

# **REPORT**

## **Alignment Analysis of Reading Extended Indicators and Operational Alternate Reading Assessment**

**Nebraska  
Grades 3-8 and 11  
2010**

**Norman L. Webb**

**October 3, 2010**



# **REPORT**

## **Alignment Analysis of Reading Extended Indicators and Operational Alternate Reading Assessment**

**Nebraska  
Grades 3-8 and 11  
2010**

**Norman L. Webb**

**October 3, 2010**

## Acknowledgements

### *Reviewers*

Cindy Jacobson	Group Leader	Wisconsin
Barth Wepking		Wisconsin
Anna Hardway		West Virginia
Lisa Sireno		Missouri
Tricia Parker		Nebraska
Judith Ruskamp		Nebraska
Amy Hill		Nebraska
Joy Martin		Nebraska

The Nebraska Department of Education funded this analysis. Jan Hoegh, Assistant Director of Statewide Assessment, was the person responsible for the study and the main contact person. Dr. Pat Roschewski, Director of Statewide Assessment, had the overall responsibility for the study.

## Table of Contents

Executive Summary .....	v
Introduction.....	1
Alignment Criteria Used for This Analysis .....	3
Categorical Concurrence.....	3
Depth-of-Knowledge Consistency.....	4
Range-of-Knowledge Correspondence.....	7
Balance of Representation .....	8
Source of Challenge.....	9
Findings.....	9
Standards.....	9
Alignment of Curriculum Standards and Assessments.....	11
Source-of-Challenge Issues and Reviewers’ Comments .....	17
Reliability Among Reviewers.....	18
Summary .....	19
References.....	20
Appendix A Nebraska Extended Reading Standards and Group Consensus DOK Values 2010 Grades 3-8 and 12	
Appendix B Data Analysis Tables Nebraska State Accountability-Alternate Assessment of Reading 2010 Grades 3-8 and 11	
Appendix C Reviewers’ Notes and Source-of-Challenge Comments Nebraska State Accountability- Alternate Assessment of Reading 2010 Grades 3-8 and 11	
Appendix D Debriefing Summary Notes Nebraska State Accountability-Alternate Assessment of Reading 2010 Grades 3-8 and 11	



## Executive Summary

A three-day alignment institute was held in Lincoln, Nebraska on September 20-22, 2020. Eight reviewers, four from Nebraska and four from other states, analyzed the Nebraska Student Accountability (NeSA) reading alternate assessment and the two reading standards (vocabulary and comprehension) and extended indicators for grades 3-8 and 10 applicable to students with significant disabilities. The assessments had been given in the spring of the 2009-2010 school year. Reviewers were reading specialist, special education teachers, and assessment specialists. One assessment of 25 items was analyzed for each grade. Six Extended Depth of Knowledge (EDOK) stages were used to identify the different levels of content complexity.

The alignment between the NeSA reading alternate assessment and standards for each grade was found to be acceptable. Fewer than five items would need to be added or replaced on each assessment to attain full alignment (see the summary table below). In the summary table, 100% indicates that the alignment criterion was acceptable for both standards. For grades 4 and 5, the standards and the assessments were found to be fully aligned. The main alignment issue was with the Depth-of- Knowledge consistency for grades 3, 6, 7, 8, and 11 and the Categorical Concurrence for grades 6 and 8. For grades 3, 6, and 7, the vocabulary items on the assessment had less than 50% of the items with an EDOK stage that was the same or higher than the EDOK stage of the assigned extended indicator. For the higher grades, 8 and 11, the comprehension items did not satisfy this criterion. For grades 6 and 8, reviewers also did not find six or more items that they judged clearly assessed the extended indicators under the vocabulary standard.

Even though the alignment for each grade was found to be acceptable when the minimum acceptable levels were considered, reviewers noted a number of ways that the assessments or the extended indicators could be improved. At least one or two items on the assessment for each grade were found uncodeable. Reviewers indicated that these items did not match any of the standards or extended indicators. For example, reviewers indicated that one grade 5 item measured students' knowledge of life skills rather than any reading skill. For all but grade 5, reviewers found at least one item that did not precisely target an extended indicator, but did in general measure knowledge under one of the standards. These items that were coded as targeting a generic extended indicator suggest possible content that was not included in the extended indicators and indicate the need to reconsider the extended indicators. Reviewers wrote notes about items that generally measured content related to an extended indicator, but did not target the content that was the main intent of the extended indicator. In addition for each grade, two or three reviewers identified a source of challenge for two to seven items. Thus, the alternate assessments and the extended indicators met the basic requirements to be considered acceptably aligned. However, there are ways that the assessments and standards could be improved to strengthen the relationship between both along with increasing the quality of the assessments in general.

Summary Table

*Percent of Nebraska Reading Standards with Acceptable Level on Each Alignment Criteria for the 2010 NeSA Reading Alternate Assessments*

Grade (N=25 items on each assessment)	<i>Categorical Concurrence</i> (Total number of items for the one stand)	<i>Depth-of-Knowledge Consistency</i> (50% at/above)	<i>Range of Knowledge</i> (50% of objectives)	<i>Balance of Representation</i> (without possible weakness)	<i>Estimated Range of Items per assessment to be Added or Replaced for Full Alignment</i>
3	100%	50%	100%	100%	2
4	100%	100%	100%	100%	0
5	100%	100%	100%	100%	0
6	50%	50%	100%	100%	2
7	100%	50%	100%	100%	2
8	50%	50%	100%	100%	4-5
11	100%	50%	100%	100%	2

Categorical Concurrence >6 items  
 Depth-of-Knowledge >50% with EDOK stage the same or higher than level of corresponding content expectation  
 Range-of-Knowledge >50% of content expectations under a strand  
 Balance of Representation A possible weakness if one or more content expectation with a relative large number of items (e.g. five or more than the content expectation with the next highest number of items)

# **Alignment Analysis of Reading Extended Indicators and Operational Alternate Reading Assessment**

## **Nebraska Grades 3-8 and 11 2010**

**Norman L. Webb**

### **Introduction**

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute of any one of these two system components. Alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997).

