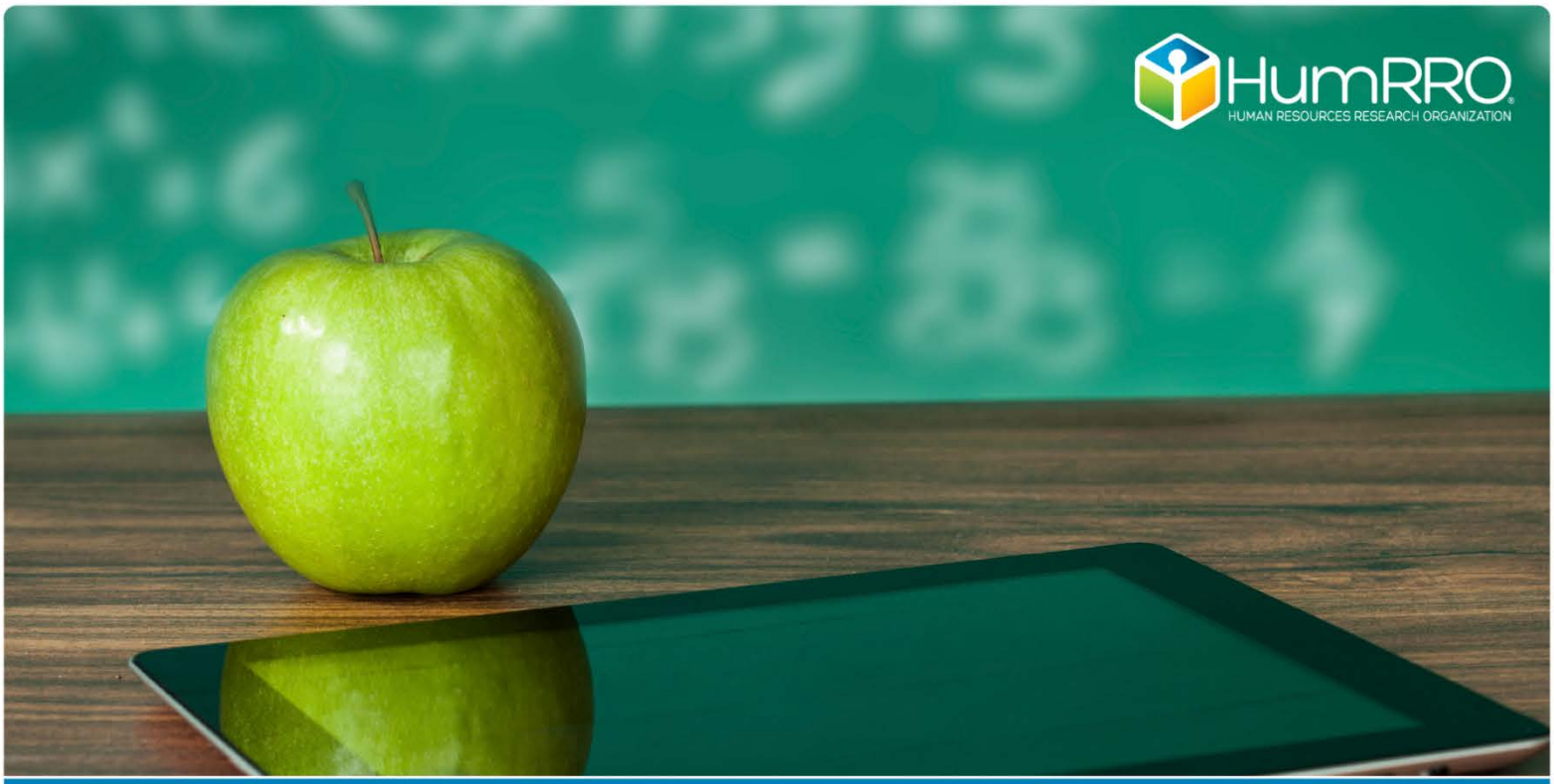

2019 No. 016

Independent Alignment Review of the Kentucky Performance Rating for Educational Progress (K-PREP) Science Assessments: Grades 4 and 7

Final Report

Prepared for:
Kentucky Department of Education
300 Sower Blvd, 5th Floor
Frankfort, KY 40601

Authors:
Arthur Thacker
Emily Dickinson
Richard Deatz

Date: May 13, 2019

Independent Alignment Review of the Kentucky Performance Rating for Educational Progress (K-PREP) Science Assessments: Grades 4 and 7

Table of Contents

List of Tables

Independent Alignment Review of the Kentucky Performance Rating for Educational Progress (K-PREP) Science Assessments: Grades 4 and 7

Executive Summary

The Human Resources Research Organization (HumRRO), an independent evaluator, conducted an alignment study for the Kentucky Department of Education (KDE) to investigate the alignment between the state's summative assessments in science for grades four and seven and the corresponding Kentucky Academic Standards (KAS) for Science. The KAS for Science are very similar to the Next Generation Science Standards (NGSS) and retain the same multidimensional structure. KAS for Science assesses students in Life Science, Physical Science, Earth and Space Science, and Engineering Design. The KAS for Science incorporates multiple dimensions of science study into these major topics. The dimensions include Disciplinary Core Ideas (DCI), Science and Engineering Principles (SEP), and Cross-Cutting Concepts (CCC). Kentucky's science assessment is similarly multidimensional to assess these complex interrelated topics.

Alignment Criteria

The Webb alignment method (1997,1999, 2005) was originally designed to align content standards with large-scale assessments. Dr. Norman Webb has researched and refined this method over time, and his approach is supported by the Council of Chief State School Officers (CCSSO).[1]

The Webb method includes four major indicators to evaluate alignment. These indicators rely on statistical analyses to assess how well items on the assessment, regardless of item type and point value, match the state's standards. The four alignment indicators are: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. While it was not appropriate to implement Webb's methodology for this study, we did use Webb's criteria to help guide our methodology and the development of criteria for judging the alignment of Kentucky's science assessments.

The table below summarizes the criteria used to evaluate the alignment of Kentucky's science assessment items to the KAS. Failure to meet a single criterion would not indicate that the test is insufficiently aligned to generate meaningful scores, but that attention to that aspect of the test should be addressed through future item development. If several of the criteria were not met, it would signal that we should be concerned with the link between the assessment and the intended measurement construct.

[1] See http://www.ccsso.org/Documents/2006/Creating_Aligned_Standards_2006.pdf for background information on alignment.

