WISCONSIN KNOWLEDGE AND CONCEPTS EXAMINATION

CONTINUOUS IMPROVEMENT PLAN

July 2009

Table of Contents

Part 1: Overview

The Wisconsin Knowledge and Concepts Examination (WKCE) has been developed by the Wisconsin Department of Public Instruction (DPI) and CTB/McGraw-Hill to assess student performance and is used to determine the adequate yearly progress of students at the school, district and state levels. In July of 2006, Dr. Norm Webb was contracted with to provide an alignment analysis of the WKCE and the assessment framework.

The goals of the alignment analysis were to identify the extent to which the WKCE and the standards were aligned in relation to categorical concurrence, depth of knowledge consistency, range-of-knowledge correspondence, and balance of representation for the reading, mathematics, and science components of the WKCE. The reports *Alignment Analysis of Reading Standards and Assessments*, *Alignment Analysis of Mathematics Standards and Assessments*, and *Alignment Analysis of Science Standards and Assessments* provide detailed summaries of the degree of alignment that was observed on the WKCE.

This document outlines the steps DPI has taken and continues to take to improve the WKCE based on recommendations from the alignment analyses. These steps have included annual item development activities and refining the test blueprint to address areas where the expectations were not clear.

Part 2: Alignment Analyses

The four alignment analyses conducted included (1) categorical concurrence, (2) depth-of-knowledge consistency, (3) range-of-knowledge correspondence, and (4) balance of representation. In addition to the four alignment analyses, a source of challenge analysis was also conducted on the science component of the WKCE. A detailed description of each of these criteria is presented below.

Categorical Concurrence
An important aspect of alignment between standards and assessments is the extent to which both address the same content categories. Categorical concurrence provides a general indication of alignment if both documents incorporate the same content. *Categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents.* In this analysis, categorical concurrence was met if the assessment had at least six items to measure content for a standard.

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 is met if the cognitive demand required on the assessment is as cognitively challenging as what students are expected to know and do as stated in the standards.* In this analysis, depth-of-knowledge consistency was met when at least 50% of the items corresponding to a standard were at or above the level of knowledge of the standard.

Interpreting and assigning depth-of-knowledge levels to standards and assessment items is an essential requirement of alignment analysis. Depth-of-knowledge levels for the reading analysis were based on Valencia and Wixson's definitions (2000, pp. 909-935):

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, slight paraphrasing of specific details from the 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 verbatim or only slightly paraphrased details from the text.
- Use a dictionary to find the meanings of words.
- Recognize 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 of 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. However, items require closer understanding of text, possibly through the item's paraphrasing of both the question and the answer. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Use context cues to identify the meaning of unfamiliar words, phrases, and expressions that could otherwise have multiple meanings.
- 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:

- Explain or recognize how author's purpose 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.

In mathematics, the depth-of-knowledge levels were defined as follows:

Mathematics Level 1 (Recall). Level 1 includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well defined, and straight algorithmic procedure should be included at this level. Other key words that signify a Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels, depending on what is to be described and explained.

Mathematics Level 2 (Skill/Concept). Level 2 includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions in how to approach the problem or activity. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret," could be classified at different levels depending on the object of the action. For example, interpreting information from a simple graph or requiring mathematics information from the graph are Level 2. Level 2 activities are not limited solely to number skills, but can involve visualization skills and probability skills. Other Level 2 activities include noticing and describing non-trivial patterns; explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Mathematics Level 3 (Strategic Thinking). Level 3 requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

Mathematics Level 4 (Extended Thinking). Level 4 requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying

significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be at Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include developing and proving conjectures; designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

In science, the depth-of-knowledge levels were defined as follows:

Science Level 1 (Recall and Reproduction) is the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple science process or procedure. Level 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (e.g. a recipe), or perform a clearly defined series of steps. A "simple" procedure is well defined and typically involves only one step. Verbs such as "identify," "recall," "recognize," "use," "calculate," and "measure" generally represent cognitive work at the recall and reproduction level. Simple word problems that can be directly translated into and solved by a formula are considered Level 1. Verbs such as "describe" and "explain" could be classified at different DOK levels, depending on the complexity of what is to be described and explained.

