the process used by the panel to determine that there are sufficient items or tasks for each assessment.

- ✓ Explain the task of sufficiency to the panel. Sufficiency ensures that there are enough items or tasks at each performance level (beginning, progressing, proficient and advanced) so that students at all levels can demonstrate their skill/knowledge.

- ✓ The independent group needs to know the performance level descriptors (PLDs) that have been developed with the assessments in order to determine sufficiency. Performance level descriptors are written illustrations of student performance at advanced, proficient, progressing, and beginning levels. *(See Criterion #6, p. 27, for a description of how to establish PLDs.)*

- ✓ Based on PLD's, each panelist will make independent decisions about the difficulty level of each assessment task/item by assigning each task/item to a proficiency level.

- ✓ Individual decisions should be recorded on sheets such as Attachment B.

- ✓ The group will come to consensus agreement on decisions of sufficiency. Decision rules for this process will need to be determined and recorded..

3. *What were the results of this process?*

Describe what the results were. This documentation should address the following questions:

- What happened as a result of the committees working together?
- Which items or types of assessments were changed? What, if necessary, is the plan for making those changes?
- What was decided? Why?

✓ Determine and record final decisions about sufficiency and any needed changes to be made on a summary record sheet similar to worksheets, but keep the individual panelists' worksheets as documentation.

✓ Needed changes in the assessments related to alignment and sufficiency should be made and reviewed for Criteria 1-6 before the assessments are administered again.

Quality Criterion #2

Students have an opportunity to learn.

- This criterion assures that the standards are present in the local curriculum and that students have been taught at least 80% of the content prior to being assessed on it.
- Districts need to document the dates of instruction and assessment.
- Documentation for meeting this criterion should include all assessments/standards.

In meeting this criterion, districts should provide documentation to answer the following questions:

Note: <u>Criterion #2 is required for all standards.</u> The other criteria, 1, 3, 4, 5, 6, are required only for the standards used for reporting.

I. <u>Who did the process?</u>

Unlike the panel used for Criterion #1 *(The assessments match the standards)*, educators who participate in this process <u>do not need to be independent</u> of the assessment writers or developers. This process needs to be done by each individual district, even if they participate with other districts in implementing and documenting their assessments.

- ✓ Convene a group of educators who teach the local curriculum.
- ✓ This group may include the assessment developers.
- ✓ This group should include teachers who give the assessments.
- ✓ The documentation for *<u>Who did the process?</u>* may or may not include the names of the panelists, but should include their years of experience, the grade level(s) they teach, and/or areas of expertise that would qualify them to conduct this process.
- ✓ Information regarding the K-12 composition of the panel and the number of teachers who participated in comparison to numbers of total staff will help describe the adequacy of the representation on the panel.
- ✓ Describe who led the process. Was it a teacher leader, an educational service unit staff developer, an administrator? Include qualifications of the leader(s) of the process.
- ✓ If forms or questionnaires were used to collect this information, include a copy.

2. *What did they do in this process?*

Timing is an important component of this criterion because it must be demonstrated that assessments are given after sufficient instruction has occurred.

Describe how the panel(s) examined the local curriculum to find where standards were taught.

- ✓ If the panel(s) examined textbooks and instructional materials, describe how they determined when the standards were covered during the school year.
- ✓ If the panel(s) collected information through lesson plans, classroom assignments, assessments, or classroom observations, explain what percentage of teachers were involved, who examined these products or made the observations, and how it was determined that the content of assessment(s) was being taught.
- ✓ If the panel(s) used surveys or questionnaires, include the forms used for that process.
- ✓ Describe how the panel(s) determined when assessment(s) would or should occur.
- ✓ The panel(s) should come to a consensus about when standards are taught.
- ✓ Panelists need to agree as a group when the assessments are given in relationship to instruction so that students have the opportunity to receive instruction on 80% or more of the standards prior to assessment. Explain the business rules or the process used to come to agreement.

3. *What were the results of this process?*

Describe the results of the review.

- ✓ The dates of instruction and assessment for all standards should be recorded on sheets such as Attachment C.
- ✓ Include the total percentage of content that is taught prior to assessment.
- ✓ Any redundancy of standards or absence of standards needs to be identified and noted. The same is true for any inappropriate timing of instruction or assessment.
- ✓ A plan and a timeline for addressing any needed changes in opportunity to learn needs to be developed and the appropriate changes made.
- ✓ Include as evidence a page from the aligned curriculum guide and sample documents used to collect instructional dates for standards. Keep products for determining instructional dates as documentation.

Quality Criterion #3

Assessments are free of bias and sensitive situations.

- This criterion assures that a review panel has examined the assessments for fairness and that nothing in the assessments or the directions is inappropriate, insensitive, demeaning, or unclear.
- Bias is more than stereotyping. It may also include issues of fairness, appropriateness of directions, graphics, and poorly written items.
- Districts should provide evidence that assessment content has been examined critically to be free of bias against any group or population of students (e.g. gender, race/ethnicity, socioeconomic status, religion).
- Assessments should have been examined to be free of offensive language and stereotyping.
- For this criterion to be met, there first needs to be an orientation to bias for the assessment developers. Then, there needs to be an examination of the content of the assessment(s) by a panel that is qualified to conduct a bias review.
- Documentation for meeting this criterion is required for all standards used for state reporting.

In meeting this criterion, districts should provide documentation to answer the following questions:

I. Who did the process?

It is recommended, but not required, that the bias review be conducted by a panel of individuals who were not the assessment writers or developers.

- ✓ The bias review process requires a staff development component so that bias review panelists are aware of sensitive situations.
- ✓ Describe who led the bias orientation training and include the qualifications of that person or those individuals.
- ✓ The documentation for *Who did the process?* may or may not include the names of the panelists, but should include an explanation of the expertise or training that would qualify them to conduct this process.

- ✓ If the bias review panel was made up of individuals from outside the school district, include the number of panelists, their qualifications and what was important about having them on the panel.

- ✓ If the bias review panel was made up of teachers from within the district, include the number of panelists, their qualifications and what was important about having them on the panel.

- ✓ If panel members were included because of diversity issues, include their qualifications and the reasons they were chosen to be part of the process.

2. What did they do in this process?

Describe the process used to insure that items/tasks in the assessment(s) are free of bias. It is not appropriate to merely "accept" the assurance of a text book company or of a test maker that items are bias-free or not sensitive. Items/tasks need to be examined locally to eliminate or change any items of an inappropriate or sensitive nature.

- ✓ Describe the bias orientation training that was provided for participants in the bias review process. Include a sample of the training material.

- ✓ As part of the bias orientation training reviewers should practice identifying examples of unfairness and offensiveness on sample assessments.

- ✓ Describe the process bias reviewers conducted in examining the assessments for bias and sensitive situations.

    - The bias reviewers should independently examine the assessments, item by item, identifying any possible instances of bias or sensitive situations.

    - Reviewers should record their independent responses for each assessment item on forms such as Attachment D.

    - The panelists should then determine collectively the instances of bias and offensiveness needing to be changed in the assessments.

    - Decide how the group will come to consensus on decisions of bias.

    - If a statistical method was used to assess bias, specify and describe the method used.

    - Needed changes should be documented and a plan for making those changes should be identified.

3. *What were the results of this process?*

Describe the results of the review.

- ✓ The final decisions of the group should be recorded for each item or task for each assessment, but keep the individual panelists' worksheets as documentation.

- ✓ Needed changes related to bias should be made and reviewed for Criteria 1-6 before the assessments are administered again.

Quality Criterion #4

The assessments are at the appropriate level.

- This criterion assures that the cognitive (thinking) level of the assessments is appropriate for the grade level being assessed.
- Districts should provide evidence that assessment content has been examined to determine the appropriateness of the level.
- This criterion can be met if there is evidence that a panel of individuals qualified to determine the level of appropriateness, examined the content and level of the assessment(s) and found them to be appropriate.
- When appropriate, the criterion may also mean an appropriate reading level for the standard being assessed.
- Documentation for meeting this criterion is required for all standards used for state reporting.