Science Assessment-to-Standards Alignment Criteria

Criteria	Description
Link to Standards	Acceptable if 50% or more of the items are directly and clearly matched to a specific KAS and at least 90% of items are matched to at least one KAS, DCI, SEP, or CCC.
DOK Adequacy	Acceptable if fewer than 10% of items are rated as DOK level 1 and more than 10% of items are rated at DOK level 3 or 4 (using Webb's DOK definitions).
Range Adequacy	Acceptable if at least 50% of CCC and SEP are aligned to test items (at least 4 CCC and 4 SEP)
Balance-of-Knowledge Correspondence (Revised for Science)	Webb's balance-of-knowledge correspondence criteria is used, computed for content domains and NGSS dimensions separately. Both must meet Webb's threshold of 0.70.
Multidimensional Adequacy	Acceptable if at least 90% of items are aligned to more than one dimension.

Method

HumRRO recruited science teachers for this alignment study from a list provided by KDE. The science teachers were geographically diverse and had recently worked on one or more special studies for KDE. We sent email invitations to approximately ten elementary and ten middle school teachers to fill 5 available slots in each panel group (grades 4 and 7). See Appendix A for the recruiting email. We filled all slots in the panel groups; however, a few days before the workshop we had one grade 4 teacher cancel. Our attempt to find a replacement at that point was unsuccessful. The 9 teachers represented all geographic areas of the state and five of the panelists had additional certifications (e.g., national board certification (2), special education, leadership or specialist training). Table 2 presents characteristics of the panelists.

HumRRO conducted the alignment study over a two-day period at a hotel in Louisville, Kentucky. In addition to the HumRRO facilitators, one HumRRO staff member was available throughout the workshop to assist with logistics. Prior to beginning their review, panelists read and signed affidavits of nondisclosure for the secure materials they would be reviewing during the workshop.

Panelists received specific instructions for rating the items. As a calibration activity, the HumRRO facilitators asked panelists to rate the first two items individually and discuss their ratings as a group. Once panelists were comfortable using the ratings, they continued the item rating activity on their own. Panelists rated the items one phenomenon at a time. This allowed them to rate groups of approximately eight items before discussing the items, ratings, and the phenomenon. Once consensus ratings were made for all the items associated with a phenomenon, panelists created a consensus statement regarding the suitability of the phenomenon for assessing the science concepts indicated by the items. Any discrepancies among the panelists for any item rating were thoroughly discussed and consensus ratings were recorded by the HumRRO facilitator. If panelists could not reach consensus, facilitators were instructed to note the lack of consensus in the spreadsheet comments section and record the majority ratings. This happened extremely rarely, and true consensus was reached for nearly every rating.

All panelists finished their rating tasks within the two days allotted for the workshop. Once panelists finished the review, they reviewed their consensus statements for each phenomenon and generated an overall consensus statement for the test as a whole. These statements were read aloud and edited/revised using a group process. The HumRRO facilitator did not participate in the consensus statement writing for the phenomena or for the test as a whole.

Results

Each panel group provided detailed consensus statements about each phenomenon, as well as about the test forms as a whole. Due to the level of item-level details in these statements, they are not included in the report, but rather will be shared with KDE separately. We do draw on elements of these statements, however, to further understand the calculated alignment statistics.

Grade 4

For both grade 4 test forms, all or nearly all items were found to be aligned to a KAS, DCI, SEP, or CCC. This provides strong evidence that science test scores reflect the intended content domain.

Neither grade 4 test form fully met the depth of knowledge criterion, but both included an appropriate percentage of Level 2 items. In fact, simply adding one item at DOK Level 3 and eliminating a DOK Level 1 item would have allowed both forms to meet this criterion. The group consensus statement indicated that depth of knowledge was fair and balanced across the clusters, but that some questions could have been raised to a higher level with some change in wording.

Form 1 from the grade 4 test fully met the range adequacy criterion but Form 2 did not. Form 2 was rated as reflecting an adequate number of SEP but not CCC. The balance criterion results indicate that on both forms, particular SEP or CCC tended to be emphasized more than others. The group consensus statement indicated that some practices that were not addressed could be included with the addition of other phenomena/storylines or other questions.

Finally, both grade 4 forms fell short of the multidimensional adequacy criterion. As noted in the group consensus statement, “The opportunities for students to actually engage in the science practices were strong throughout,” but “there were instances where engagement in the practices could have been more strongly assessed.”

Grade 7

For both grade 7 test forms, all or nearly all items were found to be aligned to a KAS, DCI, SEP, or CCC. This provides strong evidence that science test scores reflect the intended content domain.

Both grade 7 test forms fully met the depth of knowledge criterion, minimizing the number of Level 1 items and including an appropriate number of items at the higher DOK levels. The group consensus statement indicated that the inclusion of few DOK 1 items “represents a shift in science assessment from previous recall science assessment.”

Both grade 7 test forms fully met the range adequacy criterion. For both forms, there was an adequate number of CCC and SEP. The balance criterion results indicate that particular CCC or

SEP weren't overemphasized over others. The group consensus statement noted that the seven item clusters (phenomena) were "true to the intent of NGSS even if they occasionally missed the mark."

Finally, both grade 7 forms fell short of the multidimensional adequacy criterion. The group consensus statement indicated that they found that two clusters attempted to integrate engineering practices but did not do so successfully.

Discussion

Alignment is not an all-or-none judgment, but rather is a matter of degree. As stated above, failure to meet a single criterion would not indicate that the test is insufficiently aligned to generate meaningful scores, but that attention to that aspect of the test should be addressed through future item development. If several of the criteria were not met, it would signal that we should be concerned with the link between the assessment and the intended measurement construct.

Neither the grade 4 nor grade 7 assessment met all the alignment criteria evaluated. It is clear, however, that both tests are aligned to the content domain in the broadest sense. Neither test reflected the full breadth of the content domain. This is, in part, a product of the test design and item development processes, in which several items are written to the same KAS and associated dimensions. Reducing the number of items per cluster and increasing the number of clusters (and phenomena) is one approach that could increase content coverage in subsequent test versions. The grade 7 panelists also noted that simpler storylines might yield better alignment to the content and dimensions. The standards and dimensions that were measured by the item clusters tended to be balanced across the science domains but were slightly less balanced across the dimensions.

Neither grade level test fully met the multidimensional adequacy criterion, though the majority of items were rated as measuring two or more KAS and/or dimensions. Future item development efforts should place emphasis on integrating the KAS and dimensions, or on integrating multiple dimensions. It is important to note that group consensus statements reflected generally positive opinions about the science assessments at both grade levels. Teachers were pleased to see movement away from lower complexity test items and toward a test that allows students to demonstrate engagement in scientific practice.

