Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 25 Number 4, March 2020 ISSN 1531-7714

Examining the Impact of a Consensus Approach to Content Alignment Studies

Michael Russell, *Boston College*
Sebastian Moncaleano, *Boston College*

Although both content alignment and standard-setting procedures rely on content-expert panel judgements, only the latter employs discussion among panel members. This study employed a modified form of the Webb methodology to examine content alignment for twelve tests administered as part of the Massachusetts Comprehensive Assessment System (MCAS). This modification required panel members to discuss items for which there was no consensus regarding the item's depth of knowledge or targeted standard. After the discussion, panel members were allowed to change their original ratings. The number of changes that occurred were analyzed considering the number of items discussed and the size of the panel. Moreover, we evaluated the impact these changes had on the overall judgments of alignment as reported by Webb's Web Alignment Tool (WAT). Findings suggest that discussion among panel members between rating rounds positively increased agreement among panel members' ratings but had minimal effects on the overall judgments of content alignment for 11 of the 12 tests evaluated.

The validity of inferences based on scores produced by an achievement test is an essential characteristic of any assessment program. For standards-based achievement tests, information about test content is an important source of validity evidence (AERA/APA/NCME, 2014). A common methodology for collecting validity evidence about test content is a content alignment study (Webb, 2006). In a content alignment study, the key question examined is the degree to which the content sampled by a test's items aligns with and represents the content of the domain about which an achievement claim is made. There are several approaches to examining content alignment, each of which rely on a set of experts to make judgments about the standard or learning objective targeted by an item and, in most cases, the depth of knowledge required by a test taker to respond correctly to the item.

The reliance on expert judgment by content alignment study methods is similar to the use of experts during standard setting procedures. Standard setting procedures are employed by criterion-referenced testing programs to establish the cut scores that separate performance categories. Like content alignment studies, standard setting is an important component of achievement testing programs in the United States due, in part, to federal requirements to identify students whose achievement is at an acceptable level. There are several methods for identifying cut-scores that separate contiguous performance levels, most of which include procedures designed to decrease the variability among the judgments made by panel members regarding the location of each cut score. To this end, most standard setting procedures require multiple rounds of judgment. Between each round, panel members are provided an opportunity to discuss their judgments with the aim of increasing commonality in their understanding of the population tested, the items employed by the test, and

what test takers at the border of contiguous performance levels are able to do.

Given the similar reliance on expert judgment for both content alignment and standard setting, it is interesting to observe that content alignment studies typically employ a single round of judgement and, thus, do not provide an opportunity for panel members to discuss and then refine their judgments. The study presented here explored the use of two rounds of judgment during a content alignment study, with discussion about discrepancies in judgments between rounds. The primary research questions focus on the extent to which discussion leads to changes in panel members' judgments, increases agreement among judges, and in turn affects final judgements regarding alignment. Because the study employed panels that remained intact to examine content alignment for three separate grade levels, a secondary question addressed the extent to which agreement among panel members increased as the panel worked on tests for consecutive grade levels.

Background

In this section, we provide a brief description of commonly employed content alignment methods, an overview of standard setting, and a summary of prior research focused on the impact of discussions during standard setting. We also note an important distinction between tasks performed by panel members during a content alignment study compared with standard setting.

Content Alignment Methods

There are several methods to examine the alignment of test content with curricular content. As reported by the NAEP Governing Board (2009), the three most prevalent methods employed to examine the content alignment of achievement tests are Porter's (2006) Survey of Enacted Curriculum (Porter & Smithson, 2002), Achieve, Inc.'s content alignment protocol (Rothman et al., 2002) and Webb's (1997, 1999) 4-component alignment method. More recently, a fourth method was introduced by the National Center for the Improvement of Educational Assessment (NCIEA, 2016) which builds on criteria for alignment established by the Council of Chief State School Officers (CCSSO, 2014) and which was further modified by Achieve (2018). For ease of reference, we refer to this method as the Center for Assessment's method.

All four methods share a similar focus on comparing the content of a test to the content of the standards assessed by the test. In addition, all four methods rely on judgment by experts who are familiar with the test items and the targeted standards.

A distinguishing aspect of Porter's (2006; Porter & Smithson, 2002) method is the focus on the alignment of an achievement test with the curriculum that is actually enacted in the classroom. Porter's method recognizes that a school's curriculum is based on the state standards, but what is emphasized in the curriculum may result in differences between the body of standards to which students are intended to be exposed and the standards to which they are actually exposed. Enacted curriculum is an important consideration when a test is used to inform claims about school or teacher quality and/or impacts of instructional practices. However, documenting enacted curriculum across a state educational system is a challenging and expensive endeavor that may not be practical for state assessment programs that operate in states that provide local control of school curriculum. Moreover, given that states establish standards to define what students are expected to know and be able to do at a given grade level within a given content domain, a focus on enacted curriculum is less aligned with the purpose of state tests than is a focus on the standards themselves.