In meeting this criterion, districts should provide documentation to answer the following questions:

I. Who did the process?

A panel of educators familiar with the grade level and content should review the assessments for appropriate level.

- ✓ It is recommended that a span of grades be represented (e.g. 3-6 grade teachers for a 4th grade assessment.) It is helpful to include special education teachers, a school psychologist, and a counselor in the group.
- ✓ The documentation for *Who did the process?* may or may not include the names of the panelists, but should include an explanation of the expertise that would qualify them to conduct this process.
- ✓ Describe the panelists: their years of experience, the grade levels that are represented and in what configurations (e.g. all the same grade level, grades above or below the assessed grade), information indicating that these panelists are qualified judges of the appropriate level of tasks for students.
- ✓ Information regarding the K-12 composition of the panel and the number of teachers who participated in comparison to numbers of total staff will help describe the adequacy of the representation on the panel.

- ✓ Describe who led the process. Was it a teacher leader, an educational service unit staff developer, an administrator? Include qualifications of the leader(s) of the process.

2. *What did they do in this process?*

Describe the process used to review items or tasks to make sure they were at an appropriate level.

- ✓ Describe the process used for reviewing assessments (e.g. teacher judgment, readability analysis).
- ✓ If a teacher judgment method is used, the review panel should examine each task in review the assessments item by item.
- ✓ Prompts, tasks, rubrics should also be reviewed for appropriate level.
- ✓ If the evidence of the review was collected through forms, questionnaires, etc., provide examples of these collection tools.
- ✓ If a process of teacher judgment was used, describe how reviewers reached consensus about their judgments. Describe the business rules.
- ✓ If an analysis of readability level was used, indicate which one(s) was used and why the process was selected. Considerations in using a readability analysis include:
  - *The text being analyzed should be of sufficient length for the analysis to be reliable.*
  - *A textbook's statement about level of readability without supporting evidence is not considered adequate evidence.*
  - *If readability levels are inappropriate, the district needs a plan for making changes.*

3. *What were the results of this process?*

Describe the results of the review.

- ✓ Record any needed changes and recommendations regarding the final decisions about appropriate level. (See Attachment E)
- ✓ If a readability analysis was conducted, include a description of the results and the changes that were made to ensure that assessment materials were presented at an appropriate level.
- ✓ Needed changes related to appropriate level should be made and reviewed before the assessments are administered. Evidence of panel's work should be kept as documentation.

Quality Criterion #5

There is consistency of scoring.

- This criterion assures the consistency and reliability of scoring so that educators can have confidence in the inferences about student performance results generated by the assessments.

- This criterion addresses the consistency and reliability of scoring for objectively scored items as well as subjectively scored tasks, prompts, or performances. This criterion asks the district to provide evidence that scores on assessments are reliable in terms of the consistency of scores that might be expected of students who might take the assessment(s) on more than one occasion without intervening instruction.

- For subjectively scored assessments (e.g. essays, student research papers, speeches) there should be clear scoring criteria (a scoring guide or rubric), and training of raters on the scoring criteria, and show the evidence that the scoring criteria are applied consistently. This is typically demonstrated by showing that multiple scorers have independently agreed in their scoring of a sample of student work.

.

- This criterion requires that the district use an appropriate process for documenting reliability of scoring.

- A reliability measure of .70 or higher averaged across all standards (used for state reporting) is required for this criterion to be fully met.

- A reliability measure of less than .70 averaged across all standards (used for state reporting) will result in a rating of *Needs Improvement*, if the district provides a plan for improving the measure of reliability.

- A reliability measure of less than .70 averaged across all standards (used for state reporting) will result in a rating of *Not Met*, if the district provides no plan for improving the measure of reliability.

- Documentation for meeting this criterion is required for <u>all</u> standards used for state reporting.

In meeting this criterion, districts should provide documentation to answer the following questions:

1. <u>Who did the process?</u>

- ✓ Indicate the person(s) or teams who led, conducted, and participated in the reliability analysis.
- ✓ Include their qualifications.

2. *What did they do in this process?*

There are several ways to measure the reliability of scores on a district's assessments.

- ✓ Districts may calculate reliability values by individual standard, by assessment, or by strand. There are advantages and disadvantages to each approach.

Design	Advantages	Disadvantages
By individual standard	• Provides specific data about each standard's reliability/ consistency. • Provides more specific data about student performance for school improvement.	• Requires more calculations.
By groups of standards, by assessment or by strand	• Requires fewer calculations when standards are grouped.	• Provides less specific information about student performance. • Provides only "global" information about reliability.

- ✓ **A district should determine the appropriate method for calculating reliability based on the answers to three questions:**

  a) What design will provide the information needed for the decisions we want to make?

  b) What type of assessments have we designed? (objectively scored or subjectively scored)

  c) What is the best method based on the number of students we've assessed?

- ✓ Apply the appropriate method for calculating consistency or reliability of scoring. Attachment F includes a decision tree for districts to determine the appropriate method for calculating reliability.

- ✓ Providing a rationale for the method chosen is helpful.

3. *What were the results of this process?*

- ✓ Report the reliability values calculated for the assessments used.

- ✓ Provide the average reliability calculated across all standards used for state reporting.

- ✓ The measure of reliability averaged across all standards used for reporting should be .70 or higher for this criterion to be fully met.

- ✓ If an appropriate method to calculate reliability has been used but the measure averaged across all standards used for reporting is less than .70, a plan for improving the measure of reliability should be included. If the plan will likely improve reliability, the district will likely receive a rating of *Needs Improvement* for this criterion.

- ✓ If the measure of reliability averaged across all standards used for reporting is less than .70, and no plan for improving the measure of reliability is included, the district will likely receive a rating of *Not Met* for this criterion.

- ✓ Use the solutions for low reliability chart to write a plan for improving low reliability if it occurs.

- ✓ A *Needs Improvement* rating on this criterion prevents an *Exemplary* rating overall.

Methods for Calculating Reliability

The method for calculating reliability is determined by the type of assessment (objectively or subjectively scored) and the number of students assessed.

Method	Type of Assessment	Number of Students Assessed
Internal Consistency (KR-20), KR21, Coefficient Alpha, Split Half)	Objectively Scored	May need large number of students for stable results (30 or more)
Decision Consistency, Test-retest, Parallel Forms	Objectively Scored or Subjectively Scored (if the two decisions are independent	May be used with any number of students
Inter-rater Reliability	Subjectively Scored	May be used with any number of students

Internal Consistency Methods (KR20, KR21, Coefficient Alpha, Split Half)

1. These methods are most appropriate only for groups of 30 or more students and for objectively-scored assessments. Small schools could collect results over multiple years to reach 30 or join with other districts to reach sufficient numbers.

2. These methods are most easily computed using computer software. They involve entering data results into a program and generating reliability values.

3. The directions for each statistical analysis program must be learned and followed.

4. If assessments are administered multiple times to the same student, the results of the first administration should be used in the internal consistency calculations.

5. These methods do not rely on a teacher's professional judgment in the calculation.

Decision Consistency Method

1. This method is used primarily with objectively scored assessments, but may be used with subjectively scored assessments (if the two decisions about students' performance are independent.) Also, districts should be cautious about using one set of professional judgment for two processes. For example – using the same teacher judgment to calculate reliability and setting cut scores with the modified contrasting group, and using the same set for professional judgment.

2. This method is helpful to small districts but can be used with any number of students as long as the sampling of students used is representative of the student population in the district.

3. In this method you will need to have established your mastery levels and agreed upon the Performance Level Descriptors (PLDs). Beginning in 2009-10, the PLD's will be generated at the state level.

4. This method requires two independent decisions about a student performance. Basically, this method involves the calculation of the percentage of times the two independent decisions agree. The two decisions could be based upon either of the following:

    - Assessment results from two assessments measuring the same thing at the same level of difficulty. Both assessments would have to meet the Six Quality Criteria.

    - Assessment results from CRT and results from an NRT (again, the CRT would need to have been run through the Six Quality Criteria.) This approach could only be used with those standards that have been determined to match the NRT's.

