Anchor Studies for Analysis of NAEP Achievement Levels[1]

Susan Cooper Loomis

The 2016 evaluation of the National Assessment of Educational Progress (NAEP) achievement levels by the National Academies of Sciences, Engineering, and Medicine is the fourth "official" evaluation. The evaluation included several recommendations, and the National Assessment Governing Board (hereafter referred to as Governing Board) has responded with an outline of plans to carry out these recommendations (National Assessment Governing Board, 2016). This paper addresses recommendation 1 from the evaluation report. The evaluation report pointed to studies previously conducted to evaluate the alignment of achievement level descriptions with frameworks, item pools, and cut scores for Grade 12 mathematics and all three grades for reading in 2009, and they recommended "(s)imilar research is needed to evaluate alignment for the Grade 4 and Grade 8 mathematics assessments and to revise them [achievement levels descriptions] as needed to ensure that they represent the knowledge and skills of students at each achievement level. Moreover, additional work to verify alignment for Grade 4 reading and Grade 12 mathematics is needed" (p.245).

The Governing Board's response to this recommendation was positive. The Board noted that a procurement will be developed for conducting studies to evaluate the alignment among the frameworks, item pools, achievement level descriptions, and cut scores for these

[1] This paper was prepared for an Expert Panel Meeting on NAEP Achievement Levels convened by HumRRO on July 12-13, 2018 at their headquarters in Alexandria, Va. The work by HumRRO is under contract ED-NAG-17-C-0002 with the National Assessment Governing Board.

assessments of mathematics and reading (National Assessment Governing Board, 2016). This paper includes a review of studies previously conducted for this purpose. The studies have not evaluated the cut scores, per se, but the results of most of the studies would indicate whether an issue with the location of cut scores is indicated. The studies featured here all use an "anchoring" or "item mapping" methodology. Included in this review are examples of studies that were conducted for the National Center for Education Statistics (NCES). These studies are included to show the variety of study designs and purposes for which anchoring techniques have been used.

After reviewing the previous studies, issues for consideration in designing future studies for this purpose will be addressed. Anchor studies, like standard setting studies, incorporate statistical techniques along with human judgments. The results of anchor studies are ultimately judgments regarding the extent to which there is alignment of achievement levels descriptions with other key factors related to the validity of the achievement levels. The considerations discussed will focus on ways to reduce the potential subjectivity of judgments and produce more compelling evidence through the anchor study designs.

Review of Previous Studies

The studies that exemplify different methods or procedures are described in greater detail. And, the studies for reading and mathematics are described in greater detail because of the focus on these in the evaluation recommendation.

National Center for Education Statistics Studies

Anchor studies, as they are called, have been used in the NAEP program since 1984 when simply reporting the percentage of correct responses to each item was replaced by

reporting scale scores (NCES, 1992). Prior to the development of achievement levels for reporting NAEP results, anchor descriptions were produced to describe performance at various scale score intervals—200, 250, 300, 350, and so forth. The purpose of the anchor descriptions was to give meaning to the score scale by describing how students performed at different points along the scale. A response probability criterion is needed for mapping items to the score scale, and the response probability and other criteria have varied for anchor studies over time. In the early studies to develop anchor descriptions for scale score intervals, items were identified as "anchored" if they anchored or mapped with a response probability of .65. All items within relatively few points of the interval scale score (typically ±12.5 scale score points) were evaluated and considered for anchoring to a specific interval score. The performance required for correctly responding to each set of items anchoring around an interval scale score was described and used to develop a summary description for the interval scale score. A discrimination criterion was also applied to assure that the difficulty of items anchored at one scale score was sufficiently different from the next lower interval scale score. Typically, the criterion was set such that the difficulty, measured as the probability of correct response, at the next lower level was required to be at least 30 percentage points lower.

Achievement levels were first set and reported for the 1992 NAEP in reading and mathematics at Grades 4, 8, and 12.[2] NCES had anchor descriptions developed to include in reporting results for the 1992 *Nation's Report Card* for both reading and mathematics, along

[2] The Governing Board implemented an achievement level-setting procedure for mathematics in 1990, but the results were not reported at that time. After the achievement levels were set in 1992, the data for 1990 were statistically linked and performance relative to the cut scores was reported.

with the data reported for achievement levels. For the reading anchor studies, the achievement level cut scores were used as the scale scores around which items were anchored. Items with a response probability (RP) within 12.5 points of the cut score—above or below—were selected as anchor items for the cut score[3]. A discrimination criterion was not used in that study (Phillips et al., 1993).