A three-day alignment institute was held September 20-22, 2010, in Lincoln, Nebraska at the Country Inn and Suites. The purpose of the institute was to analyze the Nebraska extended indicators for reading and the 2010 alternate assessment forms for grades 3-8 and 11. One group of eight reviewers participated in the institute. Four of the reviewers were from Nebraska and four were from other states. External reviewers from other states are likely to provide a more objective view of the standards and the assessments because they have no invested interest in either. The two of the four external reviewers who participated in this analysis had participated in similar alignment studies for over eight states. Their experiences gave them a more global perspective. The Nebraska reviewers also provided important knowledge for the process. The in-state reviewers were more aware of the interpretation of the standards and extended indicators by Nebraska teachers and how the indicators are used in classrooms. The balance between external reviewers and Nebraska reviewers provided a more objective view along with important knowledge of practices from the state. The reviewers included reading content experts, special education teachers, reading teachers, and reading assessment experts.

Nebraska used the terminology *standards and extended indicators* that were applicable to students with significant disabilities. The extended indicators were adopted to be used in the 2009-2010 school year. For each grade, the reading expectations had

two standards—vocabulary and comprehension. For students with significant disabilities, these standards were further delineated by extended indicator each one written to correspond to an indicator. For each grade (3-8 and 12), the vocabulary standard had three extended indicators and the comprehension standard had eight or nine extended indicators. This analysis was done by mapping the assessment items to the extended indicators.

As part of the alignment institute, reviewers were trained to identify the extended depth of knowledge (EDOK) stages of the standards and assessment items. This training included reviewing the definitions of the six EDOK stages and reviewing examples of each. The EDOK stages assigned to each extended indicator were determined through a consensus process. For each grade, reviewers individually assigned an EDOK stage to each extended indicator. Then the group leader facilitated a process leading the group to a consensus EDOK stage for extended indicator. After analyzing four or five items as a group, the reviewers individually analyzed the assessment items in a form by assigning each item a EDOK stage and to the most appropriate extended indicator. Following individual analyses of the items, reviewers participated in an adjudication discussion by reviewing the coding for individual items with a large variance in either the assigned EDOK or content expectation.

To derive the results from the analysis, the reviewers' responses were averaged. Any variance among reviewers was considered legitimate, with the true EDOK stage for the item falling somewhere between the two or more assigned values. Such variation could signify a lack of clarity in how the standards and indicators were written, the robustness of an item that could legitimately correspond to more than one standard and/or an EDOK that falls in between two of the six defined levels. Reviewers adjudicated their results after each grade level analysis. The adjudication process included the discussion of any results in assigning items to standards without a majority of reviewers in agreement. Reviewers were not required to change their results after the discussion. Large variations among reviewers in the final results represented true differences in opinion among the reviewers and not because of coding error. These differences could be explained by different content expectations targeting the same content knowledge or for an item that did not explicitly correspond to any content expectation, but could be inferred to relate to a content expectation. Reviewers were allowed to identify one assessment item as corresponding to up to three content expectations—one primary hit (expectation) and up to two secondary hits. However, reviewers could only code one EDOK stage to each assessment item, even if the item corresponded to more than one content expectation.

Reviewers were instructed to focus primarily on the alignment between the state's standards, extended indicators, and assessments. However, reviewers were encouraged to offer their opinions on the quality of the standards, extended indicators, or of the assessment activities/items by writing a note about the item. Reviewers also could indicate whether there was a source-of-challenge issue with the item—i.e., a problem with the item that might cause the student who knows the material to give a wrong

answer or enable someone who does not have the knowledge being tested to answer the item correctly.

The results produced from the institute pertain only to the issue of alignment between the Nebraska reading extended indicators and the assessments. Note that an alignment analysis of this nature does not serve as external verification of the general quality of the state's standards, indicators, or assessments. Rather, only the degree of alignment is discussed in the results. For these results, the means of the reviewers' coding were used to determine whether the alignment criteria were met. Standard deviations are reported in the tables provided in Appendix B, which give one indication of the variance among reviewers.

This report describes the results of an alignment study of the reading Nebraska Student Accountability (NeSA) alternate assessment reading forms for grades 3-8 and 11. The study addressed specific criteria related to the content agreement between the state's standards and assessments. Four criteria received major attention: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence and balance of representation.

### **Alignment Criteria Used for This Analysis**

This analysis judged the alignment between the standards and the assessments on the basis of four criteria. Information is also reported on the quality of items by identifying items with sources-of-challenge and other issues. For each alignment criterion, an acceptable level was defined by what would be required to assure that a student had met the standards.