Example: Decision Consistency – NRT - CRT

	CRT Prof. Level	NRT Prof. Level	Agreement
1. Joe	2	2	+
2. Sue	3	2	0
3. Pam	4	4	+
4. Tom	1	1	+
5. Ned	2	2	+
		Total =	4/5 = .80

✓ For each student record, the student's performance level on the standard (as determined by the cut score process.) The cut score levels for the NRT are the same use for NRT reporting to NDE:

NRT – Percentile 1-24 = 1, 25-49 = 2, 50-74 = 3, 75-99 = 4

- ✓ After the second assessment, record for each student the performance level for each standard. Then you calculate the percentage of agreement. If a student receives the same classification (1, 2, 3, 4) on both assessments, then the results are in agreement and this is noted by a + in the table. If the assessment results are not in agreement, it is marked with a 0 in the table.

- ✓ The calculation is determined by dividing the number of agreements by the number of students. In the example shown, the calculation would be 4/5 or .80 agreement.

- ✓ If students are being classified according to performance levels (e.g., Beginning Progressing, Proficient, Advanced), calculating the measure of reliability will be based only on the number of times the decisions match exactly.

Four or Fewer Proficiency Levels

1	**2**	**3**	**4**
Basic	Progressing	Proficient	Advanced
	(must use exact match decisions)		

6. If students are being classified according to more than four mastery levels, both exact and adjacent match decisions may be included in the reliability calculation.

More than Four Proficiency Levels

1	**2**	**3**	**4**	**5**
Basic	Emerging	Satisfactory	Very Good	Exemplary
	(may use exact match and adjacent decisions)			

Teacher Judgment Decision Consistency

1. Teachers participating in this reliability method need to review the Performance Level Descriptors (PLDs) that were developed in Criterion Six and used by the independent review team in Criterion One to examine the assessment sufficiency.

2. Through this review teachers should have a common understanding of student performance at each of the levels: advanced, proficient, progressing, and beginning. Training for this process should include a means for teachers to reach this common understanding.

3. Based on the PLDs, teachers make an independent professional decision about the performance level they believe their students will achieve. This judgment needs to be made before the teachers know the assessment results.

4. The teachers' judgments should be recorded. They may be recorded by standard, by groups of standards (strands), or by assessments. (See Attachment G)

5. The actual calculations of reliability cannot be completed until the assessments are scored and mastery levels determined.

6. Once the scoring is done and mastery levels set, the rest of the worksheet can be completed.

7. The results of the actual assessments are recorded by actual mastery level achieved. If the teacher judgment and the actual results are identical " + " is recorded in the column as agreement; if not, the match is recorded as "0" in the agreement column.

8. Convert the total number of decisions that agree into a percentage. The percentage across all standards, strands, or assessments may be arranged for a total reliability calculation.

Example: Teacher Judgment and CRT Results*

*If CRT results are subjectively scored, teacher judgment must be made by someone other than the scorer.

Student	Teacher Judgment	CRT Assessment Results	Exact Agreement
1. Joe	Beginning	Progressing	-
2. Sue	Progressing	Progressing	+
3. Pam	Advanced	Advanced	+
4. Todd	Advanced	Advanced	+
5. Ned	Proficient	Proficient	+
		Total	4/5 = .80

- ✓ This an example using teacher judgment and CRT results. For each student a teacher makes a professional judgment about whether the student will score at the beginning, progressing, proficient, or advanced level.

- ✓ After the CRT results are compiled, the agreement between the teacher judgment and CRT results are calculated. Note that the mastery levels for the CRT were determined in advance and used to classify students' actual performance on the assessment.

Inter-rater Reliability (Used for subjectively scored assessments)

The inter-rater reliability method calculates a measure of consistency of scoring based on the decisions of two independent raters.

1. Subjectively scored assessments are scored with a rubric or clearly written criteria outlining specific expectations for assessment results.

2. The raters in this process must be thoroughly trained on the rubric and must be clear about the expectations of the assessment.

3. If the rubric has fewer than five score points only exact match decisions may be calculated.

4. Examples of student performances or products at all mastery levels (anchors or exemplars) should be also be used in the training of raters so that they know what the performance or product results look like at each level.

5. Raters score the assessments independently and record their scores independently.

6. The measure of rater agreement is calculated by determining how frequently the independent judgments of the raters agree about the level of proficiency on the assessment.

7. The final number of exact agreements are calculated and converted into a percentage.

8. This method can be used with a whole class or with a representative sampling of papers or performances.

9. The overall reliability is calculated by averaging the reliability across all standards, all strands, or all assessments.

Example: Calculating Inter-Rater Reliability

Student papers (or a representative sample) are scored by different raters. Make copies of the student assessments, one for each rater, or provide a cover sheet so that when papers are scored they are scored independently and neither rater can see any markings or scores of the other. See the table below for an example of how to record the results of the double scoring by raters.

	Name	Rater One Score	Rater Two Score	Level of Agreement
1.	Joe	4	4	+
2.	Sue	2	2	+
3.	Pam	1	3	0
4.	Tom	4	3	0
5.	Ned	1	1	+
			Total	3/5 = .60

Suggestions for Improving Reliability

Problem		Solution
Internal Consistency	"Outlier" (extreme score) in a distribution of student assessment scores	Delete outlier(s) provided deleted scores are less than 2% of all scores
	"Outlier" in a set of reliability coefficients	Consider using the median of reliability coefficients rather than the mean value of coefficient alpha or KR20
	Restriction of range (scores narrowly clustered together)	Add items that span a broader range of difficulty; consider "Decision Consistency" method if assessment already has a broad range of difficulty
	Low reliability (case 1; insufficient data points)	Add more assessment items to better sample the content; increase sample size of student observations if possible
	Low reliability (case 2; low quality items)	Conduct item analyses, delete poor performing items and replace with better items (e.g., use item discrimination index)
Problem		Solution
Decision Consistency	Inaccurate teacher predictions	Make predictions of student proficiency after instruction rather than before; ensure sufficient measurement opportunities for each achievement level
	Test-retest reliability low	Conduct both assessments within close temporal proximity
Problem		Solution
Inter-rater agreement	Low reliability among raters	Use a detailed "Training Protocol" with specific definitions of score points/proficiencies; include anchor performances as a validity check

Quality Criteria #6

The mastery levels are appropriately set.

This criterion is about determining "how good is good enough" in terms of levels of student achievement. It assures that mastery levels have been set appropriately, not arbitrarily, and that there is agreement about what the mastery levels mean.

- Districts should provide evidence that the student mastery levels were determined using procedures that take into account the difficulty of the items or tasks in the assessments or the classifications of students related to their achievement levels.
- The procedure used to set mastery levels should include systematic judgments about assessment content and the different levels of student performance.
- Professional judgment about students or about the assessment items/tasks need to be used to arrive at mastery level decisions.
- Documentation for meeting this criterion is required for all standards used for reporting.
- Mastery levels need to be recalculated when performance level descriptors, assessments, or student demographics change.

In meeting this criterion, districts should provide documentation to answer the following questions:

1. Who did the process?

Describe the person(s) or teams who led, conducted, and participated in the process for setting mastery levels.

- ✓ Include their qualifications.
- ✓ Indicate the grade levels represented and in what configurations.
- ✓ Include the number of teachers who participated in comparison to numbers of total staff.

2. What did they do in this process?

- ✓ A district should determine the appropriate method for setting mastery levels. *(See methods for Setting Mastery Levels, Attachment H and pages 28-32)*

- ✓ Mastery levels are appropriately set when three things are integrated in a process:
  - Performance Level Descriptors (PLDs)
  - professional judgment
  - actual student results

If districts compiled information from teachers or others and used forms or questionnaires to collect that data, include a copy of all forms that were used to include information from others.

The Establishment of Performance Level Descriptors (PLDs)

A group of teachers familiar with the content and the grade level of students being assessed can develop these descriptors. These teachers may be those who wrote or administered the assessments. Beginning with the 2009-10 administration the PLD's will be generated at the state level.