For the 1992 NAEP Mathematics Report Card, anchoring was done for the score intervals using RP .65 and a discrimination criterion. Items correctly answered by 65 percent of the students at level 2, for example, and for which at least 50% of the students in level 1 did *not* answer the item correctly and for which the probability of correct response was at least 30% lower than for level 2 met the discrimination criterion. (Phillips et al., 1993, p. 29). Further, for multiple choice items, a "correction for guessing" was used that changed the RP criterion according to the number of choices. For multiple choice items with four options, the RP criterion was .74; for items with five options, the RP criterion was .72. Again, items that satisfied these criteria and were around the interval score point (± 12.5 points) were considered as anchoring at that score level.

Anchor descriptions were also developed for reporting performance in *The Nation's Report Card* for reading (Campbell, Donahue, Reese, & Phillips, 1996), geography (Persky, Reese, O'Sullivan, Lazer, Moore, & Shakrani, 1996), and U.S. history in 1994 (Beatty, Reese, Persky, & Carr, 1996). The anchor descriptions were reported, in addition to reports of

[3] This decision to include items both above and below the cut score resulted in having anchor items from a lower achievement level included in the description of performance at the cut score being anchored. In some cases, the proportion of items from the lower achievement level was substantial.

performance relative to the achievement levels. For reading, geography and U.S. history, items were anchored at the 25th, 50th, and 90th percentile scale score values. Specifically, a range around each percentile value was used such that each anchor range was ±5 percentile points of these values. Anchor descriptions were developed to use for reporting "what students know and can do," in contrast to the achievement levels that were presented as reporting "what students should know and be able to do." NCES continues to develop anchor descriptions for reporting performance, and the information reported in *The Nation's Report Card* for performance around anchor points is approximately as extensive as that reported for performance relative to achievement levels.

Study Conducted by the National Assessment Governing Board

The Governing Board did not accept the recommendations resulting from the achievement levels-setting procedures implemented by ACT, Inc. for the 1996 Science NAEP. The Committee on Achievement Levels used data from the ALS panel meetings to reach agreement on where to set the cut scores, and the Governing Board approved those new cut scores. The Governing Board conducted a panel study to develop descriptions of performance at each achievement level, using a modified anchoring process with a response probability of .65 and the cut scores developed and approved by the Governing Board. Panelists were asked to note the performance requirements of items and write a brief description of the knowledge, skills, and abilities required for a correct response. They wrote descriptions of what students know and can do at each achievement level to use for reporting results relative to the achievement level cut scores. The study was conducted with two small groups of 4-5 panelists for each grade. Two sets of achievement level descriptions were developed for each grade, and

the recommendation was to adopt the descriptions developed by one group. That recommendation was accepted and approved. The Governing Board issued their report on NAEP performance relative to achievement levels for the 1996 Science NAEP (Bourque, 1999).

Studies for the National Assessment Governing Board

1992 Reading NAEP Revisit. This study used item mapping techniques, although it was not an anchor study. It is included here because it uses item mapping techniques and incorporates two methods for comparison of results based on the 1992 reading achievement levels. These studies were to address the question of whether the achievement levels descriptions were valid and useful for describing performance of students on NAEP. There was concern that the achievement level descriptions (ALDs) would not be interpreted the same by different people—a question of reliability. ACT designed a "Reading Revisit" (ACT, 1995) involving two mapping procedures and two panels. All panelists (60% classroom teachers at the grade level and 40% other educators) were required to be familiar with students at the grade level, trained in reading, and recognized for professional activities and involvements in professional organizations. One method was called the Item Difficulty Classification (IDC) method and the other was called the Judgmental Item Classification (JIC) method. The IDC method used statistical mapping techniques to organize items according to the probability of correct response at the cut score (lower boundary) of the achievement level and across the achievement level score interval. Items having at least a .50 probability of correct response at the cut score were categorized as "CAN do" items, items that reached a .50 probability of correct response <u>within</u> the achievement level score range were categorized as "Challenging" items, and items with less than .50 probability of correct response at the upper boundary of the

achievement level were classified as “CAN’T do” items. Panelists were to evaluate items in each category for each achievement level relative to the achievement level description. The IDC procedure was designed to address the question: Can students do what the ALDs say they should be able to do? (ACT, 1995, p. 10) Panelists noted the correspondence, or lack thereof, of items in each category to the ALD for the level. Items in the CAN do list that were not included in the ALD for that level were to be noted; items in the CAN’T do list that appeared in the ALD were to be noted, and so forth. Items in the “Challenging” category were expected to have the highest match to the ALDs since they represent performance at some point within the score interval above the lower boundary.