#### **Categorical Concurrence**

An important aspect of alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. *The criterion of categorical concurrence between standards and assessments is met if the same or consistent categories of content appear in both documents.* This criterion was judged by determining whether the assessment included items measuring content from each strand. The analysis assumed that the assessment had to have at least six items for measuring content from a strand in order for an acceptable level of categorical concurrence to exist between the strand and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors have to be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63. This indicates that about 63% of the group would be consistently classified as masters or nonmasters if two equivalent test

administrations were employed. The agreement coefficient would increase if the cutoff score is increased to one standard deviation from the mean to .77 and, with a cutoff score of 1.5 standard deviations from the mean, to .88. Usually states do not report student results by strands or require students to achieve a specified cutoff score on expectations related to a strand. If a state did do this, then the state would seek a higher agreement coefficient than .63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a strand, and as a basis for making some decisions about students' knowledge of that strand. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

### **Depth-of-Knowledge Consistency**

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the strands, as judged in this analysis, at least 50% of the items corresponding to a strand had to be at or above the depth-of-knowledge level of the corresponding standards: 50%, a conservative cutoff point, is based on the assumption that a minimal passing score for any one strand of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the corresponding standards. For example, assume an assessment included six items related to one strand and students were required to answer correctly four of those items to be judged proficient—i.e., 67% of the items. If three, 50%, of the six items were at or above the depth-of-knowledge level of the corresponding expectations, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the depth-of-knowledge level of one expectation. Some leeway was used in this analysis on this criterion. If a strand had between 40% and 50% of items at or above the depth-of-knowledge levels of the expectations, then it was reported that the criterion was “weakly” met.

*Interpreting and assigning depth-of-knowledge levels to both expectations within strands and assessment items are essential requirements of alignment analysis.* The reading levels are based on Valencia and Wixson (2000, pp. 909-935). These descriptions help to clarify what the different levels represent in reading:

#### *Reading*

The Extended Depth of Knowledge Levels for Special Education (EDOK) partitions the first DOK level (recall and recognition) into three stages—respond, reproduce, and recall. Stages 4, 5, and 6 are the same as DOK levels 2, 3, and 4. The

EDOKs were developed by Gary Cook and others. These descriptions help to clarify what the different levels represent in reading:

### *Stage 1 Respond*

Requires the ability to respond to or indicate, or acknowledge text- or discourse-related features. Example:

- ◆ Points to the letters/words/pictures on a page
- ◆ Points to letters or words
- ◆ Acknowledges a discourse interaction with an interlocutor
- ◆ Responds to a conversations
- ◆ Acknowledges someone signing
- ◆ Attends to text

### *Stage 2 Reproduce*

Requires the ability to copy, replicate, repeat, re-enact, mirror, or match text- or discourse-related features. Example:

- ◆ Copies letters
- ◆ Reproduces letters, text, or words either verbally or through writing/signing
- ◆ Matches sound/sound
- ◆ Matches letter/letter
- ◆ Matches words
- ◆ Re-enacts a story or interaction either verbally or through text (also drawing)
- ◆ Matches picture/picture
- ◆ Matches symbol/symbol

### *Stage 3 Recall and Recognition*

Requires the ability to recite or recall facts or information. Involves the ability to distinguish between text-based or discourse features. Example:

- ◆ Identifies pictures of objects (animate or inanimate) through verbal cues or text-based cues
- ◆ Identifies details in text
- ◆ Identifies correct spelling or meaning of words
- ◆ Identifies letters
- ◆ Identifies sounds
- ◆ Identifies figurative language
- ◆ Uses dictionary

### *Stage 4 (DOK Level 2, Basic Reasoning)*

Requires processing beyond recall and observation. Requires both comprehension and subsequent processing of text. Involves ordering, classifying text, as well as identifying patterns, relationships, and main points. Examples:

- ◆ Uses context to identify unfamiliar words
- ◆ Predicts logical outcome
- ◆ Identifies and summarizes main points
- ◆ Associates/identifies letters with sounds
- ◆ Indicates what comes next in a story

*Stage 5 (DOK Level 3, Complex Reasoning)*

Requires students to go beyond text. Requires students to explain, generalize, and connect ideas. Involves inferencing, prediction, elaboration, and summary. Requires students to support positions using prior knowledge and to manipulate themes across passages. Examples:

- ◆ Determines effect of author’s purpose on text elements
- ◆ Summarizes information from multiple sources
- ◆ Critically analyzes literature/text
- ◆ Expresses an opinion about text citing evidence to support reasoning

*Stage 6 (DOK Level 4, Extended Reasoning)*

Requires extended higher-order processing. Typically requires extended time to complete task, but time spent not on repetitive tasks. Involves taking information from one text/passage and applying this information to a new task. May require generating hypotheses and performing complex analyses and connections among texts. Examples:

- ◆ Analyzes and synthesizes information from multiple sources
- ◆ Examines and explains alternative perspectives across sources
- ◆ Describes and illustrates common themes across a variety of texts
- ◆ Creates compositions that synthesize, analyze, and evaluate

These extended DOK stages were derived and are compatible with the four DOK levels used for the general population. The DOK reading levels are based on Valencia and Wixson (2000, pp. 909-935). These descriptions help to clarify what the different levels represent in reading. The more detailed definitions for the reading DOK levels include:

*Reading Level 1.* Level 1 requires students to receive or recite facts or to use simple skills or abilities. Oral reading that does not include analysis of the text as well as basic comprehension of a text is included. Items require only a shallow understanding of the text presented and often consist of verbatim recall from text, or simple understanding of a single word or phrase. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Support ideas by reference to details in the text.
- Use a dictionary to find the meanings of words.
- Identify figurative language in a reading passage.

*Reading Level 2.* Level 2 includes the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Inter-sentence analysis or inference is required. Some important concepts are covered but not in a complex way. Standards and items at this level may include words such as summarize, interpret, infer, classify, organize, collect, display, compare, and determine whether fact or opinion. Literal main ideas are stressed. A Level 2 assessment item may require students to apply skills and concepts that are covered in Level 1. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Use context cues to identify the meaning of unfamiliar words.
- Predict a logical outcome based on information in a reading selection.
- Identify and summarize the major events in a narrative.