These performance level descriptors are the same set used for Criterion #1 *(The assessment match the standards*) to review for sufficiency and the same set that are used for Criterion #5 (*There is consistency of scoring*) if the Decision-Consistency Teacher Judgment method is used to calculate reliability.

To establish performance level descriptors:

- First, decide what a barely *proficient* student must know and be able to do to meet the standard.
- Next decide how the student will demonstrate proficiency.
- Then work backwards, so to speak, and decide what a *barely progressing* student would know and be able to do and how that would be demonstrated.
- Next decide what a *beginning* student would know and be able to do and how that would be demonstrated.
- Finally, decide what a barely *advanced* student would know and be able to do and how that would be demonstrated.

✓ Decide whether to use a student-based method or a test-based method for setting mastery levels.

Student-based Method

- Panelists must know the assessed students.
- Not recommended with fewer than 30 students being assessed
- Student-based methods are typically inappropriate in small schools unless multiple years of data are being calculated.

Test-based Method

- Panelists may or may not know the assessed students.
- May be used with any number of assessed students.
- Panelists need to have content knowledge and familiarity with students at the appropriate grade.

3. *What were the results of this process?*

Describe the results of the process for setting mastery levels.

- ✓ Explain the method(s) used. Provide a rationale for the method(s) used.
- ✓ Provide forms or worksheets that show the results and how they were derived.
- ✓ Provide mastery levels for all assessments used for reporting.

Methods for Setting Mastery Levels

Modified Contrasting Group Method (Student-based Method)

This method can be used for both objective and subjective items/tasks and is based upon the teacher knowing the students and their work. This method requires teacher professional judgment about students. The modified contrasting group method is not appropriate with small numbers of students (fewer than 30).

NOTE

If districts use a reliability method with professional judgment, such as decision consistency, districts should use a test-based method such as Angoff or Analytical Judgment for setting cut scores. If districts use professional judgment methods for both reliability and cut scores (i.e. decision consistency with modified contrasting group), then districts must make sure that the professional judgment used for one is independent from the professional judgment used for the other (different teachers – independent judgment.)

a. Make a list of the students to be assessed and the levels of proficiency that must be determined (i.e. beginning, progressing, proficient, advanced). (See Attachment I)

b. With the teachers who know the students, discuss and agree upon definitions of what student work would "look like" in each of those proficiency levels (e.g., what can progressing students do that beginning students cannot do?).

c. Prior to giving the assessment, but after the definitions are discussed, teachers will predict the level at which each student will score. (See Attachment J)

d. After the assessment results are in, the predictions are replaced by the actual student scores. (See Attachment K)

e. Compute the averages (means) of student scores for each proficiency level and place the average at the bottom of each column (See Attachment L). If there are extreme differences between scores, the median (middle score) may be used rather than the mean.

f. Determine the cut scores for each proficiency level by using the score that is the average midpoint between the means of adjacent groups. For example, the cut score for progressing will be the average of the means of the beginning and progressing

proficiency levels; the cut score for proficient will be the average of the means of the progressing and proficient performance levels; the cut score for advanced will be the average of the means of the proficient and advanced proficiency levels. To select the final cut score, the estimated cut score can be rounded down to the nearest whole number. (See Attachment L.)

g. With small numbers of students you may need to determine cut scores for two levels rather than four. This is done by collapsing the advanced and proficient as well as the beginning and progressing columns together, resulting in the *Met* and *Not Met* proficiency levels.

Modified Angoff Method (Test-based Method for Objectively Scored Items)

This method is best for objectively-scored assessments and works well with small numbers of students. Teachers participating must know both the test content and the characteristics of the students taking the assessment. This process consists of two rounds of activities. It is highly recommended that the facilitator of this process receive training on the modified Angoff method before using it in the district.

Round 1:

a Teachers who know the content agree upon definitions of what student performance would look like for each proficiency level (e.g., what can progressing students do that beginning students cannot do?). The panel of teachers discusses the content that is represented on the assessment items and the knowledge, skills, and abilities necessary to answer the items correctly with a focus on what distinguishes students in one proficiency level from the other proficiency levels.

b Teachers need to discuss what a Barely Progressing, Barely Proficient and Barely Advanced students look like. A Barely Progressing student is more like a beginning student except for the ability to do a few skills associated with the progressing student. The same definition applies to the Barely Proficient and Barely Advanced students.

c Looking at the assessment item by item, each panelist **independently** estimates the performance (score) of a Barely Proficient, Barely Progressing, and Barely Advanced students on one item at a time. For round 1 each teacher needs to complete a separate sheet to be collected by the leader for calculating the average cut scores and the range of cut scores for a Barely Proficient, Barely Progressing, and Barely Advanced student. (See Attachment M.)

Between Round 1 and Round 2:

a After completing round 1, estimated cut scores, range of cut scores, difficulty level, and percentage of students at each performance level are shared with the group and discussed. The estimated cut scores would be the average of the total scores for barely progressing, barely proficient, and barely advanced student for all teachers in round 1. Using the total scores for all teachers from round 1, the range of score points will be determined by using the minimum and maximum total scores for the barely

progressing, barely proficient, and barely advanced student. (See Attachments M and O.)

b After the assessments have been taken and scored, the difficulty level for each item needs to be calculated. For one point items, the difficulty level is the percentage of students getting the item correct. For multi point items, the difficulty level is the average score on the item divided by the total possible. Impact data should also include the percentage of students scoring at the beginning, progressing, proficient, and advanced levels. These values may not be very stable if they are based on a small sample of students. In this case only use the estimated cut scores and range of cut scores as impact data. If the values are based on a large sample of students (more than 30), use all four pieces of impact data to complete round 2. (See Attachment N)

Round 2:

a The panelists from round 1 meet to discuss the estimated cut scores, range of cut scores, difficulty level of each item, and the percentage of students scoring at each level. Panelists review the difficulty level and compare it to their initial item performance decision in round 1. Panelists look at the average cut score, the range of cut scores and the estimated cut score to determine the number of score points at each performance level. Using impact data about the percentage of students scoring at each level, the panelist discuss whether this matches with their view of this group of students. After discussing of the impact data, each panelist independently estimates the performance (score) of a Barely Proficient, Barely Progressing, and Barely Advanced students on one item at a time. (See Attachment O)

b The sum of the item predictions determines the final cut score for progressing, proficient, and advanced levels of performance on that assessment. The estimated cut score for the standard can be determined by finding the average of these assessment totals. The final cut score is determined from the estimated cut score and the range of cut scores from round 2. To select the final cut score the panel can round the estimated cut score down to nearest whole number or select the median cut score from the panelist's predictions.

Modified Analytical Judgment with Exemplars (Test-based Method for Subjectively Scored Items or Performance Assessments)

This method is best for performance assessments or assessments with multiple steps. It may be expanded to set multiple cut scores.

The following steps pertain to setting mastery levels when there are four proficiency levels.

a. For this method participants will examine a set of papers, products or performances that represent all score points or levels but for which the scores have been masked. These papers, products, or performances are known as exemplars. If there are more than two performance categories, then the number of papers, products, or performances at each

score point will be about the same across all score points except at the extreme low and high scores.

b. Form a panel of qualified teachers who know both the test content and the characteristics of the target students (e.g. beginning, proficient, proficient, and advanced).

c. Define what beginning, progressing, proficient, and advanced performance on the assessment means.

d. Through discussion panelists will agree on what the work of the beginning, progressing, proficient and advanced student will look like on the assessment.

e. Working with the set of exemplars, each panelist classifies them into four proficiency categories (e.g. beginning, progressing, proficient, advanced).

f. To determine the cut score for the progressing student,

- Each panelist identifies the three best papers, products or performances from the beginning category and the three poorest from the progressing category.
- Using these six papers, products or performances identified by each panelist in the beginning and progressing categories, calculate the average of the actual scores.
- Then calculate the average across all panelists. In other words, calculate an average of the averages.
- The answer becomes the final cut score for the progressing student.

g. To determine the cut score for the proficient student,

- Each panelist identifies the three best papers, products or performances from the progressing category and the three poorest from the proficient category
- For these six papers, products or performances identified by each panelist average the six papers, calculate the average of the actual scores.
- Then, calculate the average across all panelists. In other words, calculate an average of the averages.
- The answer becomes the final cut score for the proficient student.

h. To determine the cut score for the advanced student,

- Each panelist identifies the three best papers, products or performances from the proficient category and the three poorest from the advanced category
- For these six papers, products or performances identified by each panelist in the proficient and advanced categories, calculate the average of the actual scores.
- Then calculate the average across all panelists. In other words, calculate an average of the averages.
- The answer becomes the final cut score for the advanced student.