Achieve Inc.'s method and Webb's method are similar in that they focus on four aspects of alignment between the items comprising an achievement test and the state standards assessed by the test. The aspects examined through each method, however, differ in minor ways. Both methods employ panels of experts to examine the alignment of items with the state standards. In the Achieve method, the focus of analysis is on each item and the standard the item is intended to represent. In the Webb method, the targeted standard is not made known to the panelists and instead requires panelists to identify the standard with which the item aligns (in some cases more than one standard may be identified). In this way, the panelists are not informed as to what an item is intended to assess during their evaluation of the item and its alignment to the standards. Another difference between the two methods is the manner in which results are summarized. The Achieve method yields a narrative-based summary that provides a set of general statements about alignment. In contrast, Webb's method quantifies results and applies pre-specified criteria to evaluate the

strength of alignment as indicated by the resulting quantification of judgments. A final aspect of both methods worth noting is the frequency with which the Webb method has been employed by state testing programs to examine content alignment compared to the infrequency of use of the Achieve method. Perhaps due to the high frequency of use of the Webb's method, digital tools have been developed to support application of the Webb method, whereas no similar tools have been released for the Achieve method (Webb, 2005).

In contrast to each of the above methods, the Center for Assessment's approach evaluates the extent to which test and item content matches criteria for quality assessment content established by the Council of Chief State School Officers (CCSSO, 2014). More specifically, the Center's method focuses on criteria specific to the alignment of English Language Arts (ELA) and mathematics content to their respective standards. Based on the CCSSO criteria, the Center developed rubrics and scoring procedures designed to facilitate the evaluation of the extent to which a given test's content meets the criteria. Similar to other content alignment studies, a group of experts familiar with both the test taker population and the assessed domain apply the rubrics and scoring guides to evaluate each aspect of the criteria. The end product is a table, accompanied by a narrative summary, that indicates the degree to which each criterion is satisfied. Four levels of categorization are employed to reflect satisfaction of a given criterion, namely "weak", "limited/uneven", "good", and "excellent". (NCIEA, 2016).

Standard Setting

Establishing a cut score that separates two performance levels is an important component of employment and certification testing programs. Although the National Assessment of Educational Progress and some state testing programs have long relied on cut scores to categorize student performance, the passage of No Child Left Behind (2002) elevated the importance of standard setting for state achievement tests (Zieky, 2012). Today, every state testing programing employs standard setting procedures to establish at least three and sometimes four separate cut scores that categorize students into one of four or five performance levels.

The earliest method for establishing a cut score was introduced by Nedelsky (1954). This method relied on judgments about the response options that a test taker at the border of two performance levels would reasonably eliminate as incorrect. Based on the remaining "plausible" response options, the probability of guessing correctly was calculated for each item and then summed to yield the cut-score. In effect, the cut-score represented the probability of correctly guessing on each item after the borderline student eliminated response options that were obviously incorrect.

Since Nedelsky introduced this method, several approaches to establishing cut-scores were introduced (Angoff, 1971; Ebel, 1972; Ferrara & Lewis, 2012; Jaeger, 1978; Livingston & Zieky, 1982; Mitzel et al., 2001; Phillips, 2012; Zieky, 2012). There is a not enough room here to describe each method in detail (see Cizek, 2012, for detailed descriptions). Instead we focus on two methods most commonly employed by state achievement testing programs, namely the Angoff and Bookmark methods.

Both the Angoff and Bookmark methods begin by developing descriptions of the knowledge and skills students within each performance level are expected to hold. A panel of experts familiar with both the content domain sampled by the achievement test and the characteristics of the population of test takers is assembled. Training on the performance level descriptions and the standard setting procedures is then provided. For the Angoff method (1971) panel members are asked to keep in mind a test taker that is just barely above the cut-score of interest. Panel members then examine each item and make a judgement about whether or not that test taker would respond correctly or incorrectly to the item. A score of one is awarded for each item judged to be responded to correctly. The sum of item scores is calculated for each panel member and the mean of the panel member scores is said to represent the cut score.

Modifications to the Angoff method were introduced. Perhaps the most common modification shifts the focus of panel members from a single test taker deemed to be just barely above the cut score to a set of such test takers deemed to be just above the cut score. Panel members then estimate the percentage of this set of test takers that would answer a given item correctly (Cizek, 2012). The sum of each percentage is calculated to represent the cut score awarded by each panelist, and the mean of the panelists' cut scores defines the cut score.

Similar to the Angoff method, the Bookmark method (Mitzel et al, 2001) also asks panel members to focus on a test taker that is just barely within a given performance level. But, instead of estimating success (or probability of success) on each item, the Bookmark method orders all items by their observed difficulty and asks the panel members to work up through the items in order of difficulty to identify the item at which the envisioned student would no longer respond correctly. Depending on the implementation, panelists may be asked to identify the item at which the envisioned test taker has a 75%, 67%, or other specified chance of success (Zieky, 2012).

Initially, standard setting methods employed a single round during which panel members provided their judgments. Dependence on a single round of judgments often resulted in judged cut-scores that varied considerably among panel members. To decrease variation among the cut-scores established by each panel member, several recommendations were made during the 1980s to employ multiple rounds of judgment between which feedback to panel members is provided (Berk, 1986; Jaeger, 1989; Livingston & Zieky, 1982).

Feedback generally takes two forms. The first form focuses on variation among panel members. In an effort to decrease variation, after each round of judgment, panel members are shown the distribution of recommended cut scores. Panel members may also be shown variation at the item level. The second form of feedback focuses on the impact that the panel's estimated cut-score has on the classification of test takers. Most often, this feedback shows panel members the percentage of test takers that are placed into each performance level based on the cut-score estimated by the panel. For each form of feedback, panel members are provided an opportunity to discuss differences in judgment and, for impact data, the reasonableness of the resulting classifications. Panel members are then provided an opportunity to revise their judgments. Depending on the implementation, this process is repeated two or three times (Reckase, 2001).