A student answering a Level 1 item either knows the answer or does not: that is, the answer does *not* need to be "figured out," or "solved." In other words, if the knowledge necessary to answer an item automatically provides the answer to the item, then the item is at Level 1. If the knowledge necessary to answer the item does *not* automatically provide the answer, the item is at least at Level 2.

Science Level 2 (Skills and Concepts) includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is more complex than in Level 1. Items require students to make some decisions as to how to approach the question or problem. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Level 2 activities include making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Some action verbs, such as "explain," "describe," or "interpret," could be classified at different DOK levels, depending on the complexity of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, is at Level 2. An item that requires interpretation from a complex graph, such as making decisions regarding features of the graph that need to be considered and how information from the graph can be aggregated, is at Level 3.

Science Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are complex and abstract. The complexity does *not* result only from the fact that there could be multiple answers, a possibility for both Levels 1 and 2, but because the multi-step task requires more demanding reasoning. In most instances, requiring students to explain their thinking is at Level 3; requiring a very simple explanation, or a word or two, should be at Level 2. An activity that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Experimental designs in Level 3 typically involve more than one dependent variable. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve non-routine problems.

Science Level 4 (Extended Thinking). Tasks at Level 4 have high cognitive demands and are very complex. Students are required to make several connections—relate ideas within the content area or among content areas—and have to select or devise one approach among many alternatives on how the situation can be solved. Many on-demand assessment instruments will *not* include any assessment activities that could be classified as Level 4. However, standards, goals, and objectives can be stated in such a way as to expect students to perform extended thinking. "Develop generalizations of the results obtained and the strategies used and apply them to new problem situations," is an example of a grade 8 objective that is at Level 4. Many, but *not* all, performance assessments and open-ended assessment activities requiring significant thought will be Level 4.

Level 4 requires complex reasoning, experimental design and planning, and probably will require an extended period of time either for the science investigation required by an objective, or for carrying out the multiple steps of an assessment item. However, the extended time period is *not* a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student conducts a river study that requires taking into consideration a number of variables, this would be at Level 4.

Range-of-Knowledge Correspondence

For standards and assessments to be aligned, the breadth or range of knowledge required on both should be comparable. *Range-of-knowledge correspondence between standards and assessments is met if the span of knowledge expected of students by a standard is the same as, or comparable to, the span of knowledge students need in order to correctly answer the assessment items/activities.* In this analysis, range-of-knowledge correspondence was met when fifty percent of the benchmarks for a standard had at least one related assessment item. Setting the criterion at this level assumes that each benchmark for a standard should be given equal weight. It also increases the likelihood that students will have to demonstrate knowledge on more than one benchmark per standard to achieve a minimal passing score.

Balance of Representation

In addition to comparable depth- and range-of-knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range-of-knowledge criterion considers the number of benchmarks within a standard that are covered by one or more

assessment items; it does not take into consideration the distribution of assessment items/activities among the benchmarks. *Balance of representation between standards and assessments indicates the degree to which one benchmark is given more emphasis on the assessment than another.* In this analysis, an index was used to judge the distribution of assessment items. Only benchmarks with at least one corresponding assessment item were included in this index. The index was computed by considering the difference in the proportion of benchmarks and the proportion of hits assigned to the benchmark. An index value of one signifies perfect balance and is obtained if the hits (corresponding assessment items) related to a standard are equally distributed among the benchmarks for the given standard. Index values approaching zero signify that a large proportion of the hits are on only one or two of all of the benchmarks hit. Depending on the number of benchmarks and the number of hits, a unimodal distribution (most items related to one benchmark and only one item related to each of the remaining benchmarks) 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 benchmarks at least to some degree (e.g., every benchmark 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 been "weakly" met.

Source-of-Challenge – Science only
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 science skill, concept, or application. 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.