*Reading Level 3.* Deep knowledge becomes a greater focus at Level 3. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas. Standards and items at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students' application of prior knowledge. Items may also involve more superficial connections between texts. Some examples that represent, but do not constitute all of, Level 3 performance are:

- Determine the author's purpose and describe how it affects the interpretation of a reading selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

*Reading Level 4.* Higher-order thinking is central and knowledge is deep at Level 4. The standard or assessment item at this level will probably be an extended activity, with extended time provided for completing it. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking. Students take information from at least one passage of a text and are asked to apply this information to a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among texts. Some examples that represent, but do not constitute all of, Level 4 performance are:

- Analyze and synthesize information from multiple sources.
- Examine and explain alternative perspectives across a variety of sources.
- Describe and illustrate how common themes are found across texts from different cultures.

### **Range-of-Knowledge Correspondence**

For standards and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a strand is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer*

*the assessment items/activities.* The criterion for correspondence between span of knowledge for a strand and an assessment considers the number of standards within the strand with one related assessment item/activity. Fifty percent of the expectations for a strand had to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a strand. This assumes that each expectation for a strand should be given equal weight. Depending on the balance in the distribution of items and the need to have a low number of items related to any one expectation, the requirement that assessment items need to be related to more than 50% of the expectations for an strand increases the likelihood that students will have to demonstrate knowledge on more than one expectation per strand to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the expectations. However, any restriction on the number of items included on the test will place an upper limit on the number of expectation that can be assessed. Range-of-knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of strands and a large number of expectations. If 50% or more of the expectations for a strand had a corresponding assessment item, then the range-of-knowledge correspondence criterion was met. If between 40% and 50% of the expectation for a strand had a corresponding assessment item, the criterion was “weakly” met.

### **Balance of Representation**

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range-of-knowledge criterion only considers the number of expectations within a strand hit (a standard with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among these expectations. *The balance-of-representation criterion is used to indicate the degree to which one standard is given more emphasis on the assessment than another.* An index is used to judge the distribution of assessment items. This index only considers the expectations for a strand that have at least one hit—i.e., one related assessment item per expectation. The index is computed by considering the difference in the proportion of expectations and the proportion of hits assigned to the expectation. An index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a strand are equally distributed among the expectations for the given strand. Index values that approach 0 signify that a large proportion of the hits are on only one or two of all of the expectations hit. Depending on the number of expectations and the number of hits, a unimodal distribution (most items related to one expectation and only one item related to each of the remaining expectations) has an index value of less than .5. A bimodal distribution has an index value of around .55 or .6. Index values of .7 or higher indicate that items/activities are distributed among all of the expectations at least to some degree (e.g., every expectation has at least two items) and is used as the acceptable level on this criterion. Index values between .6 and .7 indicate the balance-of-representation criterion has only been “weakly” met.

## **Source-of-Challenge Criterion**

The source-of-challenge criterion is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted reading standard or expectation. Cultural bias or specialized knowledge could be reasons for an item to have a source-of-challenge problem. Such item characteristics may result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed.

## **Findings**

### **Standards**

The consensus EDOK value for each reading extended indicator for each standard can be found in Appendix A. Table 1 shows the percentages of extended indicator at each EDOK stage by grade. For each of the seven grades the Nebraska standards had 11 or 12 extended indicators. Reviewers judged that nearly all of these indicators had an EDOK stage 3 (DOK 1) or stage 4 (DOK 2). Reviewers reached consensus that the indicators expected students to use verbatim information (EDOK 3) or to determine meaning considering context (EDOK 4). Only one indicator, LAE 4.1.6.a, was judged to be an EDOK stage 5 (drawing inferences). This indicator expected students to judge the author's purpose by considering feelings of the reader. Reviewers felt that this required students to make an inference that went beyond comprehending text and verbatim reading. For grades 3 and 4, about one-third of the extended indicators were judged to be EDOK stage 3 (verbatim reading) and two-thirds of the indicator were judged to be EDOK stage 4. As would be expected, the proportion of indicators with an EDOK 4 increased to over 80% in the higher grades.

If no particular extended indicator is targeted by a given assessment item, reviewers were instructed to code the item at the level of a standard. This coding to a generic extended indicator sometimes indicates that the item is inappropriate for a grade level. However, if the item is grade-appropriate, then this situation may instead indicate that there is a part of the content not expressly or precisely described in the extended indicator. These items may highlight areas in the indicators that should be changed, or made more precise. Over half of the reviewers assigned 11 items to generic extended indicators over all seven assessment forms analyzed (Table 2). Most of these items were vocabulary items. The reviewers' comments under Notes in Appendix C state the reasons why a reviewer coded an item to a generic extended indicator. Most of these comments indicate that the main content knowledge needed to answer the question was not found in one of the grade level extended indicators. Some of the topics missing in the extended indicators were phonemic awareness (grade 3), making inferences (grade 3), abbreviations (grade 6), parts of speech (grades 6, 7, and 8), compound words (grade 7), and possessiveness (grade 11). Reviewers found grade 4 Item 7 to be confusing because

it required students to gather information from a diagram. Items assigned to generic extended indicators should be reviewed. The items may be appropriate, but the grade level extended indicators may be missing topics within the expectations. Reviewers' debriefing comments also highlight some ambiguities in the standards. These comments can be found in Appendix D.

Table 1  
*Percent of Expectations by Depth-of-Knowledge (DOK) Levels for Nebraska Alignment Analysis for Reading Alternate Assessment*

Grade	Total Number of Expectations	EDOK stage	Number of Standards by Level	Percent within Strand by Level
3	12	3	4	33
		4	8	66
4	12	3	4	33
		4	7	58
		5	1	8
5	12	3	5	41
		4	7	58
6	12	3	3	25
		4	9	75
7	11	3	2	18
		4	9	81
8	11	3	2	18
		4	9	81
12	11	3	2	18
		4	9	81

Table 2  
*Items assigned to generic extended indicators by number of reviewers and grade for Nebraska reading extended assessment alignment analyses*

Grade	Generic Content Expectation	Item Sequence Number (Number of Reviewers)
3	LA 3.1.5	9(5)
	LA 3.1.6	22(7)
4	LA 4.1.6	7(4) 23(4)
6	LA 6.1.5	13(5) 16(6) 22(6)
7	LA 7.1.5	4(7) 13(7)
8	LA 8.1.5	6(5)
11	LA 12.1.5	25(7)

## **Alignment of Curriculum Standards and Assessments**

All of the seven assessment forms had 25 multiple-choice items each. Each item had either three or four choices.