The following steps pertain to setting mastery levels when there are two proficiency levels.

a. For this method participants will examine a set of papers, products or performances that represent all score points or levels but for which the scores have been masked.

These papers, products, or performances are known as exemplars. Typically, if only two proficiency categories are being defined (met/not met) more papers in the middle range of scores will be used.

b. Form a panel of qualified teachers who know both the test content and the characteristics of the target students (i.e. beginning, proficient, and advanced).

c. Define what beginning, proficient, and advanced proficiency on the assessment means.

d. Panelists will discuss and agree on what the work of the beginning, proficient and advanced student will look like on the assessment.

e. Working with the 50 or more exemplars, each panelist separates them into three categories: beginning, proficient, advanced.

f. From this point on each panelist will work only with the papers, products or proficiencies identified as beginning or proficient.

g. Next each panelist identifies the three best papers from the group classified as beginning.

h. Then each panelist identifies the three poorest papers from the group classified as being proficient.

i. For the six papers identified by each panelist, take the average of the actual scores.

j. Finally, calculate the average across all panelists. In other words, calculate an average of the averages. The answer becomes the final cut score.

Rubric Standard Setting Method (Test-based method for subjectively scored tasks when student exemplars are not available)

a) Discuss each target student: Barely Advanced, Barely Proficient, and Barely Progressing. Discuss what their work will look like and what knowledge, skills, and abilities they will demonstrate at the respective performance level using good rubric-development language (e.g., specific, observable.)

b) Have each panelist independently write descriptions of the knowledge, skills, and abilities for students at each performance level (e.g., Barely Advanced, Barely Proficient, and Barely Progressing.)

c) Have the panelists discuss their independent descriptions for each performance level with the group to share the similarities and differences.

d) Document the final rubric's performance levels based on the panelists' consensus of these discussions. Note that score points may or may not be part of the rubric development.

e) After student exemplars become available through assessment administration, use the information to confirm or revise the descriptions to better characterize the performance of what students know and are able to do at a given performance level.

III. How Your Assessment Process Will Be Evaluated

A. The Nebraska-led Peer Review of STARS Process

- Review teams consisting of persons with assessment expertise will review each district's assessment process. These reviewers are Nebraska educators who have demonstrated experience with the STARS process and who have participated in extensive training in order to do this work. In addition, a team of external experts with expertise in assessment and measurement will assist review teams in arriving at the final assessment quality ratings.

- Each of six Quality Assessment Criteria as documented in your evidence of assessment quality will be rated as follows:

Met (no further comment necessary)
Met (some further comment necessary)
Needs Improvement
Not Met

- If any of the six Quality Assessment Criteria receive a *Met-some further comment necessary, Needs Improvement* or a *Not Met* rating, feedback will be provided about ways to strengthen performance on that criterion.

- Based on the total rating for the Quality Assessment Criteria, the overall assessment system for each grade level will be classified in one of five categories. Attachment P explains the classification rating system.
    - Exemplary
    - Very Good
    - Good
    - Needs improvement
    - Unacceptable

Nebraska Department of Education

The purpose of this review document is to assure that the assessment processes and procedures in local districts are of sufficient quality.

DISTRICT ASSESSMENT PORTFOLIO RUBRIC
2007--2008

6 Quality Criteria	Not Met	Needs Improvement	Met with Comment	Met
Criterion 1 **The assessments match the standards.**	• No qualifications of the independent reviewers are provided. • No evidence of an independent review for match to standards is provided (reviewers did not write the assessments). • No process for matching assessments to standards is described. • No results of the matching process are provided. • No sufficiency process is described. • No sufficiency results are provided (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) **Districts with local standards must designate a reading and a math standard.* o No consistency between criterion #1 and other criteria is found.	• Qualifications of the independent reviewers are unclear or incomplete. • Evidence of an independent review for match to standards unclear or incomplete (reviewers did not write the assessments). • The process for matching assessments to standards is unclear or incomplete. • Results of the matching process are unclear or incomplete. • Sufficiency process is unclear or incomplete. • Sufficiency results are unclear or incomplete (sufficiency required for both number of items/performances and levels of difficulty. Minimum 12 items or equivalent on reading standard 4.1.3, 8.1.1 and 12.1.1 and math standards 4.2.1, 8.2.2 and 12.2.1) **Districts with local standards must designate a reading and a math standard.* • Consistency between criterion #1 and other criteria is unclear or incomplete.	• Criterion has been fully met, but reviewer believes additional feedback would be helpful.	• 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) **Districts with local standards must designate a reading and a math standard.* • Consistency between Criterion #1 and other criteria is clear.

DISTRICT ASSESSMENT PORTFOLIO RUBRIC
2007--2008

6 Quality Criteria	Not Met	Needs Improvement	Met with Comment	Met
Criterion 2 **Students have an opportunity to learn.**	• No qualifications of the opportunity to learn reviewers are provided. • No process for opportunity to learn (both curriculum alignment and timing of assessment/ instruction) is described. • No results of the process for alignment of standards with local curriculum are provided. • No dates are provided when standards are taught. • No dates are provided when standards are assessed (80% of instruction should take place prior to assessment.) • No opportunity to learn information is provided for any standards. • No consistency between Criterion #2 and other criteria is found.	• Qualifications of the opportunity-to-learn reviewers are unclear or incomplete. • The process for opportunity to learn is unclear or incomplete (both curriculum alignment and timing of assessment / instruction is described.) • The results of the process for alignment of standards with local curriculum are unclear or incomplete. • Dates are provided when standards are taught but they are unclear or incomplete. • Dates are provided when standards are assessed but are unclear or incomplete • 80% of instruction should take place prior to assessment. • Opportunity to learn information provided for only some standards. • Consistency between Criterion #2 and other criteria is unclear or incomplete.	• Criterion has been fully met, but reviewer believes additional feedback would be helpful.	• Qualifications of the opportunity to learn reviewers are clear and complete. • The process for opportunity to learn is clear and complete (both curriculum alignment and timing of assessment/instruction) is described. • 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. • Opportunity to learn information provided for all standards. • Consistency between Criterion #2 and other criteria is clear and complete.

DISTRICT ASSESSMENT PORTFOLIO RUBRIC
2007--2008

6 Quality Assessment Criteria	Not Met	Needs Improvement	Met with Comment	Met
Criterion 3 **The assessments are free of bias and sensitive situations.**	• No qualifications of the bias reviewers are provided. • No bias orientation is described. • No process for bias review of assessment items is described. • No results of a bias review are provided. • No bias information provided for any standards (used for reporting). • No consistency between Criterion #3 and other criteria is found.	• Qualifications of the bias reviewers are unclear or incomplete. • The description of the bias orientation is unclear or incomplete. • The process for bias review of assessment items is unclear or incomplete. • Results of a bias review are unclear or incomplete. • Bias information provided only for some standards (used for reporting). • Consistency between Criterion #3 and other criteria is unclear or incomplete. .	o Criterion has been fully met, but reviewer believes additional feedback would be helpful.	• 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 complete. • Bias information provided for all standards (used for reporting). • Consistency between criterion #3 and other criteria is clear and complete.
Criterion 4 **The assessments are at the appropriate level.**	• No qualifications of the reviewers for appropriate level are provided. • No process for appropriate level review is described. • No results for the appropriate level review are provided. • Appropriate level information is not provided for any standards (used for reporting). • No consistency between Criterion #4 and other criteria is found.	• Qualifications of the reviewers for appropriate level are unclear or incomplete. • Process for appropriate level review is unclear or incomplete. • Results of the appropriate level review are unclear or incomplete. • Appropriate level information is provided only for some standards (used for reporting). • Consistency between Criterion #4 and other criteria is unclear or incomplete.	o Criterion has been fully met, but reviewer believes additional feedback would be helpful.	• 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. • Appropriate level information is provided for all standards (used for reporting) • Consistency between Criterion #4 and other criteria is clear and complete.