Part 3: Reading

One of the overall objectives of the WKCE program is to have enough items to be able to develop test forms that are aligned with the curriculum that students are expected to learn. The development process for all reading assessment items includes an alignment to a standard and sub-skill, linking it with a reading passage, assigning it a depth-of-knowledge level, assuring it conforms to a set of industry developed guidelines for clarity and wording, a thorough review by a committee of subject matter experts, and a successful field test prior to inclusion on an operational test form. Due to the extensive nature of these necessary steps, item development efforts have been undertaken every year since the inception of the WKCE.

Alignment Analysis Results from 2006
Several key considerations are taken into account when planning for item development each year. These include the number of existing items in the bank at each standard and sub-skill, the number of items associated with each reading passage, the statistical properties of the existing items, and the depth-of-knowledge classification for each item. In July of 2006, Dr. Norm Webb from the University of Wisconsin was contracted to review the Fall 2005 test forms that were administered and make recommendations for improving the examination. Appendix A contains a summary of the primary conclusions from this review. One recurring theme over all the grades was that a higher depth-of-knowledge (DOK) was needed for Standard 3 – Analyze Text. It was

recommended that the Analyze Text standard be assigned a DOK of 3, even though several of the corresponding assessment items on the WKCE had been identified as having a DOK of 2.

Technical Advisory Committee Recommendations
During the December 2006 meeting of DPI's Technical Advisory Committee (TAC), the alignment study results for reading were reviewed and a recommendation was made that each test form should have 65% or more items at or above the DOK level of the objective to have a fully aligned test. This TAC recommendation was higher than the 50% standard used during Dr. Webb's *Alignment Analysis* and was accepted as a guide for item development in subsequent years.

Item Development
In the fall 2007 through spring 2008 item development period, between 47 and 58 new items were written at each grade level. Table 1 provides a breakdown of the number of items written at each DOK level. Of this number, between 20 and 27 at each grade level were written for Standard 3 – Analyze Text. An emphasis was placed on writing these items to the DOK 3 level to ensure that the deficiency identified in the *Alignment Analysis* report would not continue.

In the fall 2008 through spring 2009 item development period, new reading passages were commissioned or obtained for each grade. Between 52 and 59 new items were written at each grade level. These items were written to align with the new reading passages. The focus remained on ensuring that the new items met the alignment criteria discussed earlier, especially maintaining that items be written to the appropriate DOK level. Item development also addressed adding items for other selected standards in order to expand the item pool, improve overall content coverage in the item pool, and to increase flexibility when selecting operational forms.

Test Form Assembly Checklist
The *Alignment Analysis* has led to many changes in the way the Reading portion of the WCKE is developed. In addition to the recommendation that was received from the TAC regarding DOK, it has been realized that a more detailed review of items to include on test forms is needed. Prior to the alignment analysis, the primary forms of review included verification of the statistical characteristics of items and forms (item p-values, point biserial coefficients, test characteristic curves) and verification that the test forms met the categorical concurrence standard. Since the analysis, more detailed review procedures have been established. The Test Form Assembly Checklist (see Appendix D) was developed to assist in the review process. This Checklist has allowed for a more detailed picture of each test form to be obtained and to ensure alignment criteria as well as the other traditional psychometric criteria of a test form are met.

Part 4: Mathematics

One of the overall objectives of the WKCE program is to have enough items to be able to develop test forms that are aligned with the curriculum that students are expected to learn. The development process for all mathematics assessment items includes an alignment to a standard and sub-skill, assigning it a depth-of-knowledge level, assuring it conforms to a set of industry developed guidelines for clarity and wording, a thorough review by a committee of subject matter experts, and a successful field test prior to inclusion on an operational test form. Due to

the extensive nature of these necessary steps, item development efforts have been undertaken every year since the inception of the WKCE.

Alignment Analysis Results from 2006

Several key considerations are taken into account when planning for item development each year. These include the number of existing items in the bank at each standard and sub-skill, the statistical properties of the existing items, and the depth-of-knowledge classification for each item. In July of 2006, Dr. Norm Webb from the University of Wisconsin was contracted to review the Fall 2005 test forms that were administered and make recommendations for improving the examination. Appendix B contains a summary of the primary conclusions from this review. One recurring theme over all the grades was that items needed to be developed at a higher depth-of-knowledge. This was most prevalent with standard E – Statistics and Probability where it was an issue for six of the seven tested grades, although the concern also existed for a number of other standards at various grades.