The results of the analysis for each of the four alignment criteria are summarized in Tables 3.1-3.7. More detailed data on each of the criteria are given in Appendix B, in the first three tables. With each table and for each grade, a description of the satisfaction of the alignment criteria for the given grade is provided. The reviewers' debriefing comments provide further detail about the individual reviewers' impressions of the alignment.

In Tables 3.1-3.7, "YES" indicates that an acceptable level was attained between the assessment and the learning goal on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error in the system. "NO" indicates that the criterion was not met by a noticeable margin—10% over an acceptable level for Depth-of-Knowledge Consistency, 10% over an acceptable level for Range-of-Knowledge Correspondence, and .1 under an index value of .7 for Balance of Representation.

### *Grade 3*

The NeSA reading alternate assessment for grade 3 and the grade 3 reading standards and extended indicators were acceptably aligned. Of the 25 items on the assessment, the majority of reviewers found that nine items (37%) mapped to the vocabulary standard and 15 items (63%) mapped to the comprehension standards. Five of the eight reviewers agreed that Item 1 did not match to any of the indicators under either standard and was uncodeable. The number of items that mapped to each of the two standards was sufficient to make a reliable decision about a student's proficiency of content related to each of these standards.

The Depth-of-Knowledge Consistency criterion was acceptable for the comprehension standard, but was not for the vocabulary standard. Reviewers indicated, on the average, that about two-thirds of the 15 items that mapped to the comprehension extended indicators had EDOK stages that were the same or higher than the EDOK stages assigned to the matching indicators. However, for the vocabulary standard only one-third of the items that corresponded to extended indicators had an EDOK stage that was at least as high as the EDOK stage of the assigned indicator. The low percentage of the vocabulary items with acceptable EDOK stages could result in a student being judged proficient for vocabulary without having to answer correctly any about vocabulary at a level of complexity as expected by the extended indicators.

The Range-of-Knowledge Correspondence and Balance of Representation criteria were both acceptable for each of the two reading standards. The majority of reviewers

found items that mapped to each of the three vocabulary extended indicators and to six of the nine comprehension extended indicators. This was sufficient coverage so that a student who would be judged proficient for each of the standards would likely have to answer correctly items measuring a variety of content knowledge. The reasonably high balance index values indicate that the items were fairly evenly distributed among the extended indicators under each of the standards. The number of items that mapped to the vocabulary extended indicators ranged from one to four and the number of items that mapped to the comprehension extended indicators ranged from zero to five.

Two items would need to be replaced or added to the assessment for the test to be considered fully aligned. These two items would need to map to extended indicators under the vocabulary standard and have an EDOK stage that would be at least as high as the EDOK stage of the corresponding extended indicator (stages 3 or 4). Reviewers wrote a number of notes on the items and extended indicators. These can be found in Appendices C and D. In general, reviewers felt that the alignment for grade 3 was good. There were three items that they could not fit to any extended indicators and the EDOK stage of the items that mapped to the vocabulary was not as high as expected by the extended indicators. A couple of reviewers noted that the assessment targeted the lower end of the EDOK spectrum without finding any items with an EDOK stage of 5 or 6.

Table 3.1  
*Summary of Acceptable Levels on Alignment Criteria for Nebraska Reading Extended Indicators and NeSA Reading Alternate Assessment Grade 3*

<i>NeSA Grade 3 Alternate Reading</i>	<i>Alignment Criteria</i>			
	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
<i>Content Expectations</i>				
LA 3.1.5 - Vocabulary	YES	NO	YES	YES
LA 3.1.6 – Comprehension	YES	YES	YES	YES

#### *Grade 4*

At grade 4, the NeSA reading alternate assessment and the reading extended indicators were found to be fully aligned. All four alignment criteria were judged to have an acceptable level. About one-fourth (6-7 items) of the items mapped to extended indicators under vocabulary and three-fourth (17-18 items) of the items mapped to extended indicators under comprehension. Seven of the eight reviewers thought that Item 13 was uncodeable because the vocabulary standard did not have an extended indicator addressing phonemic awareness. But even with one uncodeable item, the assessment had a sufficient number of items for each standard to make a reliable judgment about students' proficiency of each. The items also had an acceptable DOK consistency with 75% (vocabulary) and 65% (comprehension) of the items with an EDOK stage that was the same or higher than the EDOK stage of the assigned extended indicator. Range was good. At least one item mapped to each of the three extended indicators under vocabulary and to 69% of the nine extended indicators under comprehension. These items were sufficiently distributed among the extended indicators to have an acceptable balance.

Even though the results of the analysis indicated full alignment at grade 4, the reviewers’ notes bring attention to some of the items that were considered to fit only marginally an extended indicator. Reviewers also noted that additional extended indicators could be added to more appropriately cover the range of knowledge students should be expected to have. The reviewers’ notes in Appendices C and D express their concerns.