The second group of panelists were to use the achievement level descriptions to classify items as Basic, Proficient, and Advanced for the JIC study. The question to be addressed by both approaches was whether the ALDs were useful for describing performance of students on NAEP and whether the ALDs represented complete and accurate statements of student performance on NAEP. The JIC study focused on how well the ALDs served to judge/interpret performance of students on NAEP. The IDC method focused more on whether the ALDs provided sufficiently complete and accurate statements of student performance on NAEP to be interpreted accurately and effectively. The results of the two sets of classifications were compared by computing “hit rates” based on cross-tabulations between item classifications for the two panels. Overall, the hit rate was judged to be high and the information collected in the study was judged to provide “compelling evidence that the achievement level descriptions communicate clearly and accurately with respect to student performance” (ACT, 1995, p. 27).

Achievement levels set in 1992 were supported by the findings in the study, although panelists provided recommendations regarding adjustments to the achievement level descriptions. They recommended changes in the ALDs that generally focused on formatting and editing, although several substantive changes were also recommended. A general finding was that the recommended changes to ALDs based on observed performance represented a higher level of difficulty than required by the ALDs. That finding was directly counter to the conclusions of the NAE evaluation panel that had characterized the achievement levels as being set too high (ACT, 1995). As a result of the Reading Revisit, the panelists' recommended changes were reviewed again and modifications to the achievement level descriptions were incorporated for reporting results of the 1994 and subsequent NAEP reading assessments.

2002 NAEP Geography Anchor Study. The geography anchor study was the first of several anchor studies conducted for the Governing Board. ETS was contracted to implement these studies because they had developed the anchoring methodology for NAEP. Starting with this study an IRT model-based computational procedure was used for estimating the conditional probabilities for performance at scores within each achievement level, rather than the "nonparametric" approach traditionally used by ETS for estimating conditional probabilities at specific scale score points (Weiss, 2003, pp. 5-6). All students in the grade-level samples were included, rather than just the students to whom each item was administered. The Governing Board specified that the criteria used by ACT in the achievement levels-setting procedures for selection of exemplar items be used in the anchoring studies: a response probability ≥.50 was used with a discrimination criterion set for each level at the 40th percentile value or higher for

the difference between the conditional probability at the achievement level being anchored and that at the next lower level[4].

This study was conducted over a two-year time period. The first anchor panel was convened in 2000 to evaluate the alignment of the performance of students in Grades 4, 8, and 12 on the 1994 geography NAEP with the ALDs and cut scores developed for the 1994 geography NAEP. The study design was to help establish the extent to which there was evidence that students know and can do the things that the ALDs state they should know and be able to do. Panelists were persons who had worked with development of the geography NAEP framework and item pool, so they were very familiar with the NAEP program and this assessment. The items were anchored using the criteria above (RP.50 and discrimination at 40th percentile) and panelists developed descriptions of the performance on each item in the 1994 geography NAEP item pool at each achievement level in each of the three grades.

The second phase of the study was implemented in 2002 and included the original panel members from 2000 plus another set of panelists to participate in an anchor study of the 2001 Grade 8 geography NAEP. Members of the second panel were selected for their expertise in geography and for their *lack* of experience and knowledge in the geography NAEP. In 2002, both panels were convened to write descriptions of the items from the Grade 8 2001 NAEP that had been anchored for the study. The two panels worked independently. Items were presented in each of the three NAEP achievement levels, plus items that anchored below the

[4] These criteria were recommended by the Technical Advisory Committee for Standard Setting as "minimal" criteria to use in the selection of exemplar items. These criteria were used to identify items for review by ALS panelists in order to select exemplars to recommend for reporting student performance by achievement levels in The Nation's Report Card.

Basic level, did not anchor due to lack of discrimination, and did not anchor due to difficulty (did not reach the RP.50 criterion). Within the achievement levels, items were ordered by difficulty. For the 2000 study, items had been arranged by three content areas within each achievement level and then ordered by difficulty. The content classification was omitted for the 2002 study because panelists in the 2000 study had not found it to be useful (Weiss, 2003).