DISTRICT ASSESSMENT PORTFOLIO RUBRIC
2007--2008

6 Quality Assessment Criteria	Not Met	Needs Improvement	Met with Comment	Met
Criterion 5 **There is consistency of scoring.**	• No qualifications of the reliability process participants are provided. • No appropriate process for calculating reliability is described. • No reliability value is provided. (Minimum level of acceptable reliability is .70, mean or median, averaged across all standards.) • No procedure for improving reliability is provided. • Reliability is not reported for any standards (used for reporting). • No consistency between Criterion #5 and other criteria is found.	• Qualifications of the reliability process participants are unclear or incomplete. • Appropriate process for calculating reliability is unclear or incomplete. • Reliability value provided but calculations are below the minimum acceptable level. (Minimum level of acceptable reliability is .70, mean or median, averaged across all standards.) • Procedure for improving reliability is unclear or incomplete. • Reliability is reported for only some standards (used for reporting). • Consistency between Criterion #5 and other criteria is unclear or incomplete.	o Criterion has been fully met, but reviewer believes additional feedback would be helpful.	• 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. (Minimum level of acceptable reliability is .70, mean or median, averaged across all standards.) • Procedure for improving reliability is clear and complete. • Reliability is reported for all standards (used for reporting). • Consistency between Criterion #5 and other criteria is clear and complete.
Criterion 6 **The mastery levels are appropriately set.**	• No qualifications for mastery level participants are provided. • No evidence of mastery level process is provided. • No results of the mastery level process are provided. • Mastery level information is not provided for any of the standards (used for reporting). • No consistency between Criterion #6 and other criteria is found.	• Qualifications for mastery level participants are unclear or incomplete. • Evidence of a mastery level process is unclear or incomplete. • Results of the mastery level process are unclear or incomplete. • Mastery level information is provided for only some of the standards (used for reporting). • Consistency between Criterion #6 and other criteria is unclear or incomplete.	o Criterion has been fully met, but reviewer believes additional feedback would be helpful.	• 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. • Mastery level information is provided for all standards (used for reporting). • Consistency between criterion #6 and other criteria is clear and complete.

Match To Standard and Sufficiency Worksheet

Standard	Assessment Item	Match		Items/ Type	Beg.	Prog.	Prof.	Adv	Changes Needed
		Yes	No						

OPPORTUNITY TO LEARN

Standard	Dates Unit Taught	Assessment Item	Dates Assessed	Changes Needed

CRITERION THREE

ASSESSMENT HAS BEEN REVIEWED FOR BIAS

Standard	Assessment Item Examined	Changes Made

CRITERION FOUR

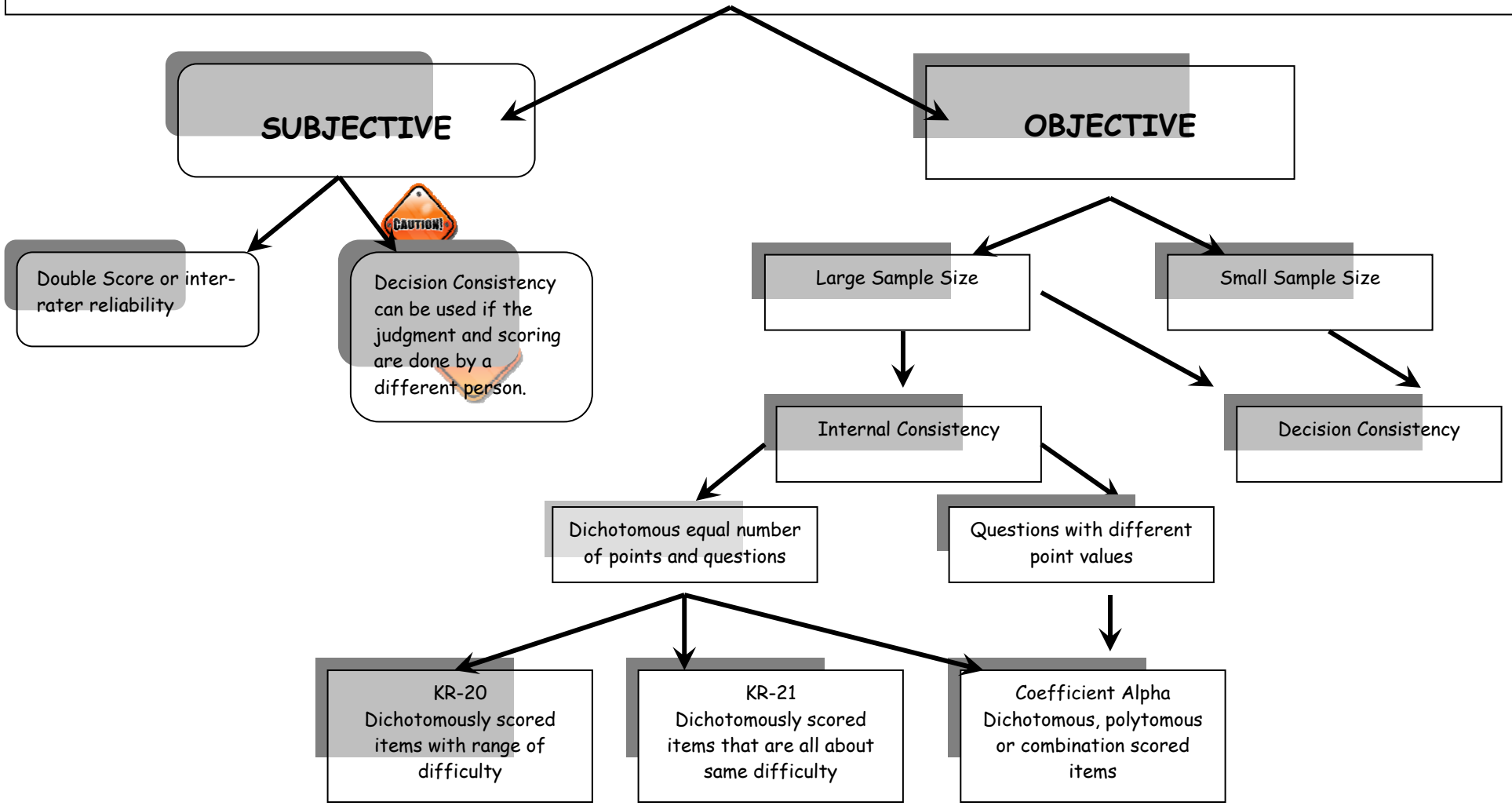
ASSESSMENTS ARE AT THE APPROPRIATE LEVEL

Standard	Assessment Used	Changes Needed

Quality Criterion #5
RELIABILITY

There is Consistency of Scoring.