Technical Advisory Committee Recommendations

During the December 2006 meeting of DPI's Technical Advisory Committee (TAC), the alignment study results for mathematics were reviewed and a recommendation was made that each test form should have 65% or more items at or above the DOK level of the objective to have a fully aligned test. This TAC recommendation was higher than the 50% standard used during Dr. Webb's *Alignment Analysis* and was accepted as a guide for item development in subsequent years. Additionally, the mathematics blueprints were modified to reflect the inclusion of a 2-point constructed-response item and the subsequent reduction of the 3-point constructed-response items from four to three. Also, the number of multiple-choice items for each reporting category was adjusted to reflect the use of multiple-choice items for reporting category A (Mathematical Processes).

Item Development

In the fall 2006 through spring 2007 item development period, 180 mathematics items were written and approved. These items were written to address content areas where not enough items existed, primarily due to the slight modification of the test blueprint. Twenty five of the new items were constructed-response items and the rest were selected-response.

In the fall 2007 through spring 2008 item development period, between 39 and 48 new items were written at each grade level. Table 2 provides a breakdown of the number of items written at each DOK level. Of this number, an average of twelve new items per grade level (70 total items) was written for standard E – Statistics and Probability. An emphasis was placed on writing these items at the DOK 3 level to ensure that the deficiency identified in the *Alignment Analysis* report would not continue.

In the fall 2008 through spring 2009 item development period, between 46 and 57 new items were written at each grade level. The focus remained on ensuring that the new items met the alignment criteria discussed earlier, especially maintaining that items be written to the appropriate DOK level. Item development also addressed adding items for other selected standards in order to expand the item pool, improve overall content coverage in the item pool, and to increase flexibility when selecting operational forms.

Test Form Assembly Checklist
The *Alignment Analysis* has led to many changes in the way the Math portion of the WCKE is developed. In addition to the recommendations that were received from the TAC regarding blueprints and DOK, it has been realized that a more detailed review of items to include on test forms is needed. Prior to the alignment analysis, the primary forms of review included verification of the statistical characteristics of items and forms (item p-values, point biserial coefficients, test characteristic curves) and verification that the test forms met the categorical concurrence standard. Since the analysis, more detailed review procedures have been established. The Test Form Assembly Checklist (see Appendix D) was developed to assist in the review process. This Checklist has allowed for a more detailed picture of each test form to be obtained and to ensure alignment criteria as well as the other traditional psychometric criteria of a test form are met.

Part 5: Science

One of the overall objectives of the WKCE program is to have enough items to be able to develop test forms that are aligned with the curriculum that students are expected to learn. The development process for all WKCE science assessment items includes an alignment to a standard and sub-skill, assigning it a depth-of-knowledge level, assuring it conforms to a set of industry developed guidelines for clarity and wording, a thorough review by a committee of subject matter experts, and a successful field test prior to inclusion on an operational test form. By utilizing these steps in the development process, an assessment is more likely to meet the requirements of an independent analysis of its alignment.

Alignment Analysis Results from 2006
In July of 2006, Dr. Norm Webb from the University of Wisconsin was contracted to review the Fall 2005 test forms that were administered and make recommendations for improving the examination. Appendix C contains a summary of the primary conclusions from this review. One recurring theme over the three grades was that there were not enough items to adequately sample the eight objectives. This was most prevalent with Standard A – Science Connections, Standard B – Nature of Science, Standard G – Earth and Space Science, and Standard H – Science in Social and Personal Perspectives. In addition, the range of knowledge criterion was not met for most standards. This indicates that the breadth of content covered on the test is not consistent with the expectations of what students should be learning in the classroom.