The study design allowed comparison of descriptions from one assessment year (1994) to another (2001) by the same panelists to assess the extent to which changes in the item pool due to removal and replacement of items impacted the alignment of the assessment to the ALDs. In addition, the design allowed comparisons of descriptions for the 2001 Grade 8 items by two different panels to evaluate the reliability of the anchor descriptions developed by panelists: "Is the process reliable and reproducible, or do the results depend heavily on the unique characteristics of individual panels" (Weiss, 2003, p. 8)?

ETS staff evaluated the comparisons. The alignment of 1994 anchor descriptions to ALDs from the 2000 panel study led to the conclusion that "(t)he alignment between the ALDs and the 1994 summary anchor descriptions (ADs) was remarkably tight" (Weiss, 2003, p. 19). For the 2002 study for which ALDs were compared to ADs for the 2001 assessment for Grade 8, the conclusion was again that there was good alignment. ETS staff further judged the alignment of ADs for the 2001 Grade 8 NAEP by the two panels to be very similar in content, although there were differences, and some differences were noted as "substantive." The ADs developed by the 2002 Panel were considerably longer, more explicit, and more detailed than those developed by the 2000 Panel. This was attributed largely to the fact that the 2002 Panel was not experienced in NAEP and had not been given the framework document or achievement

levels descriptions until after the anchor process had ended. Further, the staff did not provide any clues or advice to the 2002 panelists for writing the ADs. Members of the 2000 Panel, on the other hand, had extensive experience with NAEP and had participated in the 2000 anchor study.

ETS staff noted little or no drift in the alignment for the 1994 and 2001 item pools with the ALDs. Some differences were noted, however, and the subjective nature of the comparisons was acknowledged. No changes were made to the ALDs as a result of this study. The staff recommended using two panels for anchor studies whenever possible.

2003 NAEP Mathematics Anchor Study. An anchor study was implemented by ETS for the Governing Board to evaluate the achievement level descriptions for Grades 4 and 8 in mathematics (Braswell & Haberstroh, 2004). As was the case for the geography study, the Governing Board was interested in examining the extent to which the changes in NAEP item pools over the years since the framework was first implemented in 1990 had impacted the alignment of items with the ALDs developed in 1992. The study was designed to provide direct comparisons of anchor descriptions for the 1992 item pool and the 2003 item pool to determine the impact of minor changes to the framework and the release and replacement of items in the item pool over the decade. A panel for each Grade 4 and 8 was convened to participate in the anchor study for the 1992 item pools and again eight months later to participate in the anchor study for the 2003 item pools. This design was selected to reduce the potential for differences in descriptions due to having different panelists participating in the anchor study for the two different assessment years. ETS staff who developed and oversaw the mathematics NAEP evaluated the results. The panelists' judgments were based on a

comparison among anchor descriptions developed for the 1992 assessment, anchor descriptions developed by the same panelists for the 2003 assessment, and the ALDs originally developed for reporting student performance relative to the NAEP achievement levels originally set for the 1992 Mathematics NAEP.

The findings noted "strong similarities" between the two assessments regarding what fourth grade students could do at each level. For Grade 8, they described the finding as showing "considerable similarity" for what students could do at each level in the two assessment years. They noted that the kinds of geometry problems that 8th grade students could solve at the Basic and Proficient levels were "quite different, but in each case the demand was reasonably consistent within the achievement level" (Braswell & Haberstroh, 2004, p. 21).

They noted that because items are released and replaced for the NAEP item pools for each assessment year, it is reasonable to observe some changes in the content of item descriptions that anchor at each level in the different assessment years. They further noted reasons for which differences between ALDs and anchor descriptions (ADs) should be expected. For Grade 4, they concluded that the alignment between the ALDs and ADs was "very strong." For Grade 8, they described the alignment to be "excellent" between the 1992 ALDs and the ADs for both 1992 and 2003. They noted that for both assessment years, the ADs were very consistent with the policy definitions. (Braswell & Haberstroh, 2004, p. 28)

Studies to Develop ALDs for Reporting

2009 was a busy year with three anchor studies conducted for the Governing Board. Science, reading, and mathematics studies were all conducted in 2009 and all for different reasons than these earlier studies. The 2009 studies were conducted because of a need to

develop or modify ALDs for reporting NAEP performance results for students in one or more grade levels in these subjects.