Effect of Feedback on Panel Judgments

There is a small body of research on the effect that feedback and discussion have on the judgments made by panel members. Clauser and his colleagues (2008) conducted a generalizability study that compared the effect discussion, with and without performance data, had on panelists' ratings when using the Angoff method. They found that discussion decreased variation in panelists ratings, but did not impact the correspondence between their item judgments and actual examinee performance. In contrast, provision of impact data increased the correspondence between panel judgments and actual student item-level performance (Clauser et al., 2008). A follow-up study conducted by Clauser et al. (2009) examined the impact of providing student performance data (i.e., distribution of total scores and frequency with which multiple-choice options were chosen by examinees) on panelists' cut score judgments. Results indicated that panel members made substantial changes to their ratings in order to align their judgments to the performance data available, suggesting that panelists deferred to the performance data provided rather than relying on their knowledge and expertise.

Subsequent studies explored further the role of student performance data on Angoff standard setting procedures. Clauser et al. (2013) conducted an experimental study with two conditions: (a) full-data and (b) options-only. Panelists in full-data groups "received two types of data: (1) the proportion of examinees selecting each option and (2) plots showing the proportion of examinees selecting the correct answer by deciles defined by total test score" (p. 65). The options-only group only received the first type of data. Results indicated that judgments provided by panelists in the full-data group were in closer alignment with the empirical data compared to judgments made by panelists in the options-only group. Mee et al. (2013) examined how the accuracy of the performance data provided to panelists impacts their judgments. Inaccurate performance data was created and provided to panel members for a randomly selected sub-set of items. Panelists were warned that some of the data they received was inaccurate. Results showed that panelists did not rely on the performance data available as much as observed in previous studies (e.g., Clauser et al., 2009).

Deunk et al. (2014) also examined the effect of discussion on panel member judgments. Their analysis focused on 15 group discussions that occurred while setting cut scores for four performance levels. Discussions were found to decrease variability among panel members' judgments. Interestingly, they also found no pattern in the direction in which discussions tended to shift panel judgments – in some cases, the panel members' cut scores tended to shift up, and in

other cases they shifted down. Additional analyses found no relationship between the length or the focus of discussions and the impact on the extent to which variability among judgments decreased. A similar study by Margolis and Clauser (2014) systematically examined the impact performance data had on the judgments of 18 independent standard-setting panels for medical licensing examinations. In line with Deunk et al., Margolis and Clauser found that the availability of empirical data reduced the variability of panel members' judgments and prompted significant differences between pre- and post-feedback judgments.

Although not focused specifically on the impact of discussion on panelist judgments, a meta-analysis conducted by Hurtz and Auerbach (2003) of 113 standard setting studies included the use/non-use of discussion as one variable associated with each study. The analysis found that variation in panel judgments was smaller, on average, when discussion occurred. Together, the research on the effect of discussion on panel judgments during standard setting procedures suggests that discussion is effective for decreasing variability among the judgements made by panel members. To date, research has not examined whether discussion during content alignment studies has a similar effect on panel member judgments.

Variability in Categorization Versus Point Estimate

The Angoff and Bookmark standard setting procedures and content alignment procedures require panel members to examine items individually and make judgments based on that examination. In the Angoff and Bookmark methods, the judgement focuses on success or failure by one or more students who are deemed just within a given performance level. For content alignment, the judgment is about the standard assessed by the item and, in most cases, the level of cognition (i.e., depth of knowledge) required to answer the item correctly. In both standard setting and content alignment, panelists work through a test form item by item making these respective judgments. In both procedures, the judgments made by individual panel members are combined to yield an overall panel judgment. In these ways, standard setting and content alignment share similar procedures.

The focus of the judgments and the ultimate goal of the collective panel judgment, however, differ in important ways. As described above, the judgment made during standard setting focuses on the probability that a focal student succeeds on a given item (or succeeds at given level of probability). The ultimate goal of these item level judgments is to yield a point estimate that represents the test score that the focus student would obtain based on the combined judgements of item level success. And the ultimate goal of the collective panel judgment also is to provide a point estimate.

In contrast, the judgments made during content alignment focus on the standard with which the item seems to address and the cognitive challenge presented by the item. The panelists' judgments are combined for two purposes. First, to identify the extent to which the set of items cover (or represent) the set of standards intended to be assessed. Second, to examine the extent to which the cognitive level required to respond correctly to the item aligns with the cognitive level associated with the standard. In these ways, the focus of content alignment is on percent agreement or degree of overlap between the panelists' judgments and information associated with the standards.

This difference between yielding a point estimate and percent coverage/degree of overlap is important to note because it affects how one estimates variability among panel members. For standard setting, variability typically focuses on variation in the point estimate (i.e., cut score) yielded by each panel member. For content alignment, variability focuses on agreement/disagreement among panel members for each individual item. While one might also examine variability of judgments about the focal test-takers performance at the item level, this is not the typical practice. Despite the differences between standard setting and content alignment, given the effects of discussion on standard setting found in the research coupled with some of the similarities in the procedures employed for both standard setting and content alignment studies, it is reasonable to explore whether discussion during content alignment studies has a similar effect on the variability of judgments among panel members.