Technical Advisory Committee Recommendations
These alignment analysis results were presented to DPI's Technical Advisory Committee (TAC) during their December 2006 and May 2007 meetings. Much discussion ensued about the results and the interpretations that were made. Recommendations were put forth regarding the structure of the standards as well as depth of knowledge requirements. Specifically, a proposal was put forth and adopted to modify the test blueprint by collapsing Standard A and Standard B into a joint standard and collapsing Standard G and Standard H into another joint standard. When the test blueprint was originally developed, content experts identified the number of items from each standard that should be included on the examination. Consensus opinion was that Standards A, B, G, and H needed the least number of items (approximately four items per standard). By collapsing these four standards into two standards, the recommendations of the content experts

are maintained and the categorical concurrence alignment criterion is met. As a result of this recommendation, beginning with the Fall 2008 administration WKCE science results were reported for six standards rather than the eight standards that comprised the first three administrations of the assessment.

A related subject of discussion during the TAC meetings was the recognition that during the alignment analysis, the reviewers compared items to the standards at a finer threshold than what was originally anticipated. Reviewers analyzed the standards at the sub-skill level rather than at the objective level. As a result, the Range of Knowledge Correspondence criterion was not met for most standards. Upon further review, it was realized that by using the objective level as the benchmark, this criterion was met in most situations. In the *Alignment Analysis*, Dr Webb had mentioned this as a possible solution for the range of knowledge correspondence. The statement below from the *Alignment Analysis* confirms this position.

> *The assessment could more easily be aligned to the standards if the structure of the standards was modified by consolidating the objectives and standards into a reduced number. The current level of specificity in the standards is not necessary for guiding an accountability system.*

In addition, Dr. Webb was consulted about the reasonableness of this idea. He confirmed that during the alignment study the sub-skill level of the standards was being used to guide the analysis and that it would be appropriate to use the objective level of the standards when conducting future alignment comparisons.

The second recommendation from the TAC related to the depth of knowledge standard that should be established for the test blueprint. The depth of knowledge recommendation was that each test form should have 65% or more items at or above the DOK level of the objective to have a fully aligned test. This TAC recommendation was higher than the 50% standard used during Dr. Webb's *Alignment Analysis* and was accepted as a guide for form development in subsequent years.

Test Form Assembly Checklist

The *Alignment Analysis* has led to many changes in the way the Science portion of the WCKE is developed. In addition to the recommendations that were received from the TAC regarding blueprints and DOK, it has been realized that a more detailed review of items to include on test forms is needed. Prior to the alignment analysis, the primary forms of review included verification of the statistical characteristics of items and forms (item p-values, point biserial coefficients, test characteristic curves) and verification that the test forms met the categorical concurrence standard. Since the analysis, more detailed review procedures have been established. The Test Form Assembly Checklist (see Appendix D) was developed to assist in the review process. This Checklist has allowed for a more detailed picture of each test form to be obtained and to ensure alignment criteria as well as the other traditional psychometric criteria of a test form are met.

Part 6: Summary

The WKCE in its current format has been in place since the fall of 2005. During this time period, many steps have been taken to ensure that the assessment is adequately aligned to the state curriculum standards. First and foremost, a review of the alignment of the assessment was conducted by Dr. Norm Webb in 2006. Results from this review have led to many substantive changes in the way that the assessment is developed. Targeted tem writing coupled with a review procedure that is designed to verify the appropriateness of the items is a key component of this development effort. In addition, test blueprints have been clarified to ensure that the assessment clearly reflects the statewide curriculum standards and a test form assembly checklist has been developed to ensure that the blueprint is met when forms are assembled. These guidelines and development strategies that have been adopted have aided in the attempts at continuing to improve the assessment. These efforts will continue in an effort to maintain the WKCE as a useful tool for identifying student, school and district performance in the State of Wisconsin.

Appendix A: Reading 2006 *Alignment Analysis* Summary

Grade 3

- Higher DOK needed for Standard 3 (analyze text)
- More items needed for Standard 4 (evaluate and extend text)
- Range-of-knowledge for Standard 4 (evaluate and extend text) is "weakly" met

Grade 4

- Balance of representation for Standard 1 (determine the meaning of words and phrases in context) is "weakly" met
- Higher DOK needed for Standard 3 (analyze text)
- More items needed for Standard 4 (evaluate and extend text)

Grade 5

- Balance of representation for Standard 1 (determine the meaning of words and phrases in context) is not met
- Higher DOK needed for Standard 3 (analyze text)