Science 2009. The Grade 4 science achievement levels caused some concern in that performance relative to the levels was different from that at the other grades and different from that typical of Grade 4 performance. Most noticeably, the Grade 4 Basic cut score waslow, and that seemed inconsistent with the policy definition, as well as other Basic cut scores. The Governing Board asked ETS to implement an anchor study to develop ADs based on performance within the Grade 4 achievement levels to compare to the ALDs for the levels (Pitoniak, Chen, Holler, & Lauko, 2010). The study aimed to determine the extent to which evidence showed that Grade 4 student performing within each achievement level had the knowledge and skills specified in the achievement levels descriptions. In order to also evaluate whether the ALDs were appropriately calibrated, the study was to include comparisons of the ADs to the policy definitions. This study was only one of several used to develop cut scores for science that appeared more reasonable and consistent.

The anchor study was conducted by ETS using .67 as the RP criterion. A modified bookmark procedure (Mapmark with Whole Booklets) was used in the achievement levels-setting process for the 2009 science NAEP, and .67 was the response probability used for mapping items to the score scale (ACT, 2010). No correction for guessing was used for mapping items in the ALS process, and there was no correction for guessing used in the anchor study. Although no discrimination criterion was used in the ALS study, the value at the 40^{th} percentile for the Basic level was used in the anchor study.

The distribution of items that anchored at each level is helpful for understanding the results of the ALS process for Grade 4. Only 14% (28) of the Grade 4 items anchored at the Basic level, while 26% (50) anchored at the Proficient level and 38% (75) at Advanced. Additionally, 22% (43) of the items did not anchor: 6 items anchored below the Basic level; 2 items at the Proficient level and 3 at the Advanced did not anchor because they failed to meet the discrimination criteria; 32 items—16% of the Grade 4 item pool—did not anchor because they were too difficult, i.e. did not meet the RP criteria even at the Advanced level. Given the items that did anchor at Advance and those too difficult to anchor at Advanced, the Grade 4 item pool had 54% of the items with difficulty at or above the Advanced level (Pitoniak, et al., 2010, p. 3). Based on other research to evaluate the science achievement levels at all three grades, the Advanced level cut scores were judged to be too high.

Anchor descriptions were compared to the policy definitions, to the preliminary achievement levels descriptions in the framework, and to the ALDs used in setting cut scores for Grade 4. Panelists were asked to rate the alignment for each as *weak*, *moderate*, or *strong*. Two sets of ratings were collected for each. The first rating was provided independently, and then discussed before collecting the second rating. The alignment of the Basic ADs to the policy definition was judged as strong by all four panelists in the second round of ratings. The advanced ADs were judged to be least well-aligned to the policy definition for Advanced performance. With respect to the ALDs, the ADs were rated as having only weak or moderate alignment at the Basic level: two panelists gave each rating in the second round. For both the Proficient and Advanced levels, three panelists rated the alignment as moderate and one as strong in the second round. The overall results were judged to be “inconclusive.”

Several other studies were conducted to compare the cut scores resulting from the ALS process with external data including state assessment data and international assessment data. Based on the compilation of research, as well as direction from the members of COSDAM, the cut scores for the Basic and Proficient achievement levels for Grade 4 were raised and cut scores for the Advanced level for all three grades were lowered. The cut score recommendations were approved by the Governing Board and used for reporting the results for the 2009 and subsequent science NAEP achievement levels.

The modifications to cut scores required an evaluation of the ALDs relative to the new cut scores. Eight members of the expert content panel that had originally developed the ALDs for the 2009 Science NAEP were again convened to review items that had changed from one level to another and modify the achievement levels descriptions as needed. The content facilitators for the ALS process were among the members of the expert review panel, and they were especially familiar with the descriptions and items. Items were again anchored to achievement levels for evaluation in this process. Only items that changed from one level to another as a result of cut score changes were considered in this study. The ALDs used for the ALS process were modified by the content facilitators on the basis of recommendations by the review panel. The modifications were evaluated and agreed to by the entire expert review panel before they were presented for approval by the Governing Board.

2009 Mathematics NAEP at Grade 12. The anchor study for the 2009 mathematics NAEP was for Grade 12 only. Changes to the framework first implemented in 2005 had added objectives requiring the assessment of mathematics beyond algebra II. The Governing Board had begun research aimed at reporting preparedness of 12th grade students for college and

careers, based on NAEP performance. The change to the mathematics framework was to reflect more fully the requirements of preparedness for post-secondary endeavors—college or careers. The new framework also eliminated some objectives from the 2005 framework. The decision had been made to maintain the scale score trend started in 2005 and to continue reporting performance at achievement levels using the cut scores set for the 2005 NAEP at Grade 12. Thus, modified or completely new achievement levels descriptions were needed for reporting student performance relative to the cut scores. Specifically, the achievement levels descriptions for reporting performance on assessments to be developed for the 2009 framework would need to represent the content of the new framework that best operationalized the policy definitions for each level. The anchor study was to determine the extent to which the ALDs developed for the 2005 framework would need to be modified to represent the 2009 framework and to recommend appropriate modifications to the ALDs. The anchor panel was then to make the modifications and produce a draft set of ALDs to be distributed widely for public comment and review. The recommendations collected from the public comment review would be evaluated and final revisions made for approval by the Governing Board.