Table 3.2  
*Summary of Acceptable Levels on Alignment Criteria for Nebraska Reading Extended Indicators and NeSA Reading Alternate Assessment Grade 4*

<i>NeSA Grade 4 Alternate Reading</i>	<i>Alignment Criteria</i>			
<i>Content Expectations</i>	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
LA 4.1.5 – Vocabulary	YES	YES	YES	YES
LA 4.1.6 – Comprehension	YES	YES	YES	YES

*Grade 5*

As for grade 4, the NeSA reading alternate assessment for grade 5 was fully aligned with the grade 5 reading extended standards. Of the 25 items, the majority of reviewers found eight items (35%) that corresponded to extended indicators under the vocabulary standard and 15 items (65%) that corresponded to extended indicators under the comprehension standard. Two items, 14 and 21, were judged to be uncodeable. Item 14 required students to use dictionary skills or alphabetizing and Item 21 was judged by reviewers to target life skills rather than any of the two reading standards. Even with these two uncodeable items, the assessment had a sufficient number of items for each standard to have an acceptable level for the Categorical Concurrence criterion.

Table 3.3  
*Summary of Acceptable Levels on Alignment Criteria for Nebraska Reading Extended Indicators and NeSA Reading Alternate Assessment Grade 5*

<i>NeSA Grade 5 Alternate Reading</i>	<i>Alignment Criteria</i>			
<i>Content Expectations</i>	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
LA 5.1.5 – Vocabulary	YES	YES	YES	YES
LA 5.1.6 – Comprehension	YES	YES	YES	YES

Reviewers judged that nearly all of the grade 5 items, over 80%, had an EDOK stage that was the same or higher than the EDOK stage of the assigned extended indicator. Thus, the DOK consistency was high for grade 5. The majority of reviewers judged that at least two items mapped to each of the three vocabulary extended indicators and from one to three items mapped to six of the nine comprehension extended

indicators. This was sufficient to have an acceptable range. The balance was highly acceptable for grade 4 with most of the extended indicators assessed having two or three corresponding items. Other than the two items that were judged to be uncodeable, reviewers commented that the grade 5 assessment and standards were better aligned than grades 3 and 4. In their debriefing notes (Appendix D), reviewers did suggest changes in how the extended indicators could be improved.

*Grade 6*

The alignment between the extended indicators and the reading alternate assessment for grade 6 was acceptable, but less than for the prior grades. Only six of the reviewers agreed that four items on the assessment targeted vocabulary extended indicators. Five reviewers agreed on two other assessment items as measuring vocabulary knowledge. However, two reviewers found one of these items (Item 13) to be uncodeable. The average number of items found by the eight reviewers fell short of the six items needed to have an acceptable level for Categorical Concurrence. The reviewers could not agree that six items on the grade 6 assessment solidly measured expectations under the vocabulary standards. In fact, three of the five corresponding items were coded to the standard level rather than to one of the extended indicators. Reviewers did agree that 18 items targeted extended indicators under the comprehension standard, but they did not fully agree on the precise indicator the item measured. All reviewers judged that Item 1, requiring letter recognition, was uncodeable. Not only did the assessment have fewer vocabulary items than needed, only one third of the five items had an EDOK stage that was at least as high as the EDOK stage of the corresponding extended indicator. Range was good for both standards with nearly 70% of the underlying extended indicators with at least one corresponding item. Balance of Representation was acceptable for both standards.

Table 3.4  
*Summary of Acceptable Levels on Alignment Criteria for Nebraska Reading Extended Indicators and NeSA Reading Alternate Assessment Grade 6*

<i>NeSA Grade 6 Alternate Reading</i>	<i>Alignment Criteria</i>			
<i>Content Expectations</i>	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
LA 6.1.5 - Vocabulary	NO (5.62)	NO	YES	YES
LA 6.1.6 – Comprehension	YES	YES	YES	YES

For grade 6 the alignment was acceptable with only two items needed to be replaced or added to attain full alignment. Both of these items should clearly target an extended indicator under the vocabulary standard and should have an appropriate EDOK stage. Even though the results of the analysis produced acceptable results, the reviewers’ comments (see Appendix D) indicate that there were more grade 6 items that did not fully address the content as expressed in the extended indicators than on the assessments for the earlier grades. One reviewer wrote, “There were many items that were attempting to get at the standard but fell short, and as a result, had to be coded to a different standard.”

The grade 6 assessment could be improved by replacing about five or six items.

*Grade 7*

Alignment between the grade 7 NeSA reading alternative assessment and the grade 7 reading standards and extended indicators was acceptable. Reviewers, on the average, found over six items that targeted extended indicators under vocabulary and 18 items that targeted extended indicators under comprehension. Two of the vocabulary items did not map precisely to any of the extended indicators and were coded to the standard. The average EDOK stage of the vocabulary items was below the EDOK stage assigned to each of the vocabulary extended indicators. Overall, only two of the vocabulary items had an EDOK stage that was the same or higher than the EDOK stage of the assigned extended indicator. The EDOK stages of the 18 comprehension items were acceptable based on the decision rules used for this study, but only barely (51%). Over 60% of the extended indicators for each standard had at least one corresponding item. This was enough to have an acceptable level for range. The balance was acceptable, but a high proportion of the comprehension items (50%) targeted one extended indicator (LAE 7.1.6. j). One reviewer explained, “There were a number of items written for standard 7.1.6.j (literal/inferential questions) which seems to be a "catch all" when no match is clearly found for another standard.”

Overall, the alignment for grade 6 was acceptable. Two items would need to be added or replaced to attain full items. Both of these items would need to target vocabulary extended indicators and have an acceptable EDOK stage. Reviewers indicated that the grade 7 items were a better fit with the grade level extended indicators than for the prior grades. However, the reviewers made a number of comments about missing expectations and some content that should have been assessed that was not (e.g. antonyms, cause/effect, or compare/contrast).