Independent Alignment Review of the Kentucky Performance Rating for Educational Progress (K-PREP) Science Assessments: Grades 4 and 7

Introduction

Alignment studies address a vital question related to the validity of test scores—does the test content adequately reflect the content knowledge and skills that students are expected to learn as outlined in the state standards? School curriculum must be designed to meet the goals specified by the state standards and consequently assessments should measure the same content. This requirement is part of Federal law under Title I Section 1111(b)(3)[2] and Title VI[3] programming provided for within the Elementary and Secondary Education Act.

The Human Resources Research Organization (HumRRO), an independent evaluator, conducted an alignment study for the Kentucky Department of Education (KDE) to investigate the alignment between the state's summative assessments in science for grades four and seven and the corresponding Kentucky Academic Standards (KAS) for Science. The KAS for Science are very similar to the Next Generation Science Standards (NGSS) and retain the same multidimensional structure. KAS for Science assesses students in Life Science, Physical Science, Earth and Space Science, and Engineering Design. The KAS for Science incorporates multiple dimensions of science study into these major topics. The dimensions include Disciplinary Core Ideas (DCI), Science and Engineering Principles (SEP), and Cross-Cutting Concepts (CCC). Kentucky's science assessment is similarly multidimensional to assess these complex interrelated topics.

The study required convening a workshop consisting of panels of Kentucky educators and content experts. The panelists reviewed and evaluated the KAS for Science and operational test items from the K-PREP Science assessments for grades four and seven to evaluate the extent to which the operational test items reflect content knowledge and skills at the breadth and depth outlined in the content domain. Kentucky's assessment uses complex science phenomena, linked to test items sets, to assess the complex interrelated KAS for Science standards. This report describes the alignment method and results, along with discussion of the overall alignment of the assessments to the content standards.

Method

Several methods of alignment are in current use (e.g., Porter, 2002; Webb, 1997, 1999, 2005). These methods involve panelists evaluating several aspects of the content standards and test items. [4] The data from panelists' evaluations are analyzed statistically to determine the extent of alignment. HumRRO developed the methodology described herein to account for the multidimensional nature of the science standards and for the phenomenon-based assessments used by Kentucky.

[2] See https://www2.ed.gov/policy/elsec/leg/esea02/pg2.html for Federal law, Title I Section 1111(b)(3)
[3] See https://www2.ed.gov/about/offices/list/ocr/docs/hq43e4.html for Federal law, Title VI
[4] See http://programs.ccsso.org/projects/surveys_of_enacted_curriculum/understanding_alignment_analysis/

Alignment Criteria

The Webb alignment method (1997,1999, 2005) was originally designed to align content standards with large-scale assessments. Dr. Norman Webb has researched and refined this method over time, and his approach is supported by the Council of Chief State School Officers (CCSSO).[5]

The Webb method includes four major indicators to evaluate alignment. These indicators rely on statistical analyses to assess how well items on the assessment, regardless of item type and point value, match the state's standards. The four alignment indicators are: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. While it was not appropriate to implement Webb's methodology for this study, we did use Webb's criteria to help guide our methodology and the development of criteria for judging the alignment of Kentucky's science assessments. Below, we briefly describe Webb's criteria, and the similar criteria used for the Kentucky science assessments.

Webb's ***Categorical concurrence*** is a basic measure of alignment between content standards and test items. This term refers to the proportion of overlap between the content stated in the standards document and that assessed by items on the test. Webb's criterion is based on the minimum number of items required to achieve acceptable reliability for reporting. We prefer to directly examine the reliability of the science assessments, which will be available in the forthcoming technical report[6] for the K-PREP. Reliability of scores should be evaluated for overall science scores at the student level and for sub-scores computed at the aggregate level for schools, districts, or the state.

Webb's categorical concurrence criterion is derived by determining if there are at least six items per reporting category on the assessment. Kentucky does not report sub-categories for students on the science assessments in grades four and seven. So, at the most basic level, Kentucky meets Webb's criteria if at least six items per form can be matched to any science standards. This would not be a robust criterion.

The KAS are written as performances or tasks through which students can demonstrate understanding of the content. These expectations were developed based on the DCI, SEP, and CCC the students are expected to have learned at each grade level. Test items might directly address the KAS, or they might address the supporting DCI, SEP, or CCC. Ideally, an item would be linked to both a KAS and some number of DCI, SEP, or CCC, but that may not always be possible given the relatively discrete nature of selected-response test items. It may be necessary to address all aspects of a standard through multiple test items.

For this criterion, we report the proportion of items that panelists match to the KAS for science. The proportions also indicate the number of items not judged to relate to any KAS (32 items per form). To be judged acceptable, at least 50% of the test items should be directly matched to a KAS. We use 50% match to KAS as one component of this criteria because we expect some items to be matched only to DCI, SEP, or CCC. Ideally, all items would match at least one KAS, DCI, SEP, or CCC. However, it is possible for an assessment to have acceptable alignment with one or two weak items (as judged by panelists). To be judged acceptable for the second component of this criterion, at least 90% of items should be matched to either a KAS, DCI, SEP,

[5] See http://www.ccsso.org/Documents/2006/Creating_Aligned_Standards_2006.pdf for background information on alignment.
[6] The technical report will be authored by Pearson and publicly available through KDE.

or CCC. To be judged acceptable, the test form must meet both components. We will refer to this criterion as **Link to Standards**.

Webb's ***Depth-of-knowledge*** (DOK) ***consistency*** statistic measures the type of cognitive processing required by items compared to the cognitive processing required by the matched content standards. For example, is a student expected to simply identify or recall basic facts, to use reason to manipulate information, or to strategize how to best solve a complex problem? Using science as an example, a student may be asked to identify the planets of our solar system among several answer choices. This task should be less complex than comparing the composition of the planets in preparation for landing unmanned probes.

The purpose of using DOK as a measure of alignment is to determine whether a test item and its corresponding standard are written at the same level of cognitive complexity. In Webb's method, panelists make two separate judgments about cognitive complexity, one rating for the standard and one rating for the item. These two judgments are compared to determine whether the item is written at the same level as the standard to which it is linked. Webb (1997) refers to this comparison as *Depth-of-Knowledge consistency*.