ETS was again contracted by the Governing Board to conduct the anchor study (Pitoniak, Dion & Garber, 2010). A panel of six mathematics experts with extensive experience in the development of NAEP mathematics assessments was convened for the study that lasted four days. Cut scores for Grade 12 mathematics had been set for the 2005 assessment using a modified bookmark procedure called Mapmark with Domains (ACT, Inc., 2005). For this study, items were anchored using RP .67 (the same as used in the ALS process) and a discrimination

measure equal to the value of the 40th percentile of difference in the probability of correct response at the given achievement level and the next lower level. No discrimination criteria had been used in the ALS process.

The summary AD was compared to the policy definition for each achievement level, using the same two-round rating plan previously described. The comparison of ADs to policy definitions was to determine if the calibration of performance in each achievement level seemed appropriate. In the second round, two panelists rated the alignment for the Basic AD and policy definition as moderate and four rated it as strong. At the Proficient level, all six rated the alignment as moderate. At the Advanced level, five panelists rated the alignment as moderate and one as weak (Pitoniak, et al., 2010, p. 11).

Next, the ADs were compared to the achievement level definitions, the goal was to ultimately identify the extent to which there was overlap or a lack thereof in the knowledge, skills, and abilities described in the two sets of descriptions. The results for the second round of ratings revealed fewer strong ratings. At the Basic level, three panelists rated the alignment as moderate and three as strong. At both the Proficient and Advanced levels, the ratings were four moderate and two strong (Pitoniak, et al., 2010, p. 12). This would suggest that the alignment was not as good at the Proficient and Advanced levels when ADs were compared to both the policy and achievement level descriptions.

For drafting the new ALDs, panelists identified 11 themes from the ADs to be addressed in the descriptions. Panelists started from the 2005 ALDs and modified the descriptions according to their judgments of the need for change. The achievement level description for the Basic level was not changed. The Proficient and Advanced ALDs both included some

descriptions of knowledge, skills, and abilities judged to be more appropriate at a lower level. The descriptions for these levels were modified.

When asked to evaluate their satisfaction with the products of the panel meeting, the panelists responses indicated a generally high level of satisfaction (Pitoniak, et al., 2010, p. 13).

- Item level ADs 3 very satisfied; 3 satisfied
- Summary ADs 4 very satisfied; 2 satisfied
- Revised ALDs 5 very satisfied; 1 satisfied

The revised ALDs were vetted through the Governing Board's website and by direct communication with stakeholders in the field of mathematics education. Four sets of comments were collected via the website and 62 from the direct request (Pitoniak, et al., 2010, p.14). ETS Staff compiled comments and categorized them by question and response before submitting them to the anchor panel for review and discussion. The panel noted whether action should be taken on the comments and recorded notes about the comments. This discussion took place via webinar. Then a group of the anchor panelists worked on revisions to the ALDs according to the panel's agreed-upon judgments regarding needed changes. The revised ALDS were distributed for review and revised several times before reaching final agreement that the ALDs were ready to submit for the approval of the Governing Board. Their recommendations were approved and used for reporting results for Grade 12 mathematics assessments in the NAEP program.

2009 Preliminary Reading Study at Grades 4 and 8. A new framework for reading was developed and first implemented in the 2009 administration. Implementation of the new framework had been delayed because of interest in maintaining trend during the period of "No

Child Left Behind." Indeed, despite the clear call to establish a new scale for reading with the implementation of the new framework, there was interest in maintaining the current score scale and trend line for reporting results of the new assessment for reading. Given that goal, COSDAM determined that the cut scores for the achievement levels should also be maintained. In order to facilitate reporting results for the 2009 reading NAEP as soon as possible, the Governing Board asked ETS to implement an anchor study to examine the feasibility of using cut scores set for the 1992 reading NAEP for reporting performance for assessments based on the new reading framework. Data for the 2009 operational NAEP were not yet available for use, so the 2008 field trial data collected for Grades 4 and 8 only were used for the study.