Table 3.5  
*Summary of Acceptable Levels on Alignment Criteria for Nebraska Reading Extended Indicators and NeSA Reading Alternate Assessment Grade 7*

<i>NeSA Grade 7 Alternate Reading</i>	<i>Alignment Criteria</i>			
	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
<i>Content Expectations</i>				
LA 7.1.5 - Vocabulary	YES	NO	YES	YES
LA 7.1.6 – Comprehension	YES	YES	YES	YES

*Grade 8*

The alignment between the grade 8 NeSA reading alternate assessment and the reading extended indicators was found to be barely acceptable. The majority of reviewers only found four items that mapped to extended indicators under the vocabulary standards. This is below the six items needed to have an acceptable level for the Categorical Concurrence criterion. Reviewers mapped 18 items to extended indicators under the

comprehension standard. All of the reviewers judged that Items 2 and 5 were uncodeable. Item 2 was considered too far below grade level to match any grade 8 extended indicator. Item 5 required students to use letter recognition that was also below grade level.

The EDOK stages of the four items that mapped to the vocabulary standard were reasonably comparable to the EDOK stage of the corresponding extended indicators. However, only seven of the 18 items (40%) that mapped to comprehension extended indicators had an EDOK stage that was at least as high as the EDOK stage of the corresponding extended indicators. Thus, the assessment and the standards only weakly met the DOK consistency criterion for the grade 8 comprehension standard. Both range and balance were acceptable. Items on the assessment targeted over 65% of the underlying extended indicators for both standards and were nearly evenly distributed among the extended indicators.

Overall, the alignment for grade 8 was acceptable. Four or five items would need to be added or replaced to have full alignment between the assessment and the standards. Two additional vocabulary items would be needed along with two comprehension items. All of these items need to have an EDOK stage that is at least the same as the EDOK stage of the corresponding extended indicator. Reviewers’ debriefing comments note that the reviewers found the extended indicators to be reasonable. The assessment had high percentage of items at an EDOK stage 3. Reviewers thought that an assessment for grade 8 students would have more items at a higher level of complexity. One reviewer wrote, “Assessment at this grade level seems too easy with so many Stage 3 literal items. Adding the student written poem is good. Like that the stories and paragraphs are longer.” Another reviewer supported this view, “There are so many pictures that cue the right answer. Many of the items become matching items rather than comprehension items.”

Table 3.6  
*Summary of Acceptable Levels on Alignment Criteria for Nebraska Reading Extended Indicators and NeSA Reading Alternate Assessment Grade 8*

<i>NeSA Grade 8 Alternate Reading</i>	<i>Alignment Criteria</i>			
	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
<i>Content Expectations</i>				
LA 8.1.5 - Vocabulary	NO (4.12)	YES	YES	YES
LA 8.1.6 – Comprehension	YES	WEAK	YES	YES

*Grade 11*

For grade 11 the alignment between the NeSA reading alternate assessment and the grade 12 reading extended indicators was acceptable. Reviewers found six items that mapped to vocabulary extended indicators and 16 items that mapped to comprehension extended indicators. These were a sufficient number to have an acceptable level for the Categorical Concurrence criterion. The EDOK stages of the items, however, were not as high as they should be to have a fully aligned assessment with the standards. Only 51%,

on the average, of the vocabulary items and 44% of the comprehension items had an EDOK stage that was the same or higher than the EDOK stage of the assigned extended indicator. This proportion just met an acceptable level for DOK consistency for the vocabulary standard, but only weakly met the criterion for the comprehension standard. The range and balance were acceptable for both standards. Although balance was acceptable for the comprehension standard, the majority of reviewers coded eight items to one extended indicator—LAE 12.1.6 f. Although balance was within the acceptable limits, having nearly half of the items mapped to one extended indicator seems to be excessive.

Overall, the alignment for grade 11 was acceptable. Only two items would need to be added or replaced. These items should target comprehension extended indicators and should have appropriate EDOK stages. More than one reviewer indicated that the items on the grade 11 assessments were less complex than expected. These reviewers did not see a progression across the grades from less complex to higher complexity. One reviewer commented, “I have noted that some of the test items seem too easy for grade level expectations. Also, a couple of them are addressing indicators which are not directly stated in the extended indicators but have been in previous grade levels (singular/plural).” Although the results for the grade 11 indicated that the alignment was acceptable, some consideration should be given to the vertical alignment across the grades.

Table 3.7

*Summary of Acceptable Levels on Alignment Criteria for Nebraska Reading Extended Indicators and NeSA Reading Alternate Assessment Grade 11*

<i>NeSA Grade 11 Alternate Reading</i>	<i>Alignment Criteria</i>			
	<i>Categorical Concurrence</i>	<i>Depth-of-Knowledge Consistency</i>	<i>Range of Knowledge</i>	<i>Balance of Representation</i>
<i>Content Expectations</i>				
LA 12.1.5 - Vocabulary	YES	YES	YES	YES
LA 12.1.6 – Comprehension	YES	WEAK	YES	YES

### **Source of Challenge Issue and Reviewers’ Comments**

Reviewers were instructed to document any source-of-challenge issue and to provide any other comments they may have. These comments can be found in Tables (grade).5 and (grade).7 in Appendix C. Two or three reviewers noted source of challenge issues for two to seven items for each grade. Reviewers found the most source issues for the grades 7 and 8 assessments. Generally reviewers noted issues with graphics or the item format that may cause students to be misled or not use the knowledge the item was designed to measure. Single reviewers identified a number of other items as having a source of challenge issue. All of the items listed in on Table 5 for each grade in Appendix C should be reviewed in case one reviewer noted something that the others did not. After coding each assessment form, reviewers were asked to respond to five debriefing questions. All of the comments made by the reviewers are given in Appendix D. The

notes in general offer an opinion on the item or give an explanation of the reviewer’s coding.