The mean or median of all assessments must be greater than 0.70

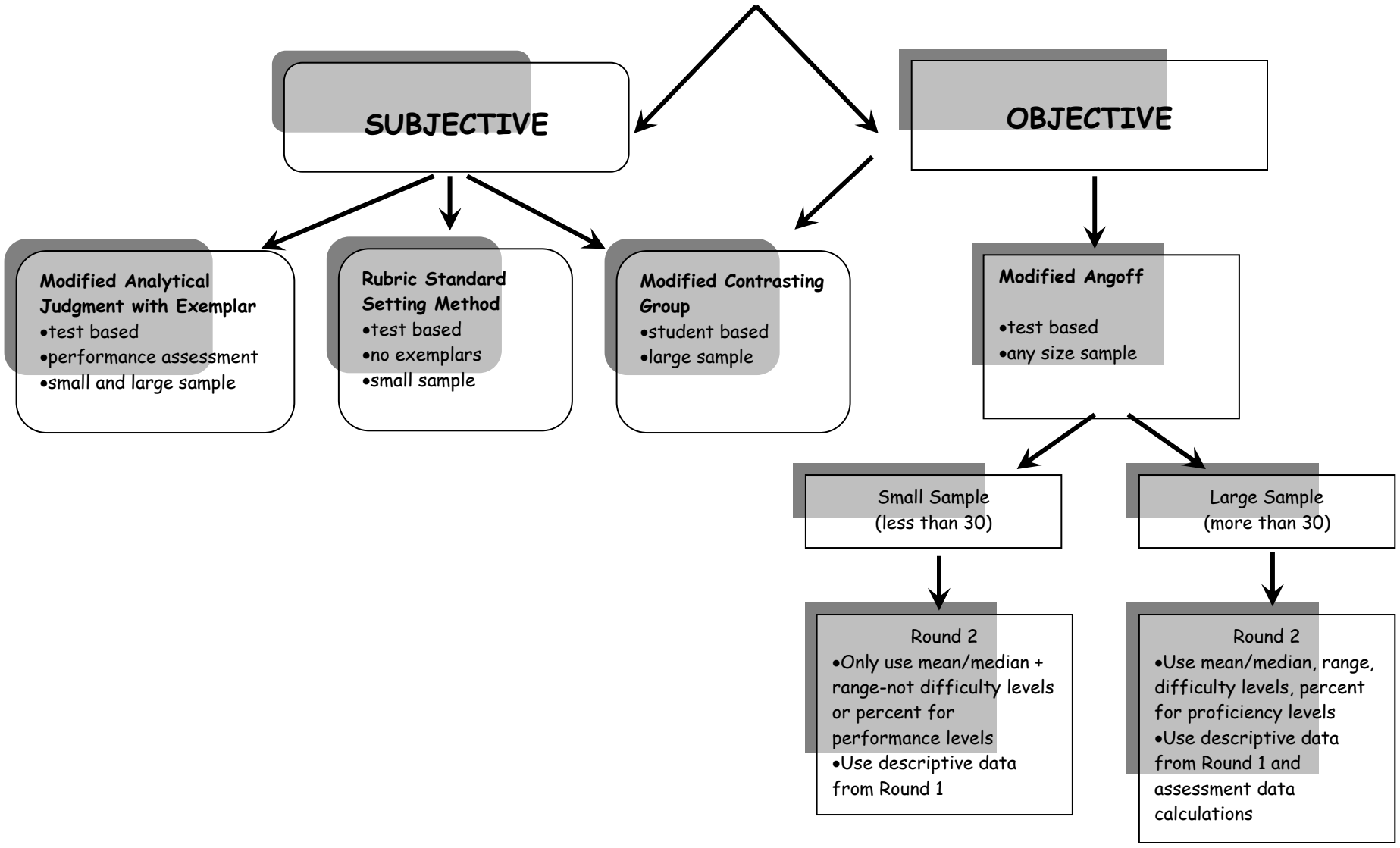


TEACHER JUDGMENT WORKSHEET

Student Name	Tchr Judge	Results	Agree	Tchr Judge	Results	Agree	Tchr Judge	Results	Agree	Tchr Judge	Results	Agree	Tchr Judge	Results	Agree	Total Agree

Quality Criterion #6
MASTERY LEVELS

Mastery Levels are appropriately set.



Attachment I

Modified Contrasting Group Method

Performance Level Descriptors	
Level	**Definition**
Beginning	
Progressing	
Proficient	
Advanced	

Student Name	Beginning	Progressing	Proficient	Advanced
Student One	_____	_____	_____	_____
Student Two	_____	_____	_____	_____
Student Three	_____	_____	_____	_____
Student Four	_____	_____	_____	_____
Student Five	_____	_____	_____	_____
Student Six	_____	_____	_____	_____
Student Seven	_____	_____	_____	_____
Student Eight	_____	_____	_____	_____
Student Nine	_____	_____	_____	_____
Student Ten	_____	_____	_____	_____
Student Eleven	_____	_____	_____	_____
Student Twelve	_____	_____	_____	_____
Student Thirteen	_____	_____	_____	_____
Student Fourteen	_____	_____	_____	_____
Student Fifteen	_____	_____	_____	_____

Modified Contrasting Group Method

Performance Level Descriptors	
Level	Definition
Beginning	
Progressing	
Proficient	
Advanced	

Student Name	Beginning	Progressing	Proficient	Advanced
Student One			XX	
Student Two	XX			
Student Three				XX
Student Four		XX		
Student Five				XX
Student Six		XX		
Student Seven			XX	
Student Eight	XX			
Student Nine			XX	
Student Ten		XX		
Student Eleven			XX	
Student Twelve	XX			
Student Thirteen			XX	
Student Fourteen		XX		
Student Fifteen				XX

Modified Contrasting Group Method

Performance Level Descriptors	
Level	Definition
Beginning	
Progressing	
Proficient	
Advanced	

Student	Beginning	Progressing	Proficient	Advanced
Student One	_____	_____	86	_____
Student Two	71	_____	_____	_____
Student Three	_____	_____	_____	86
Student Four	_____	67	_____	_____
Student Five	_____	_____	_____	98
Student Six	_____	80	_____	_____
Student Seven	_____	_____	83	_____
Student Eight	48	_____	_____	_____
Student Nine	_____	_____	78	_____
Student Ten	_____	74	_____	_____
Student Eleven	_____	_____	88	_____
Student Twelve	55	_____	_____	_____
Student Thirteen	_____	_____	84	_____
Student Fourteen	_____	63	_____	_____
Student Fifteen	_____	_____	_____	95

Modified Contrasting Group Method

Performance Level Descriptors	
Level	Definition
Beginning	
Progressing	
Proficient	
Advanced	

Student Name	Beginning	Progressing	Proficient	Advanced
Student One			86	
Student Two	71			
Student Three				86
Student Four		67		
Student Five				98
Student Six		80		
Student Seven			83	
Student Eight	48			
Student Nine			78	
Student Ten		74		
Student Eleven			88	
Student Twelve	55			
Student Thirteen			84	
Student Fourteen		63		
Student Fifteen				95
Mean (Average):	58	71	84	91
Cut Scores:	**64.5 ≈ 65**	**77.5 ≈ 78**	**87.5 ≈ 88**	

* Average between the adjacent groups.

Modified Angoff Method – Round 1

Item	Barely Progressing	Barely Proficient	Barely Advanced	Total Possible
1	1	1	1	1
2	0	1	1	1
3	0	0	1	1
4	0	0	0	1
5	0	0	0	1
6	1	2	3	3
7	0	1	2	3
8	0	0	1	3
9	1	1	2	3
10	0	2	2	3
Estimated Total Right	3	8	13	20

Impact Data	Beginning	Progressing	Proficient	Advanced
Average Cut Scores		4.1	8.3	14.8
Range of Cut Scores		3-5	8-10	13-17
Cut Score		4	9	15
Score Points	0-3	4-8	9-14	15-20

Modified Angoff Method Example (Impact Data)

A. Mean / Median

Mean – average score of students taking assessment

$$\text{Mean} = \frac{\text{total of all scores}}{\text{number of students}} = \frac{345}{25} = 13.8$$

Median – the score that divides the list of scores exactly in half. To determine median list scores from largest to smallest and find the score that is halfway (add number of scores.) With even number, find the two scores exactly in the middle and the median is the average of the two middle scores.