For this study, the Governing Board again instructed ETS to use the same criteria that had been used for selecting exemplar items in the 1992 reading ALS process: RP .50 and discrimination at the 40th percentile (as described for previous studies above). Four panelists participated in the anchor study for each grade. ETS assessment development staff for the reading NAEP facilitated the panels (Donahue, Beaulieu, Freund, & Pitoniak, 2009).

The reading NAEP is a cross-grade assessment. For the preliminary Grade 4 anchor study using Grade 4 field trial data, there was a total of 132 items to analyze with 77 administered for Grade 4 only and 55 administered for both Grades 4 and 8. A total of 19.4% of the items did not anchor due to lack of discrimination or failure to reach RP.50 at any achievement level: 13% of the "Did not Anchor" items were at the Grade 4 level only and 18% were administered for Grades 4 and 8 (Donahue, et al., 2009, p. 5). There were 186 items in the field trial for Grade 8, and these included 76 items for Grade 8 only, plus 55 that were administered for Grades 4 and 8 and 55 administered for Grades 8 and 12. For Grade 8, 10.8%

of the items did not anchor: 5.5% administered for Grades 4 and 8; 10.5% administered for Grade 8 only, and 16.4% administered for Grades 8 and 12 (Donahue et al., 2009, p. 6).

The items were arranged in two notebook sets: (1) items organized as administered in reading passages, and (2) items, scoring guides, and anchor data arranged within achievement level by content area (literary or informational) and ordered from least to most difficult.

The process followed the same procedure described for other 2009 anchor studies. Panelists worked independently to develop descriptions of each item anchored at a level. They then discussed the items, grouped those with similar content together, and developed summary anchor descriptions for each level. As for previous studies, panelists arrived at a sense of agreement on whether there was sufficient evidence of a particular type of knowledge, skill, or ability to warrant a summary anchor description (Donahue et al., 2009, p. 11). Once the summary descriptions were all agreed upon, panelists were next to evaluate sequentially these anchor descriptions relative to the policy definitions, the 1992 achievement levels descriptions, the preliminary achievement levels descriptions included in the 2009 reading framework, and the descriptions of performance in each achievement level developed in the 2002 anchor study. As for previous studies, panelists were to rate the alignment of each achievement level comparison as *weak*, *moderate*, or *strong*. The first round of ratings was completed independently before discussing with the group. After the discussion, panelists had the opportunity to change ratings for any or all comparisons

The first comparison of ADs based on the 2008 field trial items was to determine if the performance in the ranges of the cut scores appears to be calibrated according to the policy definitions. In general, the panelists found this to be the case, especially for Grade 4. The

consensus among Grade 4 panelists was that student performance on the two assessments would be classified into the same achievement levels, so the level of performance would be the same although the descriptions of the levels differed somewhat. No one rated the alignment of the ADs to policy definitions as weak for either grade. Only at the Grade 8 Proficient level were there no strong ratings (Donahue et al., 2009, p.13).

The next comparison was of the ADs to the 1992 ALDs: what was the extent of overlap between the two sets of descriptions for each achievement level? Could the 1992 ALDs be modified or would it be necessary to start afresh and write entirely new descriptions? There were known differences in the frameworks, so differences between the two sets of descriptions were expected. The alignment here was rated as lower, especially for Grade 8. For Grade 4, all the ratings were moderate except for one strong rating at both the Proficient and advanced levels. For Grade 8, however, there were weak ratings for all three levels and only the Basic level had two panelists rating the alignment as "strong." Specific aspects of the alignment revealed that some descriptions in the 1992 ALDs were missing from the 2009 ADs and some 2009 ADs were missing from the 1992 ALDs (Donahue, et al., 2009, p. 16).

For the remaining two comparisons, the Grade 8 panelists were unable to complete the tasks within the time available to them. The ratings by Grade 4 panelists for the alignment of 2009 ADs to the preliminary ALDs in the 2009 framework were all strong except for one moderate. For the comparison of ADs from 2009 to ADs from 2002, Grade 4 panelists rated the alignment as mostly strong, with 4 moderate ratings across the three levels (Donahue et al., 2009, pp. 17-18).

Overall, the panelists agreed that while there were differences in the knowledge, skills, and abilities described in the 2009 ADs and the comparison descriptions, the performances for 2009 were appropriately classified by level. Their comments indicated that the 1992 ALDs could be revised and edited for use in reporting performance on the 2009 reading assessment[5].