### **Reliability Among Reviewers**

The overall intraclass correlation among the reading reviewers’ assignment of EDOK stages to items was high for eight reviewers for all seven analyses (Table 4). An intraclass correlation value greater than 0.8 generally indicates a high level of agreement among the reviewers. The intraclass correlation for assigning EDOK stages to items for all seven analyses were higher than 0.80. A pairwise comparison was used to determine the degree of reliability of reviewer coding at the extended indicator level and at the standard level. The pairwise comparison was computed by considering for each item the coding assigned by each reviewer compared to the coding by each of the other seven reviewers. With six reviewers, for example, a total of 15 comparisons would be computed for each item. The pairwise extended indicator agreements were all above 0.60 and five were about 0.70. These values are very reasonable and comparable for most alignment studies. The pairwise standard agreement values were all about 0.90 which is normal for most alignment study. Reviewers had high agreement in assigning items to both extended indicators and standards. These reliability measures were computed for the final results after the eight reviewers adjudicated their mappings. In general, mappings without majority agreement were discussed. Reviewers then decided if they wanted to change the value. Each reviewer was left to make their own decision about whether or not the value they coded was appropriate or should be changed.

Table 4  
*Intraclass and Pairwise Comparisons, Nebraska Reading Alternate Alignment Analysis for Grades 3-8 and 11 Extended Reading Indicators and NeSA Reading Alternate Assessments 2010*

Grade	Number of Reviewers	Intraclass Correlation	Pairwise Comparison:	Pairwise: Content Expectation	Pairwise: Strand
3	8	0.95	0.59	0.70	0.94
4	8	0.92	0.66	0.71	0.97
5	8	0.93	0.71	0.77	0.97
6	8	0.92	0.57	0.69	0.95
7	8	0.93	0.61	0.76	1.00
8	8	0.95	0.65	0.62	0.93
11	8	0.92	0.65	0.73	0.95

## Summary

A three-day alignment institute was held in Lincoln, Nebraska on September 20-22, 2020. Eight reviewers, four from Nebraska and four from other states, analyzed the Nebraska Student Accountability (NeSA) reading alternate assessment and the two reading standards (vocabulary and comprehension) and extended indicators for grades 3-8 and 10 applicable to students with significant disabilities. The assessments had been given in the spring of the 2009-2010 school year. Reviewers were reading specialist, special education teachers, and assessment specialists. One assessment of 25 items was analyzed for each grade. Six Extended Depth of Knowledge (EDOK) stages were used to identify the different levels of content complexity.

The alignment between the NeSA reading alternate assessment and standards for each grade was found to be acceptable. Fewer than five items would need to be added or replaced on each assessment to attain full alignment (see the summary table below). In the summary table, 100% indicates that the alignment criterion was acceptable for both standards. For grades 4 and 5, the standards and the assessments were found to be fully aligned. The main alignment issue was with the Depth-of- Knowledge consistency for grades 3, 6, 7, 8, and 11 and the Categorical Concurrence for grades 6 and 8. For grades 3, 6, and 7, the vocabulary items on the assessment had less than 50% of the items with an EDOK stage that was the same or higher than the EDOK stage of the assigned extended indicator. For the higher grades, 8 and 11, the comprehension items did not satisfy this criterion. For grades 6 and 8, reviewers also did not find six or more items that they judged clearly assessed the extended indicators under the vocabulary standard.

Even though the alignment for each grade was found to be acceptable when the minimum acceptable levels were considered, reviewers noted a number of ways that the assessments or the extended indicators could be improved. At least one or two items on the assessment for each grade were found uncodeable. Reviewers indicated that these items did not match any of the standards or extended indicators. For example, reviewers indicated that one grade 5 item measured students' knowledge of life skills rather than any reading skill. For all but grade 5, reviewers found at least one item that did not precisely target an extended indicator, but did in general measure knowledge under one of the standards. These items that were coded as targeting a generic extended indicator suggest possible content that was not included in the extended indicators and indicate the need to reconsider the extended indicators. Reviewers wrote notes about items that generally measured content related to an extended indicator, but did not target the content that was the main intent of the extended indicator. In addition for each grade, two or three reviewers identified a source of challenge for two to seven items. Thus, the alternate assessments and the extended indicators met the basic requirements to be considered acceptably aligned. However, there are ways that the assessments and standards could be improved to strengthen the relationship between both along with increasing the quality of the assessments in general.

Summary Table

*Percent of Nebraska Reading Standards with Acceptable Level on Each Alignment Criteria for the 2010 NeSA Reading Alternate Assessments*

Grade (N=25 items on each assessment)	<i>Categorical Concurrence</i> (Total number of items for the one stand)	<i>Depth-of-Knowledge Consistency</i> (50% at/above)	<i>Range of Knowledge</i> (50% of objectives)	<i>Balance of Representation</i> (without possible weakness)	<i>Estimated Range of Items per assessment to be Added or Replaced for Full Alignment</i>
3	100%	50%	100%	100%	2
4	100%	100%	100%	100%	0
5	100%	100%	100%	100%	0
6	50%	50%	100%	100%	2
7	100%	50%	100%	100%	2
8	50%	50%	100%	100%	4-5
11	100%	50%	100%	100%	2

Categorical Concurrence >6 items  
 Depth-of-Knowledge >50% with EDOK stage the same or higher than level of corresponding content expectation  
 Range-of-Knowledge >50% of content expectations under a strand  
 Balance of Representation A possible weakness if one or more content expectation with a relative large number of items (e.g. five or more than the content expectation with the next highest number of items)

**References**

Subkoviak, M. J. (1988). A practitioner’s guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25(1), 47-55.

Valencia, S. W., & Wixson, K. K. (2000). Policy-oriented research on literary standards and assessment. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Vol. III*. Mahwah, NJ: Lawrence Erlbaum.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and Mathematics education*. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.