Webb's DOK consistency category is nearly impossible to implement when the standards are multi-dimensional. Doing so would require panelists to determine the DOK for each potential combination of standard and dimension. For science, it is also the case that the test standards can be interpreted in multiple ways and each combination of standard and dimension would represent a range of cognitive complexities depending on the specific knowledge, skills, and abilities that were being addressed. So, even if we could generate the number of DOK ratings required by the science standards, our ratings would likely be vague, unreliable, and inflated (Webb's rule is to assign the higher DOK level if the standard is ambiguous). No attempt was made to match item DOK with standard DOK for this study.

Kentucky's science assessment is based on clusters of items centered around scientific phenomena. It might be more appropriate to consider whether the assessments reached the desired cognitive complexity at the cluster level as well. Each cluster contains items that may exhibit a range of cognitive complexity, and the overall complexity of the cluster may be greater than all the individual items. It may also be more appropriate to choose a different scale for cognitive complexity than DOK if these ratings were made at the cluster level. However, for this study, we used DOK since it is the most commonly used metric for cognitive complexity reported in alignment studies and because DOK was considered during item development and is included in the meta-data provided by the contractor.

It is still, however, important to determine if science test items reflect the level of cognitive complexity indicated by the science standards. If we look at the standards more globally, we find that they focus on requiring students to use their science knowledge and skills to investigate potentially unfamiliar phenomena. Focusing on science in this way means that students are expected to engage in more complex reasoning than simply recalling science terms or generating simple answers using familiar algorithms. We therefore reasoned that Kentucky's science assessment should include few, if any, low-complexity items. Webb uses a four-point scale for DOK. For an assessment based on KAS for Science, we would expect no more than 10% of items to be rated at level one. Webb's scale also includes a level four rating, which is seldom met on summative tests. This level of cognitive processing requires deep engagement of the students with the content, in multiple ways, typically over an extended period of time. This level is similar to producing a thesis or generating an extensive investigation of some scientific phenomenon a student observes, collects data on, and reports about. We do not expect

Kentucky's assessments to include level four items. We would expect the assessments to be primarily a mix of DOK level two and three items. We would also expect more level two items than level three items. Level three items require more input or time for students to respond, and it would not be practical to include primarily level three items on a summative assessment. We set Kentucky's DOK acceptability criterion such that no more than 10% of items are rated at level 1 and no less than 10% of items are rated at level 3. If there are more than 10% of items at level 1 or fewer than 10% of items at level 3, the DOK level of the items as a group would be judged too low to adequately represent the KAS for science. We will refer to this criterion as **DOK Adequacy**.

Webb's ***Range-of-knowledge correspondence*** examines the extent to which the test items reflect the full range of knowledge, skills, and abilities contained in the standards document. Where categorical concurrence notes whether a sufficient number of items on the test covers each general content topic (reporting category), the range-of-knowledge correspondence measure indicates the number of specific content objectives within each broader topic that are assessed by the test items.

Webb's range-of-knowledge correspondence criterion requires that at least 50% of the standards from each reporting category are addressed on the assessment. We stated above that Kentucky intends to report students' overall science scores, but not finer-grained sub-scores (e.g., physical science, life science). Meeting Webb's range-of-knowledge criterion would thus require that at least half of the KAS for science be represented on the tests. Given the three-dimensional nature of the standards, this criterion is not practical. The number of potential combinations of domains and dimensions represent too many standards to address in any single testing event, and that would not be necessary to sample the standards, in any regard. The standards emphasize students making meaning from information gathered from new or unfamiliar phenomena. They are expected to have a deep understanding of SEPs and CCCs, and that knowledge is expected to provide tools to use across DCIs in all content domains. We will focus on SEPs and CCCs for this criterion rather than on trying to address the full breadth of the KAS for science.

Because students are expected to use their knowledge of SEPs and CCCs across multiple standards and content domains, we would expect these dimensions to be high priorities on Kentucky's science assessments. We also expect for there to be few, if any, items on the tests that measure only a single SEP or CCC, and that these concepts are measured in context with DCIs from legitimate scientific phenomena. Items are coded to indicate if they measure an SEP or CCC, or both. We would expect at least 50% of the SEPs and CCCs to be directly measured by items on the tests. There are eight SEPs and seven CCCs. The assessments should contain items that address at least 4 SEPs and 4 CCCs to meet this criterion. We will refer to this criterion as **Range Adequacy**.

Webb's ***Balance-of-knowledge representation*** focuses on content coverage in yet more detail. In this case, the number of items matched to the content objective does matter. The balance of representation criterion determines whether the assessment measures the content objectives equitably within each content topic using only those content objectives identified by panelists as measured by the test item. Based on Webb's (1997) method, items should be distributed evenly across the objectives per content topic for good balance. The balance-of-knowledge representation is determined by calculating an index, or score, for each content topic. Each topic should meet or surpass a minimum index level to demonstrate adequate balance.

It would not be possible to compute a single interpretable balance-of-knowledge representation index for a three-dimensional assessment. The interaction of the dimensions and domains would yield too many objectives to include on a summative test form. It does, however, make sense to consider that each content domain should be represented rather evenly, or purposefully, on an assessment. It might also be sensible to declare that the three dimensions should be represented rather evenly, or purposefully, on an assessment. Acceptability for Kentucky's science test will be determined using the same metric as Webb uses for balance-of-knowledge correspondence with the notable exception that it will be computed twice; once for domain, and again for dimension. Acceptability for each will be set at the same level Webb uses for traditional assessments (0.70). Both balance criteria must be met for the assessment to be considered adequately aligned. We will refer to this criterion as **Balance-of-Knowledge Correspondence (Revised for Science)**, or simply as **Balance**.

In addition, Kentucky's test items are written to be multi-dimensional. They are intended to measure more than isolated science content knowledge and are expected to address CCC and SEP in addition to DCI and/or specific KAS. To address whether the items accomplish this goal, we evaluate whether panelists agree that items are related to multiple science concepts across DCI, CCC, and SEP. To be judged acceptable, at least 90% of items should address more than one dimension. We will refer to this criterion as **Multidimensional Adequacy**.

Table 1 summarizes the criteria used to evaluate the alignment of Kentucky's science assessment items to the KAS. Failure to meet a single criterion would not indicate that the test is insufficiently aligned to generate meaningful scores, but that attention to that aspect of the test should be addressed through future item development. If several of the criteria were not met, it would signal that we should be concerned with the link between the assessment and the intended measurement construct.