Methodology

The study presented here examines the effect of discussion on variability among panelists' content alignment judgments and on the final composite judgment regarding alignment. To this end, this study

employed a modified form of the Webb methodology to examine content alignment for twelve tests administered as part of the Massachusetts Comprehensive Assessment System (MCAS). Specifically, the modification required panel members to discuss each item for which less than 70% of panel members agreed regarding the standard and/or depth of knowledge assessed by an item. Following discussion, panel members were provided an opportunity to modify their judgments. As described in greater detail below, analyses focused on the extent to which panel members changed their judgments following discussion, the extent to which these changes affected agreement among panel members, and, finally, the degree to which the collective judgment changed following discussion and the second round of judgments. In the sections below, we describe in greater detail implementation of the Webb method and the analytic methods employed.

Implementation of the Webb Method

The Webb method considers four aspects of alignment, namely categorical concurrence, depth of knowledge consistency, range of knowledge, and balance of representation. Categorical concurrence focuses on the extent to which the categories of content covered by a set of standards corresponds with the categories of content covered by test items. For the analyses presented here, the domains covered by the standards represent the categories of the standards of interest. The primary question addressed through this aspect is the extent to which the items of the test address each domain addressed by the grade level content area standards.

Depth of knowledge consistency focuses on the extent to which the depth of knowledge at which each test item assesses a targeted standard aligns with the depth of knowledge associated with the standard itself. This aspect requires identification of a) the depth of knowledge required to achieve the standard; b) the standard targeted by each item; and c) the depth of knowledge at which the item addresses the targeted standard. For each item, a comparison is made between the depth of knowledge assigned to the item and the depth of knowledge assigned to the standard targeted by the item. Note that for the study presented here, depth of knowledge was defined by Massachusetts' three cognitive levels which include: Level 1 – Identify and/or Recall; Level 2 – Infer/Analyze; and Level 3 – Evaluate/Apply.

Range of knowledge focuses on the extent to which the full set of standards associated with a given domain are represented by the items targeting the given domain. Here the question is not whether the domain is represented, but instead the extent to which all of the standards associated with the domain are represented. As noted above, range of knowledge is influenced by the number of test items and the number of standards. Further, full representation of the standards typically cannot be obtained when the number of standards exceeds the number of operational items comprising the test. The criteria established by Webb classify range of knowledge as adequate when at least half of the standards within a given domain are represented by the items on a test.

Balance of representation focuses on the extent to which the standards addressed by the test items that target a given domain cover the standards in a balanced manner. In other words, given the standards within a domain deemed to be addressed by items, are the standards represented evenly across the items.

For each aspect of alignment, the Webb method calculates a value that indicates the extent to which the aspect of alignment is met. Based on the value, the Webb method then categorizes the extent to which the aspect is met into three levels which are labeled "Yes," "Weak," and "No." "Yes" indicates that the aspect of alignment is fully satisfied and that the resulting test information is sufficient for representing student achievement with respect to the given aspect of alignment. "Weak" indicates representation that is also minimally acceptable for representing student achievement, but could be strengthened. "No" indicates that alignment with respect to the given aspect is not sufficient for adequately representing student achievement. In all cases, the aspects of alignment are examined at the domain level. Thus, the Webb method provides information about the extent to which coverage of each domain is sufficient to represent student achievement within that domain.

In a standard application of the Webb method, panelists review standards and items individually and then code them accordingly. The panelists codes are then examined collectively to make judgements about each of these four aspects of content alignment.

The method employed for this study differed in that after panelists made their initial judgements, the panel leader examined ratings to identify standards and/or

items for which fewer than 70% of the panelists agreed on a rating. Discussion then focused on each standard and/or item for which panel agreement of at least 70% was not reached. Panelists were then given a second opportunity to code the discrepant standard or item. The final ratings were used to examine each aspect of content alignment. Although 100% agreement is clearly desirable, obtaining this level of agreement would likely require substantial time and greatly increase the cost of content alignment. While the 70% minimum level of acceptable agreement is arbitrary, it was chosen as a reasonable threshold that is consistent with thresholds commonly employed in other bodies of literature that rely on panel judgements. In particular, this threshold is consistent with acceptable levels of inter-rater reliability between graders in large-scale assessments, and the chance of responding correctly to a selected response item for a borderline examinee at the cut-off item chosen by a panel member in the Bookmark standard-setting method.

Our implementation of the Webb method entailed the following components:

1. **Panel selection:** Four panels were formed. Two panels focused on ELA and two focused on mathematics. For each content domain, one panel focused on grades 3-5 and the second on grades 6-8. All panel members were teachers who taught the subject area that was the focus of their panel. Members of each panel were selected to represent the geographic/demographic diversity of the state. All panel members had prior knowledge of the state standards associated with their grade level and content area.
2. **Pre-Materials:** All panel members were provided informational materials prior to the panel meeting. These materials described the purpose of the study and introduced key concepts that were covered in greater detail during training.
3. **Whole-Group Training:** All panelists were presented with background information on the purpose of the content alignment study, the definition of alignment employed for this study, definitions of depth of knowledge employed for this study, and the general procedures used to examine and judge alignment. Panelists also engaged in a consensus building activity designed to familiarize panelists with each other and to practice consensus building as a panel.
4. **Panel Training:** Each panel was led by a panel leader who provided additional training that focused on:
    a. Depth of Knowledge as it applied to the content area of focus by the panel
    b. Procedures for coding standards and items for depth of knowledge
    c. Practice coding sample standards and items for depth of knowledge
    d. Issues to consider when identifying the standard(s) addressed by a given item
    e. Practice identifying the standard addressed by sample items
    f. Procedures for discussing discrepancies and for moving towards consensus
    g. Use of the software employed to record depth of knowledge ratings and standard aligned with a given item.
5. **Coding standards for Depth of Knowledge:** Panel members worked individually to examine each standard within a grade level and then assigned a depth of knowledge code to the standard. Panel members focused on only one grade level at a time. After all panel members completed their initial coding, the panel leader examined the level of agreement for each standard. Standards for which fewer than 70% of the panel members assigned the same depth of knowledge were deemed to have not reached consensus agreement. These non-consensus standards were then discussed individually by the panel during which panel members were asked to make a case for each depth of knowledge assigned by one or more members. Additional discussion then occurred as needed before panel members were given an opportunity to recode the standard if desired. After all non-consensus standards were discussed and recoded, the resulting codes were employed to determine if panel consensus was reached and to determine the depth of knowledge of each standard. In cases where the panel consensus was not reached, the depth of knowledge level coded by