Grade 6

- Reasonable overall alignment
- Balance of representation for Standard 1 (determine the meaning of words and phrases in context) is "weakly" met
- DOK for Standard 3 (analyze text) is "weakly" met

Grade 7

- Reasonable overall alignment
- DOK level for Standard 3 (analyze text) is "weakly" met

Grade 8

- Fully aligned

Grade 10

- Fully aligned[1]
- Balance of representation for Standard 1 (determine the meaning of words and phrases in context) is "weakly" met

[1] Even though the assessment overemphasizes Objective 1.1 when compared to number of items that correspond to the other two objectives under Standard 1, this is not considered a major alignment issue. Therefore, the grade10 standards and assessment are considered to be fully aligned.

Appendix B: Mathematics 2006 *Alignment Analysis* Summary

Grade 3

- DOK level for Standard A (mathematical processes) is "weakly" met
- Higher DOK needed for Standard E (statistics and probability)

Grade 4

- Balance of representation for Standard D (measurement) is "weakly" met
- Higher DOK needed for Standard E (statistics and probability)
- Balance of representation for Standard F (algebraic relationships) is "weakly" met

Grade 5

- Higher DOK needed for Standard A (mathematical processes)
- Higher DOK needed for Standard E (statistics and probability)
- DOK level for Standard F (algebraic relationships) is "weakly" met

Grade 6

- Higher DOK needed for Standard A (mathematical processes)
- DOK level for Standard B (number operations and relationships) is "weakly" met
- Higher DOK needed for Standard E (statistics and probability)

Grade 7

- DOK level for Standard B (number operations and relationships) is "weakly" met
- DOK level for Standard C (geometry) is "weakly" met
- Higher DOK needed for Standard E (statistics and probability)
- DOK level for Standard F (algebraic relationships) is "weakly" met

Grade 8

- Balance of representation for Standard D (measurement) is "weakly" met
- Higher DOK needed for Standard E (statistics and probability)

Grade 10

- Higher DOK needed for Standard A (mathematical processes)
- Balance of representation for Standard D (measurement) is "weakly" met
- DOK level for Standard F (algebraic relationships) is "weakly" met

Appendix C: Science 2006 *Alignment Analysis* Summary

Grade 4

- Not enough items aligned to Standard A – Science Connections to conduct alignment analysis
- Not enough items aligned to Standard B – Nature of Science to conduct alignment analysis
- Standard C – Science Inquiry does not meet the alignment criteria for depth of knowledge consistency, range of knowledge, and balance of representation
- Range-of-knowledge for Standard D (Physical Science) is not met
- More items are needed for Standard E (Earth and Space Science) and range-of-knowledge for the standard is "weakly" met
- Range-of-knowledge for Standard F (Life and Environmental Science) is not met
- More items are needed for Standard G (Earth and Space Science) and range-of-knowledge for the standard is not met
- More items are needed for Standard H (Science in Social and Personal Perspectives)

Grade 8

- Not enough items aligned to Standard A – Science Connections to conduct alignment analysis
- More items are needed for Standard B (Nature of Science) and range-of-knowledge for the standard is not met
- Range-of-knowledge for Standard C (Science Inquiry) is not met
- More items are needed for Standard D (Physical Science) and range-of-knowledge for the standard is "weakly" met
- Range-of-knowledge for Standard E (Earth and Space Science) is not met
- Depth of knowledge consistency and range-of-knowledge for Standard F (Life and Environmental Science) are not met
- More items are needed for Standard G (Earth and Space Science) and range-of-knowledge for the standard is not met
- Not enough items aligned to Standard H – Science in Social and Personal Perspectives to conduct alignment analysis

Grade 10

- Not enough items aligned to Standard A – Science Connections to conduct alignment analysis
- More items are needed for Standard B (Nature of Science) and range-of-knowledge for the standard is not met
- Depth of knowledge consistency and range -of-knowledge for Standard C (Science Inquiry) are not met
- Range-of-knowledge for Standard D (Physical Science) is not met
- Depth of knowledge consistency for Standard E (Earth and Space Science) is "weakly" met
- Range-of-knowledge for Standard F (Life and Environmental Science) is not met
- More items are needed for Standard G (Earth and Space Science) and depth of knowledge consistency and range-of-knowledge for the standard is not met
- Standard H – Science in Social and Personal Perspectives doesn't meet any of the alignment criteria