Table 1. Science Assessment-to-Standards Alignment Criteria

Criteria	Description
Link to Standards	Acceptable if 50% or more of the items are directly and clearly matched to a specific KAS and at least 90% of items are matched to at least one KAS, DCI, SEP, or CCC.
DOK Adequacy	Acceptable if fewer than 10% of items are rated as DOK level 1 and more than 10% of items are rated at DOK level 3 or 4 (using Webb's DOK definitions).
Range Adequacy	Acceptable if at least 50% of CCC and SEP are aligned to test items (at least 4 CCC and 4 SEP)
Balance-of-Knowledge Correspondence (Revised for Science)	Webb's balance-of-knowledge correspondence criteria is used, computed for content domains and NGSS dimensions separately. Both must meet Webb's threshold of 0.70.
Multidimensional Adequacy	Acceptable if at least 90% of items are aligned to more than one dimension.

Scope of Alignment Evaluation

The alignment evaluation performed for this study involved a comparison of the science operational test items to the KAS. Highly qualified educators provided alignment ratings for the evaluation. To maintain the independent and external nature of the study, KDE and their testing contractor, Pearson, did not take an active part in this process. Although Pearson staff did provide access to the assessments and to relevant item data and statistics, the alignment process was conducted and directed solely by HumRRO.

Panelists

HumRRO recruited science teachers for this alignment study from a list provided by KDE. The science teachers were geographically diverse and had recently worked on one or more special studies for KDE. We sent email invitations to approximately ten elementary and ten middle school teachers to fill 5 available slots in each panel group (grades 4 and 7). See Appendix A for the recruiting email. We filled all slots in the panel groups; however, a few days before the workshop we had one grade 4 teacher cancel due to a family situation. Our attempt to find a replacement at that point was unsuccessful. The 9 teachers represented all areas of the state and 56% of the panelists had additional certifications (e.g., national board certification (2), special education, leadership or specialist training). Table 2 presents characteristics of the panelists.

Table 2. Professional and Demographic Characteristics of Panelists

Number of Panelists	Average Years Experience (SD)	Percent Master's Degree or higher	Gender		School Community		
			Female	Male	Rural	Suburban	Urban
9	21 (7.71)	89%	89%	11%	44%	22%	33%

Review of Content Alignment

The review involved three major tasks: (a) determining the item DOK for each test item, (b) evaluating the science test items by verifying KAS the item is intended to measure, plus any DCI, CCC, or SEP the item is intended to address, and (c) creating a consensus summary statement regarding the adequacy of each of the scientific phenomena used on the test forms for eliciting information regarding students' knowledge of the content. All ratings were recorded on specially designed Excel® spreadsheets. Panelists independently made all ratings for test items, then discussed to reach final consensus ratings. Initial ratings were kept by panelists and the HumRRO facilitator recorded the consensus ratings. Summary statements for phenomena were developed and recorded entirely by panelists.

Training

Training is an essential part of any alignment study, for both facilitators and panelists. Even though the HumRRO facilitators were very experienced, every alignment study is unique. Therefore, facilitators were required to attend a 90-minute alignment training session before the alignment workshop convened. Facilitators were trained on the specifics of the assessment

system, and they conducted a step-by-step walkthrough of the alignment tasks for which they were to guide panelists to complete.

Panelists' training began the first day of the alignment workshop. After introducing HumRRO staff, panelists received familiarization training on assessment-to-standards alignment terms and general processes, in addition to a detailed discussion of DOK. Panelists were also trained on the access and use of the test items. Then, panelists were dismissed to their individual panel groups to receive additional training, which is discussed in more detail below.

Materials

During the alignment workshop panelists viewed test items (as students see them). Items are grouped within phenomena, each of which supports approximately 8 items. The panelists evaluated the alignment of the items and completed Excel®. rating forms. . The test items and rating forms are discussed next.

Test Items. Panelists evaluated all Kentucky operational test items for grades 4 and 7 science from the Spring 2018 assessment. It is important to note that the operational test forms contain both common and matrix items. All students completed a common set of items related to one phenomenon (eight items). The remainder of the items differed by form. Kentucky administered six forms in 2018. Each form contained the common phenomenon (eight items), plus three additional phenomena (approximately 24 items). Kentucky used a total of seven phenomena for each grade level. Six forms were created by altering the order of the phenomena by form. For alignment purposes the order of the items does not impact the computation of the ratings or the generation of the criteria statistics. We were therefore able to treat six forms as if they were two. Alignment computations were made for two forms, representing two sets of items (8 common items plus 24 unique items (in differing orders) per form).

A complete test form contains seven or eight additional items (pilot test items) that were not reviewed or included in the analysis because those items are not included in a student's score. Alternate assessment test items were not reviewed for this study. Because the test items are secure, this report does not include any examples of items or references to specific item content.

Rating Forms and Instructions. Panelists were given instruction describing the rating tasks, the codes to be used, and the data entry Excel® files for data entry of consensus and individual ratings. Table 3 provides the spreadsheet headings the panelists saw. Panelists completed the non-shaded fields from Table 3. The shaded fields contained item meta-data and other information to assist panelists in making their ratings. The Excel® spreadsheet included data validation in the fields to limit coding errors. The actual coding form contains information related to the scientific phenomena described on the tests, which is confidential, and is therefore not provided in this report.

Table 3. Fields from Data-Coding Spreadsheet Used by Alignment Panelists

Spreadsheet Data Field	Description
Grade	Either Grade 04 or 07
UIN	Unique Item Number
Item Type	MC-Multiple Choice, MS-Multiple Select, OR-Open Response
Max Points	Total number of points the item is worth
Phenomenon	Brief descriptor of the phenomena associated with the item
Assign an item depth of knowledge (DOK) rating. 1=Recall 2=Skill/Concept 3=Strategic Thinking 4 = Extended Thinking	
KY Academic Standard (KAS) to which item was written	Code from item meta-data
Do you agree that the item measures this KAS? 0 = No 1 = Yes	
Disciplinary Core Idea (DCI) #1	Code from item meta-data
Do you agree that the item measures this DCI? 0 = No 1 = Yes	
Disciplinary Core Idea (DCI) #2	Code from item meta-data
Do you agree that the item measures this DCI? 0 = No 1 = Yes	
Cross Cutting Concept (CCC) #1	From item meta-data
Do you agree that the item measures this CCC? 0 = No 1 = Yes	
Cross Cutting Concept (CCC) #2	From item meta-data
Do you agree that the item measures this CCC? 0 = No 1 = Yes	
SEP #1	From item meta-data
Do you agree that the item measures this SEP? 0 = No 1 = Yes	
SEP #2	From item meta-data
Do you agree that the item measures this SEP? 0 = No 1 = Yes	
SEP #3	From item meta-data
Do you agree that the item measures this SEP? 0 = No 1 = Yes	
Comments	Panelists were encouraged to explain any “No” rating and to comment if there were any issues with the item. Panelists did not review items for quality, bias, or sensitivity, but were asked to note such issues here if they saw them.