the largest number of panel members was assigned to the standard. In cases of a tie, the higher-level depth of knowledge was assigned per Webb's recommendation. Discussions to reach consensus regarding the depth of knowledge of the reviewed standards are part of Webb's method.

6. **Coding Standard Aligned to Item and Depth of Knowledge of Item**: Panel members worked individually to examine each item within a grade level and to identify the standard assessed by the item. In addition, panel members were instructed to only assign more than one standard to an item if they determined that both standards were addressed equally by the item. In this way, the procedures attempted to reduce over-stating representation of standards that might occur if any and all standards that seemed related to the item were identified.

   After panel members assigned one or more standards to an item, they identified the depth of knowledge at which the item assessed the targeted standard(s). Once all panel members completed coding all items within a grade level, the panel leader examined the resulting codes to identify items for which less than 70% of the panel assigned the same standard and/or depth of knowledge. These items were then discussed by the panel. Panel members were given an opportunity to recode the item if desired.

7. **Grade level progression:** Each panel repeated steps 5 and 6 for each grade level to which they were assigned, progressing from the lowest grade level to the highest (e.g., Grade 3, then 4, and finally 5).

8. **Analysis and Summary of Findings:** Once panel sessions were concluded, the tools built into the Webb Alignment Tool (WAT) were used to generate tables that summarize results for each grade level and content area.

Analytic Methods

To examine the extent to which discussions lead to changes in panel member's judgements several metrics were calculated based on the total number of changes in ratings that occurred. Items were discussed for one of three reasons: a) agreement regarding the standard to which the item was aligned was below 70%; b) agreement regarding the depth of knowledge assessed by the item was below 70%; or c) agreement for both category of codes was below 70%. Given that the standard(s) aligned with the item was discussed separately than the depth of knowledge assessed by the item, the number of items discussed for each cause of discrepancy varied. When items were reviewed, panel members were allowed to make changes to their original ratings. On several occasions, panelists also changed one or more of their ratings for items that were not discussed having been informed by discussions for other items.

The first metric calculated corresponds to the mean number of reviewers who made a change per item. This metric was computed as the ratio between the total number of changes observed and number of items for which changes were observed (which often was greater than the number of items discussed). This average has an upper bound equal to the total number of reviewers in the panel. For ease of interpretation of this metric, Tables 1 and 2 in the results section present the observed minimum and maximum number of reviewers who made changes to their judgments.

The second metric reflects the mean number of changes made per reviewer. This metric was computed as the ratio of the total number of changes observed and the number of reviewers in the panel. This metric has an upper bound equal to the total number of items for which changes occurred. To facilitate interpretation, Tables 1 and 2 in the results section also present the observed minimum and maximum number of items for which a reviewer made changes.

The third metric represents the mean percentage of changes made by each reviewer given the number of items discussed. This metric was based on the total number of changes observed, the number of reviewers in a panel and the total number of items discussed. Because this metric places the mean number of changes made by reviewers in relation to the number of items discussed, it provides a standardized metric that can be directly compared across tests and panels. In effect, this metric expresses the proportion of observed changes relative to the number of opportunities for changes (i.e., number of reviewers multiplied by the number of items discussed).

To assess the extent to which discussion increased the agreement between judges, the proportion of items for which consensus (i.e., 70% or more) was

reached was calculated for both the initial and final rounds of panel member judgments.

The impact of the changes made by panel members on the overall judgments of alignment provided by the WAT was evaluated by determining the proportion of judgements that changed. As described above, Webb's alignment method evaluates 4 dimensions of alignment (categorical concurrence, depth of knowledge consistency, range of knowledge, and balance of representation) for each content domain forming the state standards. Therefore, the total number of judgements differed by panel according to the number of content domains represented in the grade level standards targeted by the test. Table 4 presents the number of content domains and the total number of judgements the WAT provides per panel per grade level.

Finally, the extent to which agreement increased through consecutive grade levels within a panel is assessed by comparing several of the metrics described above, particularly the mean percentage of changes made by each reviewer given the number of items discussed and the percent consensus on the initial round of panel ratings.

Results

This study examined the use of discussions during an operational content alignment study of twelve state achievement tests. Of particular interest was the effect discussions had on changes to the codes provided by panel members, agreement among panel members, and ultimately the overall judgments regarding alignment provided by the Webb method. The study focused on four panels each of which performed a content alignment analysis for three tests. The effect of discussions on depth of knowledge ratings and on standard alignment judgments were examined separately.