Appendix D: Test Form Assembly Checklist

Subject: _____Grade level: _____Administration date: _____

Criteria	Yes	No	Comments
1. Correct # of test items? Identify number: _____			
2. Correct # of CR and SR items?			
3. Correct # of total points on test form? Identify number: _____			
4. Is test characteristic curve similar to previous years?			
5. Is test form difficulty similar to previous years?			
6. Do all items have acceptable point-biserial coefficients?			
7. Are items free of bias flags?			
8. Are items for each standard similar in difficulty to previous years?			
9. Correct # of reading passages for the reading test?			
10. Are the different types of reading passages represented on the test?			
11. Appropriate number of items per reading passage?			
12. Correct # of items per standard (categorical concurrence)?			
13. Are 65% of the items assigned to each standard at or above the indicated DOK level for the standard (DOK consistency)?			
14. Do at least half of the benchmarks for each standard have at least one related item (range of knowledge correspondence)?			
15. Is there an approximately equal distribution of items across each individual standard (balance of representation)?			
16. Do test items appear to be independent of each other?			
17. Does each test session have an approximately equal number of items?			
18. Are correct answer choices distributed approximately equally?			

Reviewer 1 Signature: _____ Date: _____

Reviewer 2 Signature: _____ Date: _____

Table 1: Reading Item Development

2007-08 New Reading Items				
	DOK 1	DOK 2	DOK 3	Total items
Grade 3	7	30	19	56
Grade 4	5	36	17	58
Grade 5	4	23	20	47
Grade 6	10	25	14	49
Grade 7	16	23	14	53
Grade 8	15	13	19	47
Total	**57**	**150**	**103**	**310**

2008-09 New Reading Items				
	DOK 1	DOK 2	DOK 3	Total items
Grade 3	12	22	23	57
Grade 4	10	26	23	59
Grade 5	8	32	18	58
Grade 6	5	22	25	52
Grade 7	4	26	25	55
Grade 8	7	22	26	55
Total	**46**	**150**	**140**	**336**

Table 2: Mathematics Item Development

2007-08 New Mathematics Items				
	DOK 1	DOK 2	DOK 3	Total items
Grade 3	5	20	14	39
Grade 4	5	18	16	39
Grade 5	2	25	13	40
Grade 6	1	21	19	41
Grade 7	13	19	7	39
Grade 8	12	27	9	48
Total	**38**	**130**	**78**	**246**

2008-09 New Mathematics Items				
	DOK 1	DOK 2	DOK 3	Total items
Grade 3	13	24	15	52
Grade 4	16	29	12	57
Grade 5	12	28	13	53
Grade 6	2	34	12	48
Grade 7	11	23	12	46
Grade 8	7	29	10	46
Total	**61**	**167**	**74**	**302**

References

CTB/McGraw-Hill. (2009). *Fall 2008 WKCE Technical Manual*. Monterey, CA: Author.

CTB/McGraw-Hill. (2008). *Fall 2007 WKCE Technical Manual*. Monterey, CA: Author.

CTB/McGraw-Hill. (2007). *Fall 2006 WKCE Technical Manual*. Monterey, CA: Author.

CTB/McGraw-Hill. (2006). *Fall 2005 WKCE-CRT Technical Manual*. Monterey, CA: Author.

CTB/McGraw-Hill. (2005). *WKCE November 2004 Testing Technical Report*. Monterey, CA: Author.

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. (2006). *Alignment Analysis of Mathematics Standards and Assessments*. A document submitted to the Wisconsin Department of Public Instruction. Madison, Wisconsin: Author.

Webb, N. L. (2006). *Alignment Analysis of Reading Standards and Assessments*. A document submitted to the Wisconsin Department of Public Instruction. Madison, Wisconsin: Author.

Webb, N. L. (2006). *Alignment Analysis of Science Standards and Assessments*. A document submitted to the Wisconsin Department of Public Instruction. Madison, Wisconsin: Author.