Procedures

HumRRO conducted the alignment study over a two-day period at a hotel in Louisville, Kentucky. In addition to the HumRRO facilitators, one HumRRO staff was available throughout the workshop to assist with logistics. Prior to beginning their review, panelists read and signed affidavits of nondisclosure for the secure materials they would be reviewing during the workshop.

Before beginning each of the rating tasks, the HumRRO facilitators trained panelists on the procedures to complete the task, discussed the rating criteria, and facilitated a short calibration activity to ensure panelists were comfortable applying ratings. HumRRO facilitators provided general suggestions and comments when appropriate; however, they emphasized that their role was not to provide explicit direction on how to rate items because panelists were valued as the content experts. Each panelist was assigned a laptop loaded with rating forms. Items were accessed via paper forms (the same format used by students in 2018).

Panelists received specific instructions for rating the items. As a calibration activity, the HumRRO facilitators asked panelists to rate the first two items individually and discuss their ratings as a group. Once panelists were comfortable using the ratings, they continued the item rating activity on their own. Panelists rated the items one phenomenon at a time. This allowed them to rate groups of approximately eight items before discussing the items, ratings, and the phenomenon. Once consensus ratings were made for all the items associated with a phenomenon, panelists created a consensus statement regarding the suitability of the phenomenon for assessing the science concepts indicated by the items. Any discrepancies among the panelists for any item rating were thoroughly discussed and consensus ratings were recorded by the HumRRO facilitator. If panelists could not reach consensus, facilitators were instructed to note the lack of consensus in the spreadsheet comments section and record the majority ratings. This happened extremely rarely and true consensus was reached for nearly every rating.

All panelists finished their rating tasks within the two days allotted for the workshop. Once panelists finished the review, they reviewed their consensus statements for each phenomenon and generated an overall consensus statement for the test as a whole. These statements were read aloud and edited/revised using a group process. The HumRRO facilitator did not participate in the consensus statement writing for the phenomena or for the test as a whole.

Results

The following section summarizes the results from the analysis of panelists' ratings.

Results for Science Alignment Criteria

All of Webb's (1997) measures begin with calculations for each panelist and build up to a summary of results across panelists. For science, it is problematic to summarize ratings across panelists. The CCC and SEP are, by design, highly integrated. For example, a student might be required to determine the expected proportion of pea plants with a specific characteristic in the third generation given the genotype of the parents. This task would require mathematics, proportional reasoning, and identification of patterns, as well as science knowledge regarding the way genotypes combine to produce specific phenotypes in offspring. Depending on how the test item was constructed, students might be required engage in multiple thinking tasks addressing several aspects of the science standards. However, it is usually the case that a test

item is designed to elicit a particular strategy or method, and therefore access some concepts more directly than others. We find that it is better to have panelists discuss the way students would approach the item and come to conclusions based on the interaction among the members regarding the "best" alignment of the item. For that reason, panelists did generate individual ratings, but they then shared those ratings among the group. This allowed for rich discussion and consensus regarding the alignment of the item to the standards. All criteria are computed based on the consensus ratings.

Link to Standards

Link to Standards describes the extent to which the science items, regardless of item type and point value, align to specific KAS. This criterion was evaluated by computing the percentage of items panelists indicated as "directly measuring aspects of the intended standard" and compared to the total number of operational test items on the form. Panelists were provided the KAS that the item was intended to measure from item meta-data and they were trained to indicate a match if the item directly addressed aspects of the standard. Items were not expected to cover the full breadth of the standard but should address key standard components.

Table 4 summarizes the results for the Link to Standards criterion. Only Form 2 from grade seven contained any item that was not matched to either a KAS, DCI, SEP, CCC, or some combination. The great majority of items were matched to a KAS as well as one or more dimensions. Panelists considered several items from one phenomenon in grade 7 to not adequately address the KAS on which the phenomenon was based, but most of those items did address one or more dimensions of science. This criterion was met for all forms for grades 4 and 7.

Table 4. Summary of Link to Standards Results

Grade/Form	Percent of Item Matched to KAS	Meets Component #1	Percent of Item Matched to KAS, DCI, SEP, or CCC	Meets Component #2	Meets Criterion
Grade 4/Form 1	93.75%	Yes	100%	Yes	Yes
Grade 4/Form 2	90.63%	Yes	100%	Yes	Yes
Grade 7/Form 1	90.63%	Yes	100%	Yes	Yes
Grade 7/Form 2	71.85%	Yes	97%	Yes	Yes

Depth-of-Knowledge Adequacy

Analyses of depth-of-knowledge (DOK) measure the type of cognitive processing required of students (Webb, 1997, 2005). DOK adequacy indicates that the DOK level of the assessment items is sufficient to elicit the kind of cognitive processing indicated by the standards. These analyses are typically done by comparing the DOK of items to that the standard to which that item is matched. For the KAS, and other three-dimensional standards, establishing the DOK of the standards is impractical. Instead, we rely on the standards to inform item DOK more generally. The standards expect students to go beyond simple recall of information toward reasoning and problem solving in unfamiliar contexts. For that reason, we expect DOK Level 1 items to be rare (less than 10%). We also expect the assessments to include some high DOK Level items (Level 3), which require students to engage deeply with the item content to reason, use evidence, and generate hypotheses. We expect at least 10% of items to be at Level 3 or higher. Level 4 items

are impractical to include on a summative assessment, as they typically require an extended period of time to answer. Panelists rated item DOK but did not rate DOK for KAS.

To make their ratings, panelists used a rating scale (adapted from Webb, 2005) with four levels of cognitive complexity.

- Level 1 Recognition – simple recall of information (i.e., facts, terms); sequencing; more automatic.
- Level 2 Skills/Concepts – beyond habitual response; applying concepts; problem-solving.
- Level 3 Strategic Thinking – requires basic reasoning, planning, or use of evidence; generating hypotheses.
- Level 4 Extended Thinking – complex reasoning; evaluation of multiple sources or independent pieces of evidence; often over an extended period of time.[7]

Table 5 summarizes the DOK adequacy results. Both grade seven forms met this criterion, but the grade 4 forms did not. However, adding one item at DOK Level 3 and eliminating a DOK Level 1 item would have allowed both grade four forms to meet this criterion.