Table 1 focuses on the depth of knowledge ratings provided for each test. Table 1 shows the grade levels and subject area tests examined, the number of panel members that participated in the content alignment for each test, and the total number of items that formed each test. Also shown in Table 1 is the number of items for which less than 70% of panel members agreed on the depth of knowledge for a given item. It is these items for which discussions occurred. As noted in the methodology section, panel members were not limited to making changes to only the discussed items, but could make changes to the DOK rating for any item based on learning that occurred during discussions. The total number of items that exhibited changes is shown in the fifth column of Table 1.

As shown in Table 1, the number of items discussed ranged from 3 to 15. ELA saw fewer items discussed. This likely occurred because the ELA test contained fewer items. Meanwhile, the number of items for which panelists changed their original DOK rating ranged from 6 to 18. In two-thirds of the panels the number of items that exhibited changes exceeded the number of items discussed.

Across all items on a test for which one or more change occurred, the average number of reviewers who made a change ranged from 1.63 for 5th grade ELA to 3.50 for 8th grade ELA. As indicated by the minimum and maximum changes per item, there were some mathematics items that were discussed but experienced no changes. For most tests, however, a discussion resulted in at least one change per item.

Table 1 also reports the mean number of items that were changed by panel members. These means ranged from a low of only 1.63 item changes per reviewer to 6.13. Most mean item changes per reviewer were between 3 and 4. This table also indicates that for several tests there was at least one panel member that recorded no changes or only one change. For the grade 3 mathematics test, there was one panel member who changed 10 items.

Finally, Table 1 indicates that the mean percentage of items changed per reviewer given items discussed ranged from 23% to 66%. Comparing these mean percentages across tests, the mean percentages are generally higher for ELA than for mathematics. There appears to be no relationship between the order in which a given panel examined the three tests assigned to them and the percentage of changes made.

Table 2 presents the same descriptive statistics for changes in the assignment of items to standard(s). The first four columns present the same information as Table 1. Column five reports the number of items for which panelists changed their original selection of targeted standards by each panel for each test. Note that the number of items that exhibited changes for alignment to standard (ranging from 8 to 23) is noticeably higher than the number of items which were targeted for discussion which ranged from 3 to 15. As

Table 1. Changes to DOK Ratings

	# Reviewers	Total Items	Items Discussed	Items where changes occurred	Reviewers who made a change per item			Changes made by a reviewer			Percent of Changes Given Opportunity
					Mean	Min	Max	Mean	Min	Max	
Math											
3rd	8	40	14	18	2.72	1	4	6.13	1	10	44%
4th	8	40	11	12	2.50	1	4	3.75	2	7	34%
5th	8	40	9	9	2.89	2	4	3.25	2	5	36%
6th	7	34	15	14	1.71	0	3	3.43	2	6	23%
7th	8	34	11	17	2.12	1	4	4.50	3	6	41%
8th	9	34	9	9	3.11	0	5	3.11	0	6	35%
ELA											
3rd	8	25	7	7	2.57	2	4	2.25	0	5	32%
4th	8	25	3	8	1.63	1	3	1.63	1	3	54%
5th	8	25	5	6	3.00	1	5	2.25	0	5	45%
6th	8	25	9	13	2.62	1	5	4.25	1	7	47%
7th	8	25	9	13	2.46	1	4	4.00	3	6	44%
8th	8	25	8	12	3.50	1	7	5.25	1	8	66%

noted above, this difference resulted from panel members being allowed to make changes to any item ratings based on a given discussion. The increase in the number of items that exhibited changes for standards compared to depth of knowledge is not surprising given that there were only three depth of knowledge levels into which items were categorized. In contrast there were 30-50 standards to which an item could be aligned.

Table 2 indicates that, across all items for a given test that were discussed, the mean number of reviewers who changed a rating for an item ranged from 3.00 to 4.53. There were three tests (Math 5, 6, and 7) for which there was at least one item that was discussed but which did not experience any changes. There were also three tests (ELA 4, 7, and 8) for which at least one item that was discussed was changed by all reviewers.

Table 2 also shows that the mean number of changes made by each panel member for a given test ranged from 4.25 to 10.75. All panel members made at least one change and some made between 10 and 14 changes.

Finally, Table 2 indicates that the mean percentage of changes that occurred given an opportunity to make a change following a discussion ranged from 43% to 84%. The mean percentage of changes made by reviewers for items discussed was higher, on average, for ELA than for mathematics. There is not a notable relationship between the order in which tests were reviewed by a panel and the percentage of changes made given the opportunity for change.

Table 3 reports the level of agreement among panel members separately for DOK ratings and standard alignment following the first and second round. In all cases, the percentage of agreement increased noticeably after discussion. Recall that each panel worked first with the lowest grade level to which they were assigned (i.e., third grade or sixth grade) and progressed upwards to the highest grade level. It is interesting to observe that there was not a consistent pattern in how the level of agreement changed during the initial or final round as the panels progressed through the tests to which they were assigned. In some cases, the level of agreement increased as the panel moved up through their assigned grade levels, but in most cases, this did not occur.