B. Range of Scores - the largest score to the smallest score.

Ex: Range of 21 to 23

C. Difficulty Level (DL)

One point items, difficulty level is the percentage of students getting item correct (represent as two-digit decimal)

$$\text{DL Item \#1:} \quad \frac{\text{\# of students getting item correct}}{\text{total \# of students}} = \frac{24}{25} = .95$$

Items with more than one point, difficulty level is the average score on the item divided by the total possible score (represent as two-digit decimal.

$$\text{DL Item \#6} \quad \frac{\text{Average score on item}}{\text{total possible score}} = \frac{2.4}{3} = .80$$

D. Percentage of Students Scoring at Each Performance Level

Using the cut scores for each performance level from round 1, the percentage of students scoring at each level can be determined. In the example, a total of 25 students are used.

Performance Level	Score Points	No. of Students	Percentage of Students
Advanced	15-20	10	40%
Proficient	9-14	10	40%
Progressing	4-8	4	16%
Beginning	0-3	1	4%
		25 (total)	

Modified Angoff Method Example

Item	Barely Progressing	Barely Proficient	Barely Advanced	Total Possible	Difficulty Level
1	1	1	1	1	.95
2	0	1 (0)	1	1	.65
3	0	0	1	1	.60
4	0	0	0	1	.25
5	0	0	0 (1)	1	.45
6	1 (2)	2 (3)	3	3	.80
7	0	1	2	3	.40
8	0	0 (1)	1	3	.67
9	1	1	2	3	.60
10	0	2	2 (3)	3	.33
Estimated Total Right	3 (4)	8 (9)	13 (15)	20	

(Possible changes in parenthesis)

Impact Data (round 2)	Beginning	Progressing	Proficient	Advanced
Average Cut Scores		4.1 (5.2)	8.3 (10.1)	14.8 (16.2)
Range of Cut Scores		3-5 (4-6)	8-10 (9-12)	13-17 (15-19)
Estimated Cut Score		4 (6)	9 (11)	15 (16)
Score Points	0-3 (0-5)	4-8 (6-11)	9-14 (11-15)	15-20 (16-20)

(Changes for round 2 are shown in parenthesis.)

QUALITY CRITERIA RATING CHART
FOR THE DISTRICT ASSESSMENT PORTFOLIO

Each grade level portfolio from the district will receive one of five ratings:

Quality Criteria for Assessment	Exemplary	Very Good	Good	Needs Improvement	Unacceptable
1. Matches Standards	Met	Met	Met	Met	Not Met
					or
2. Opportunity to Learn	Met	Met	Met	Met	Not Met
3. Bias Review	Met	Met	Met	Not Met	Not Met
			or		
4. Appropriate Level	Met	Met	Met	Not Met	Not Met
5. Score Consistency	Met	Met	Met	Not Met	Not Met
		or	or		
6. Mastery Levels	Met	Met	Met	Not Met	Not Met

Districts may receive one of four comments:

1) "Met"
2) "Met some further comment necessary"
3) "Needs Improvement"
4) "Not Met"

Each grade level portfolio from the district will receive one of five ratings:

Exemplary	The district has met all 6 quality assessment criteria.
Very Good	The district has met quality assessment criteria 1, 2, 3, 4, and either 5 or 6.
Good	The district has met quality assessment criteria 1 and 2 and either 3 or 4 as well as either 5 or 6.
Needs Improvement	Districts have met criteria 1, 2, 3, 4 and not met either 5 or 6. or The district has met criteria 1 and 2 and any one of the other four quality assessment criteria. or The district has met only quality assessment criteria 1 and 2. or The district has met criteria 1, 2,5 and 6, and not met criteria 3 or 4.
Unacceptable	The district has met only one of either quality assessment criteria 1 or 2. or The district has not met either quality assessment criteria 1 or 2. or The district has not submitted a portfolio.

IV. Nebraska-led Peer Review of STARS Appeals Process

The Appeals Process

There will be an appeals process for the results of the Nebraska-led Peer Review for Part I as this is the only "rated" portion of the review. This appeals process is currently in place with the District Assessment Portfolio Review and the Statewide Writing Assessment Process. Appeals forms are included in STARS Update# 21 as Attachment F. Districts will have 10 days from the notification of their rating to file a request for appeal. A review team will be convened to review the appeal and the evidence presented to the Nebraska Department of Education. Districts will be notified of the results of the appeal and any accompanying actions that a district needs. Any needed improvements or corrections a district may need to make will need to be completed before June 30, 2007 if the Peer Review occurred in the fall of 2006. Districts that are reviewed after January 1, 2007 will have until September 30, 2007 to make the needed improvements. The rating classification for those districts will be marked as "Continuing Review" on the State of the Schools Report until the needed changes have been received and approved.

Appeals Process
Nebraska-led Peer Review of STARS

An appeal/resubmission form (Attachment Q) should have been submitted for any district that

a) received a "Continuing Review" rating during the Peer Review of STARS.

b) wants to raise the assigned rating to a higher level.

The appeal/resubmission forms were to be faxed to the Statewide Assessment Office according to the following schedule:

Review Week	Appeals Window	Appeal Due	Evidence Due On or Before
Oct. 30-Nov. 2, 2006	Nov. 20-Dec. 1, 2006	December 1, 2006	June 30, 2006
Jan. 22-26, 2007	Febr. 12-23, 2007	February 23, 2007	September 30, 2007
March 5-9, 2007	Mar. 26-Apr. 6, 2007	April 6, 2007	September 30, 2007
April 23-27, 2007	May 14-25, 2007	May 25, 2007	September 30, 2007

Procedure for Resubmission

Two review sessions will be held for the re-examination of evidence. The first will be in July, and the second will be in early October.

Although districts must resubmit by either June 30, 2007 or September 30, 2007 as outlined above, districts may submit prior to the assigned due date. In fact, districts are encouraged to complete the evidence and submit at their earliest convenience.

Feedback and revised ratings from the July review will be available to districts by August 1 during the 10 day window used for reviewing STARS and AYP data.

Feedback and revised ratings from the October review will be available to districts by mid-October during a second 10-day window used as a "second look" at STARS and AYP data.

To resubmit evidence, districts should do the following:

1) Materials are to be sent by the individual district.

2) Send a copy of the two-page previously submitted appeal/resubmission form along with the evidence required by the reviewer feedback. Districts are encouraged to send only the evidence specified in the reviewer comments. Do not send the entire portfolio.

3) If districts are submitting evidence or clarification for more than one criterion, organize evidence by criterion number, clearly labeled.

4) Materials should be hard copy as NDE will be filing them.

5) The materials may be mailed to: (please do not fax)

Nebraska Department of Education
STATEWIDE ASSESSMENT
301 Centennial Mall South, PO Box 94987
Lincoln, NE 68509
Phone: 402 471-2495

Nebraska Department of Education

DISTRICT ASSESSMENT DOCUMENTATION APPEAL FORM

(Complete a form for each re-review requested.)

This form is a request for a re-examination of specified criteria documenting quality of the District Assessment Documentation.

Date _____

SCHOOL DISTRICT	COUNTY DISTRICT NUMBER
SUPERINTENDENT	SIGNATURE
LOCAL ASSESSMENT CONTACT if different from superintendent:	SIGNATURE
SCHOOL ADDRESS	CITY, ZIP
PHONE	FAX
EMAIL:	Portfolio Grade Level:

The appeals process may occur only between _____

Please indicate the date new evidence will be resubmitted.

_____

Return to:

Nebraska Department of Education
STATEWIDE ASSESSMENT
301 Centennial Mall South
Lincoln, NE 68509-4987
Fax : 402 471-4311
Phone 402 471-2495

School District Name: **Grade Level:**

INSTRUCTIONS:
If your district received a "not met" or a "needs improvement" on a criterion that would change the assessment quality rating and your district wants to appeal the rating, check the box of the criterion to be re-examined. A re-examination will be conducted only on criteria for which you have listed a reason for appeal. Note: A review of any criterion marked "Met with comment" will NOT change the rating.

CRITERION #	REASON FOR APPEAL
☐ 1. Assessments reflect state or local standards.	
☐ 2. Students have had an opportunity to learn the content.	
☐ 3. Assessments are free from bias or offensive situations.	
☐ 4. Assessment levels are appropriate for students.	
☐ 5. There is consistency of scoring.	