Table 5. Summary of DOK Adequacy Results

Grade/Form	DOK 1	DOK 2	DOK 3/4	Meets Criterion
Grade 4/Form 1	12.5%	78.1%	9.4%	No
Grade 4/Form 2	9.4%	81.3%	9.4%	No
Grade 7/Form 1	9.4%	78.1%	12.5%	Yes
Grade 7/Form 2	9.4%	75.0%	15.6%	Yes

Range Adequacy

The *range adequacy* criterion examines in greater detail the breadth of knowledge covered by the assessment. For science, we define this criterion in terms of the CCC and SEP addressed by test items. To be rated as acceptable, test items should directly assess at least 50% of the CCC and SEP described in the standards. This means that each test form should have items that directly assess at least 4 CCC and 4 SEP.

This criterion was determined based on panelists' indications of the whether items assessed the intended CCC and/or SEP indicated by the item meta-data. If panelists indicated that they agreed, then that CCC or SEP was counted. The total number of unique CCC and SEP for each form is included in Table 6. Both numbers must be greater than 4 to meet the criterion. Only form 2 from grades 4 failed to meet this criterion. This may be because the phenomena tended to focus on a single CCC. Each form included only four phenomena, which may limit the range of the assessment for addressing all the dimensions included in the KAS.

[7] The full item DOK guidance used by panelists is included in the appendices.

Table 6. Summary of Range Adequacy Results

Grade/Form	Unique CCC	Unique SEP	Meets Criterion
Grade 4/Form 1	6	5	Yes
Grade 4/Form 2	3	6	No
Grade 7/Form 1	7	4	Yes
Grade 7/Form 2	6	5	Yes

Balance-of-Knowledge Representation (Revised for Science)

Webb's (1997) method includes a *balance-of-knowledge representation statistic*. This measure describes the distribution of items linked to each standard within each strand. The number of items should be distributed rather evenly between the strands to achieve good balance.

The content balance is determined by calculating an index, or score, for each strand.[8] According to Webb (1997), the minimum acceptable index for a single content strand is 70 (on a scale of 0 to 100 with 100 representing perfect balance). An index of 70 or higher suggests that items broadly assess the standards within a strand instead of clustering around one or two standards.

It is important to note that only those standards that were indicated by panelists as being aligned to an item are included in calculations of the balance index. A given strand may include more standards than were verified by panelists as being linked to items. Recognizing this feature of the balance index is important in cases when the range measure and balance measure produce seemingly contrasting results.

For science, we must also account for the multi-dimensional nature of the standards in our consideration of balance. Kentucky's science assessment is designed to assess the DCI, SEP and CCC through phenomena that also address the life sciences, physical sciences, and earth and space sciences. Ideally, the test would be balanced such that neither of these aspects of science are emphasized more than the other.

Tables 7- 14 present the results for balance-of-knowledge representation. Results for balance are presented for science domains and dimensions separately. Given the nature of Kentucky's assessment, it is not surprising that the balance criterion is not always met. Kentucky uses clusters of items grouped by scientific phenomena. Each test form includes only 4 clusters. The clusters are necessarily focused on a few standards. It may not always be possible to create an assessment form that is balanced across science domain and dimension. When forms do not meet the balance criterion, table notes indicate the areas of emphasis on the test forms. Only Form 1 from grade 7 meets the balance criterion for all domains and dimensions. The other forms tend to be well-balanced across most components.

[8] The exact formula for calculating the balance index is explained in detail in Webb's (2005) alignment training manual: http://www.wcer.wisc.edu/WAT/index.aspx.

Table 7. Content Domain Balance-of-Knowledge Representation: Grade 4 Science (Form 1)

Science Domain	Standards Linked with Items	Number of Items Linked	% of Items Linked to Domain	Balance Index	Balance Index Target Met
Physical Science	4	9	30.0	72	Yes
Life Science	2	10	33.3	90	Yes
Earth and Space Science	3	8	26.7	83	Yes
Engineering, Technology and the Application of Science	1	3	10.0	100	Yes

Table 8. Science Dimension Balance-of-Knowledge Representation: Grade 4 Science (Form 1)

Science Dimensions	Elements Linked with Items	Number of Items Linked	% of Items Linked to Component	Balance Index	Balance Index Target Met
DCI	10	30	93.8	77	Yes
SEP	5	30	93.8	60	No
CCC	7	29	90.6	62	No

SEPs emphasize Constructing Explanations and Designing Solutions. CCCs emphasize Cause and Effect and Patterns.

Table 9. Content Domain Balance-of-Knowledge Representation: Grade 4 Science (Form 2)

Science Domain	Standards Linked with Items	Number of Items Linked	% of Items Linked to Domain	Balance Index	Balance Index Target Met
Physical Science	5	11	37.9	78	Yes
Life Science	1	7	24.1	100	Yes
Earth and Space Science	3	10	34.5	83	Yes
Engineering, Technology and the Application of Science	1	1	3.4	100	Yes

Table 10. Science Dimension Balance-of-Knowledge Representation: Grade 4 Science (Form 2)

Science Dimensions	Elements Linked with Items	Number of Items Linked	% of Items Linked to Component	Balance Index	Balance Index Target Met
DCI	9	31	96.9	75	Yes
SEP	6	30	93.8	70	Yes
CCC	4	30	93.8	60	No

CCCs emphasize Cause and Effect and Patterns.

Table 11. Content Domain Balance-of-Knowledge Representation: Grade 7 Science (Form 1)

Science Domain	Standards Linked with Items	Number of Items Linked	% of Items Linked to Domain	Balance Index	Balance Index Target Met
Physical Science	3	12	41.4	100	Yes
Life Science	2	7	24.1	79	Yes
Earth and Space Science	3	7	24.1	90	Yes
Engineering, Technology and the Application of Science	2	3	10.3	83	Yes

Table 12. Science Dimension Balance-of-Knowledge Representation: Grade 7 Science (Form 1)

Science Dimensions	Elements Linked with Items	Number of Items Linked	% of Items Linked to Component	Balance Index	Balance Index Target Met
DCI	7	29	90.6	81	Yes
SEP	4	22	68.8	77	Yes
CCC	7	24	75.0	83	Yes

Table 13. Content Domain Balance-of-Knowledge Representation: Grade 7 Science (Form 2)

Science Domain	Standards Linked with Items	Number of Items Linked	% of Items Linked to Domain	Balance Index	Balance Index Target Met
Physical Science	3	10	41.7	87	Yes
Life Science	1	8	33.3	100	Yes
Earth and Space Science	2	4	16.7	100	Yes
Engineering, Technology and the Application of Science	1	2	8.3	100	Yes

Table 14. Science Dimension Balance-of-Knowledge Representation: Grade 7 Science (Form 2)

Science Dimensions	Elements Linked with Items	Number of Items Linked	% of Items Linked to Component	Balance Index	Balance Index Target Met
DCI	5	25	78.1	68	No
SEP	5	16	50.0	90	Yes
CCC	6	21	65.7	74	Yes

DCIs emphasize LS2.A.