Table 2. Changes to Standards Ratings

	# Reviewers	Total Items	Items Discussed	Items where changes occurred	Reviewers who made a change per item			Changes made by a reviewer			Percent of Changes Given Opportunity
					Mean	Min	Max	Mean	Min	Max	
Math											
3rd	8	40	14	15	3.80	1	6	7.13	1	9	51%
4th	8	40	11	8	4.25	1	7	4.25	3	5	61%
5th	8	40	9	10	3.70	0	7	4.63	2	7	46%
6th	7	34	15	16	3.25	0	6	7.43	3	12	44%
7th	8	34	11	18	3.78	0	7	8.50	6	11	57%
8th	9	34	9	14	3.00	1	7	4.67	2	10	58%
ELA											
3rd	8	25	7	12	3.17	1	7	4.75	1	8	43%
4th	8	25	3	23	3.74	1	8	10.75	7	14	67%
5th	8	25	5	17	3.12	1	6	6.63	3	11	44%
6th	8	25	9	21	3.43	1	7	9.00	6	12	64%
7th	8	25	9	17	4.53	1	8	9.63	7	13	69%
8th	8	25	8	16	4.19	1	8	8.38	3	13	84%

Table 3. Changes in consensus proportion per panel per grade level

	DOK ratings		Standard Ratings	
	% Consensus Initial Round	% Consensus Final Round	% Consensus Initial Round	% Consensus Final Round
Math				
3rd	64%	100%	65%	91%
4th	57%	91%	74%	94%
5th	80%	94%	76%	100%
6th	65%	93%	56%	79%
7th	51%	92%	68%	94%
8th	64%	91%	74%	88%
ELA				
3rd	72%	100%	56%	88%
4th	88%	92%	36%	76%
5th	79%	100%	40%	76%
6th	64%	96%	44%	92%
7th	64%	100%	44%	96%
8th	68%	96%	60%	100%

Finally, Table 4 compares the overall judgements regarding alignment for each test based on ratings provided during round 1 and separately during round 2. Recall that for each domain the overall judgement is reported on four areas of alignment (categorical concurrence, depth of knowledge consistency, range of knowledge, and balance of representation). Table 4 shows the number of domains covered by the standards assessed by each test. Multiplying the number of domains by 4 areas of alignment yields the total number of opportunities for an alignment judgement to change. In addition, there are three levels of alignment for each of the four categories of alignment (Yes, Weak, and No). For this reason, a change in alignment may either increase or decrease the strength of alignment. Table 4 indicates that very few of the opportunities for change

Table 4. Changes on Overall Alignment Judgments

Panel	Content Domains	Opportunities	Stronger Alignment	Weaker Alignment	Total Changes	Percent
Math						
3rd	5	20	1	1	2	10%
4th	5	20	1	1	2	10%
5th	6	24	1	0	1	4%
6th	5	20	0	1	1	5%
7th	5	20	0	1	1	5%
8th	5	20	1	0	1	5%
ELA						
3rd	4	16	1	0	1	6%
4th	4	16	1	0	1	6%
5th	4	16	0	0	0	0%
6th	4	16	0	0	0	0%
7th	4	16	1	0	1	6%
8th	4	16	2	2	4	25%

experienced a change. For mathematics, four of the six tests saw only one change, while two tests experienced two changes. In addition, the direction of the changes varied across grade levels such that there was no clear pattern to the direction of the changes. For ELA two tests experienced no change, three tests saw only one change, and one test saw four changes. In most cases the changes strengthened the alignment. With the exception of grade 8 ELA, the percent of opportunities for change that actually experienced a change was 10% or less. Collectively this suggests that the effect of discussions on the overall alignment ratings was relatively small.

It is interesting to note that there was no pattern to the categories of alignment that experience changes following discussions. As Table 5 shows, each category

Table 5. Quantity and Direction of Changes in Alignment Categories

	CC		DOK		ROK		BOR	
Panel	Stronger	Weaker	Stronger	Weaker	Stronger	Weaker	Stronger	Weaker
Math								
3rd	1			1				
4th	1			1				
5th	1							
6th						1		
7th				1				
8th					1			
ELA								
3rd			1					
4th					1			
5th								
6th								
7th			1					
8th				1			2	1

**Note*. CC = Categorical Concurrence, DOK = Depth of knowledge consistency, ROK = Range of Knowledge, and BOR = Balance of Representation.

experienced at least three changes. For three of the four categories, both positive and negative changes occurred. The one exception was categorical concurrence for which the three changes all strengthened alignment. These changes occurred on the grade 3, 4, and 5 math tests. Recall that each test assesses four to six domains. Thus, for each grade level there are between four and six opportunities for change for each category of alignment. Given the small number of changes experienced compared with the opportunities for change, it is difficult to interpret whether this pattern is meaningful.

Discussion

The study found that the use of discussion for items for which less than 70% of panel members ratings agreed did lead to changes in panel members ratings. There was some variability in the degree to which discussions impacted changes across subject areas. In general, discussion led to higher percentages of changes for the ELA tests for both depth of knowledge and standards ratings. It is unclear, however, why this pattern occurred. One possibility is that the sub-domains into which the mathematics standards are clustered more clearly divide content into discrete skills and knowledge. In turn, the discrete nature of mathematics standards may have reduced differences in opinion regarding the alignment between a given test item and the targeted standard.

The study also found no pattern in changes among grade levels or as panel members progressed through the grade level tests to which they were assigned. In addition, while panel members did make several changes to their ratings following discussion, these changes did not have a meaningful impact on the overall judgments regarding alignment. Recall that there are four categories of alignment judged for each domain assessed by a given target. In most cases, a test saw a change for only one category of alignment across all the domains assessed. In only one case (grade 8 ELA) did a substantial percentage (25%) of alignment judgements changed following discussion. For this test, two changes strengthened alignment and two changes weakened alignment, thus the net effect was zero.