Multidimensional Adequacy

This criterion indicates if an acceptable proportion of items measures more than one science dimension (DCI, CCC, or SEP). To be considered multidimensional an item may measure either more than one dimension (e.g. a DCI plus a CCC), or it may measure more than one standard within a single dimension (e.g. two SEP). An assessment is considered acceptable if more than 90% of items are multidimensional.

Panelist indicated if items measured each intended DCI, CCC, and SEP based on the information provided in the item meta-data. If panelists agreed that an item measured a particular CCC, then that CCC counted toward the item's total number of measured dimensions. If panelists indicated that the item did not measure the CCC, it was not counted toward the item's total number of measured dimensions. An item was considered multidimensional if panelists indicated two or more dimensions were measured by the item. Results are presented in Table 8.

Table 15. Multidimensional Adequacy Results

Grade/Form	Percent of Multidimensional Items	Meets Criterion
Grade 4/Form 1	81.3%	No
Grade 4/Form 2	84.4%	No
Grade 7/Form 1	87.5%	No
Grade 7/Form 2	81.3%	No

Summary

Each panel group provided detailed consensus statements about each phenomenon, as well as about the test forms as a whole. Due to the level of item-level details in these statements, they are not included in the report, but rather will be shared with KDE separately. We do draw on elements of these statements, however, to further understand the calculated alignment statistics. In this section, we will summarize across the criteria for each grade level test, incorporating high level details from consensus statements as appropriate.

Grade 4

For both grade 4 test forms, all or nearly all items were found to be aligned to a KAS, DCI, SEP, or CCC. This provides strong evidence that science test scores reflect the intended content domain.

Neither grade 4 test form fully met the depth of knowledge criterion, but both included an appropriate percentage of Level 2 items. In fact, simply adding one item at DOK Level 3 and eliminating a DOK Level 1 item would have allowed both forms to meet this criterion. The group consensus statement indicated that depth of knowledge was fair and balanced across the clusters, but that some questions could have been raised to a higher level with some change in wording.

Form 1 from the grade 4 test fully met the range adequacy criterion but Form 2 did not. Form 2 was rated as reflecting an adequate number of SEP but not CCC. The balance criterion results indicate that on both forms, particular SEP or CCC tended to be emphasized more than others.

The group consensus statement indicated that some practices that were not addressed could be with the addition of other phenomena/storylines or other questions.

Finally, both grade 4 forms fell just short of the multidimensional adequacy criterion. As noted in the group consensus statement, “The opportunities for students to actually engage in the science practices were strong throughout,” but “there were instances where engagement in the practices could have been more strongly assessed.”

Grade 7

For both grade 7 test forms, all or nearly all items were found to be aligned to a KAS, DCI, SEP, or CCC. This provides strong evidence that science test scores reflect the intended content domain.

Both grade 7 test forms fully met the depth of knowledge criterion, minimizing the number of Level 1 items and including an appropriate number of items at the higher DOK levels. The group consensus statement indicated that the inclusion of few DOK 1 items “represents a shift in science assessment from previous recall science assessment.”

Both grade 7 test forms fully met the range adequacy criterion. The balance criterion results also indicate that particular CCC or SEP weren’t overemphasized over others. The group consensus statement noted that the seven item clusters were “true to the intent of NGSS even if they occasionally missed the mark.”

Finally, both grade 7 forms fell just short of the multidimensional adequacy criterion. The group consensus statement indicated that they found that two clusters attempted to but did not successfully integrate engineering practices.

Discussion

Alignment is not an all-or-none judgment, but rather is a matter of degree. As stated above, failure to meet a single criterion would not indicate that the test is insufficiently aligned to generate meaningful scores, but that attention to that aspect of the test should be addressed through future item development. If several of the criteria were not met, it would signal that we should be concerned with the link between the assessment and the intended measurement construct.

Neither the grade 4 nor grade 7 assessment met all the alignment criteria evaluated. Some were met fully, and others were met for a subset of domains or dimensions. It is clear, however, that both tests are aligned to the content domain in the broadest sense. Neither test reflected the full breadth of the content domain. This is, in part, a product of the test design and item development processes, in which several items are written to the same KAS and associated dimensions. Reducing the number of items in a cluster and increasing the number of clusters is one approach that could increase content coverage in subsequent test versions. The grade 7 panelists also noted that simpler storylines might yield better alignment to the content and dimensions.

The standards and dimensions that were measured by the item clusters tended to be balanced across the science domains but were less balanced across the dimensions. In grade 4, there was an overemphasis on Constructing Explanations and Designing Solutions on one form. Both forms were rated as having an overemphasis on Cause and Effect and Patterns, as compared

to other CCC. The grade 7 test forms were generally well-balanced, though Form 2 overemphasized one life sciences DCI compared to the other aligned DCI.

Neither grade level test fully met the multidimensional adequacy criterion, though the majority of items were rated as measuring two or more KAS and/or dimensions. Future item development efforts should place emphasis on integrating the KAS and dimensions, or on integrating multiple dimensions. It is important to note that group consensus statements reflected generally positive opinions about the science assessments at both grade levels. Teachers were pleased to see movement away from lower complexity test items and toward a test that actually allows students to demonstrate engagement in scientific practice.

References

Porter, A. C. (2002, October). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3–14.

Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in Science and Science education*. Washington, D.C.: Council of Chief State Schools Officers.

Webb, N. L. (1999). *Alignment of Science and Science standards and assessments in four states. (Research Monograph 18)*. Madison, WI: National Institute for Science Education and Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED440852).

Webb, N. L. (2005). *Webb alignment tool: Training manual*. Madison, WI: Wisconsin Center for Education Research. Available: http://www.wcer.wisc.edu/WAT/index.aspx