The study presented here was conducted in an operational, rather than experimental, context that employed formal training procedures and recruited diverse and representative sets of panel members. The study also focused on alignment for twelve operational state achievement tests examined by four panels. Given the operational nature of the study, the number of panels employed, and the number of tests examined, these findings present a robust opportunity to examine the impact of discussion on content alignment ratings.

Collectively, these findings indicate that discussions did lead to changes in ratings for panel members and had a positive effect on agreement among panel members ratings following these changes. However, with the exception of only one of the 12 tests examined, the impact of discussions on the overall judgments of alignment was minimal. This suggests that while discussions are effective for decreasing variability in panel members ratings, and in some cases lead the panel to substantially change a rating for a given item, discussions did not have a meaningful impact on the overall judgement of alignment. Given the time and resources required to conduct discussions and modify ratings, it is unclear whether the investment in discussion provides a meaningful benefit, beyond increasing agreement in ratings, when employing the Webb content alignment method. Nonetheless, additional research is needed before reaching a definitive conclusion about the value of discussions during content alignment studies.

References

Achieve. (2018). *Independent analysis of the alignment of the ACT to the common core state standards*. https://www.achieve.org/files/ACTReport.pdf

American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). American Council on Education.

Berk, R. (1986). A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research*, *56*, 137-172.

Cizek, G. J. (2012). *Setting Performances Standards: Foundations, Methods, and Innovations* (2nd ed.). Routledge

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2008). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education*, *22*(1), 1-21.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judge's use of examinee performance data in an Angoff standard-setting exercise for medical licensing examination: An experimental study. *Journal of Educational Measurement*, *46*(4), 390-407

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data into Angoff judgments. *International Journal of Testing, 13*, 65-85.

Council of Chief State School Officers. (2014). *Criteria for procuring and evaluating high quality assessments*. Retrieved June 3, 2019, from http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf .

Deunk, M. I., van Kuijk, M. F., & Bosker, R. J. (2014). The effect of small group discussion on cutoff scores during standard setting. *Applied Measurement in Education*, *27*(2), 77-97.

Ebel, R. L. (1972). *Essentials of Educational Measurement* (2nd ed.). Prentice-Hall.

Ferrara, S., & Lewis, D. M. (2012). The Item-Descriptor (ID) Matching Method. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations Methods, and Innovations* (2nd ed., pp. 255-282). Routledge.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, *63*(4), 584-601.

Jaeger, R. M. (1978). *A proposal for setting a standard on the North Carolina high school proficiency test* [Paper presentation]. North Carolina Association for Research in Education Annual Meeting, Chapel Hill, NC, United States.

Jaeger, R. M. (1989). Certification of student competencies. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). Macmillan.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Educational Testing Service.

Margolis, M. J. & Clauser, B. E. (2014). The impact of examinee performance information on judges' cut scores in modified Angoff standard-setting exercises. *Educational Measurement: Issues and Practice*, *33*(1), 15-22.

Mee, J. M., Clauser, B. E., & Margolis, M. J. (2013). The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. *Educational Measurement: Issues and Practice*, *32*(3), 27-35.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 249-281). Erlbaum.

National Assessment of Educational Progress Governing Board. (2009). *Design of Content Alignment Studies in Mathematics and Reading for 12th Grade NAEP and Other Assessments to be Used in Preparedness Research Studies*. NAEP Governing Board.

National Center for the Improvement of Educational Assessment. (2016). *Guide to evaluating assessments using the CCSSO criteria for high quality assessments: Focus on test content*. Retrieved November 3, 2019, from https://www.nciea.org/sites/default/files/publications/CFA-Guide-FocusOnTestContent-R1_0.pdf .

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*, 3-19.

Phillips, G. W. (2012). The Benchmark Method of Standard Setting. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations Methods, and Innovations* (2nd ed., pp. 323-346). Routledge.

Porter, A. (2006, December). Measuring alignment. *National Council on Measurement in Education Newsletter*, *14*(4).

Porter, A. C. & Smithson, J. L. (2002, April). *Alignment of assessments, standards and instruction using curriculum indicator data* [Paper presentation]. American Education Research Association Annual Meeting, New Orleans, LA, United States.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: the role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations Methods, and Innovations* (2nd ed., pp. 159-174). Routledge.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing.* National Center for Research on Evaluation, Standards, and Student Testing.

Webb, N. L. (1997). *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education, Research Monograph No. 6*. Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of Science and Mathematics Standards and Assessments in Four States*. Council of Chief State School Officers.

Webb, N. L. (2005). Web Alignment Tool (WAT) Training Manual. http://wat.wceruw.org/

Webb, N. L. (2006). Identifying Content for Student Achievement Tests. In S. M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp. 155-180). Routledge.

Zieky, M. J. (2012). So much has changed: an historical overview of setting cut scores. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations Methods, and Innovations* (2nd ed., pp. 19-52). Routledge.

Citation:
Russell, M. & Moncaleano, S. (2020). Examining the Impact of a Consensus Approach to Content Alignment Studies. *Practical Assessment, Research & Evaluation*, *25*(4). Available online: https://scholarworks.umass.edu/pare/vol25/iss1/4/

Corresponding Author

Michael Russell
Lynch School of Education and Human Development
Boston College
Chestnut Hill, MA, USA

email: russelmh [at] bc.edu