2009 Reading NAEP at Grades 4, 8, and 12. The linking studies were successful, and the decision was made to report performance of students in the 2009 reading NAEP on the score scale originally established for the reading NAEP in 1992 in order to maintain trend. Results of the linking analyses all supported the technical feasibility of maintaining trend, and the preliminary anchor study for Grades 4 and 8 supported the feasibility of using the cut scores set for the 1992 assessment to revise ALDs for reporting performance for the new 2009 reading NAEP. New achievement levels descriptions would be needed for reporting performance on the new assessment, and an anchor study was implemented using the same anchoring procedures described for the studies for Grade 4 science and Grade 12 mathematics. However, for this study, the RP criterion was .67 and the 40th percentile discrimination criterion was again used for anchoring items.

Sixteen panelists were convened for a four-day study: 5 panelists for each Grade 4 and 8 and 6 for Grade 12 (Donahue, Beaulieu, & Pitoniak, 2010). The panelists were all reading experts, and all but two had extensive experience with the assessment of reading for the NAEP

[5] The original plan was to have Grade 8 panelists complete their ratings at home and submit them for the two comparisons for which they had not had time to complete. The results of the preliminary study were sufficient for determining that ALDs could be developed for reporting results, however. A new study using the operational data for 2009 for all three grades was planned to produce the ALDs to use in reporting the results for 2009 for all three Grades.

program. Six of the panelists in this study had participated in the preliminary feasibility study just described.

ETS NAEP reading program staff had learned in the preliminary anchor study that having a systematic classification of passages would facilitate the judgment task. Prior to the anchor meeting, two content experts developed a complexity rubric to use in coding each of the reading passages. Complexity was judged with respect to the following attributes of the reading passages (Donahue et al., 2010, p. 7):

- Vocabulary
- Sentence structure
- Text structure/author's craft
- Background knowledge
- Cognitive demand
- Density of ideas

Further, four of the persons identified from the anchor panel were sent the rubric and instructions to use for coding the passages as below average difficulty for grade level, average difficulty, or above average difficulty for the grade level. The classifications were summarized by ETS staff. Any passages for which agreement was less than 3 of the four judges were flagged for adjudication, as were any classifications that were non-adjacent. ETS staff convened a webinar with the "complexity" panel to adjudicate the flagged classifications, and final classifications were shared with the panelists for use in the anchor study.

When the full panel of 16 members was convened, they reviewed the items and scoring guides and then discussed the skills demonstrated by students correctly responding to

questions or scoring at each of the different levels of the scoring rubric. They then started writing descriptions of the performance required for correctly responding to each item or achieving each rubric score level. Materials were organized slightly differently for the reading panels because of the need to be able to consider the items with respect to the reading passage. As was the case for the preliminary anchor study, panelists were given two sets of items: (1) items organized by anchor level (as for the other subjects) and ordered by difficulty from lowest to highest and (2) items organized according to the order of administration within each passage set, with the anchor level noted.

The distribution of items across grades was approximately the same in number as for the preliminary study previously described. Because a higher response criterion was used, however, the number of items that were classified into levels shifted such that fewer items anchored at lower levels, and more items did not anchor. For Grade 4, 27.3% of the items did not anchor (p. 5), for Grade 8 a total of 16.4% items did not anchor, and for Grade 12 a total of 17.7% did not anchor. Combining Grade 4 items that anchored at the Advanced achievement level and those that did not anchor because they were too difficult, approximately 43% of the items were at or above the Advanced level. For Grades 8 and 12 the total is around 30% performing at or above the Advanced level for each grade when items anchored at Advanced are added to those that did not anchor due to difficulty (Donahue, et al., 2010, p. 6).

After independently completing anchor descriptions for all of the items at their grade level, panelists worked together to summarize student performance at each subscale in reading for each achievement level. They then evaluated their anchor descriptions relative to the policy definitions for each achievement level, the 1992 achievement level descriptions for each level,

and the preliminary achievement level descriptions included in the framework for the 2009 reading NAEP.

Panelists provided two sets of ratings for each comparison of alignment with the summary ADs, as for the previous studies. For the comparison of ADs to policy definitions across the three grades and three levels 6% were rated as *weak* alignment, 54% as *moderate* alignment, and 40% as *strong* alignment. The only weak alignment ratings were for Grade 8 Basic and Grade 12 Advanced. (Donahue, et al., 2010, p. 15) For the comparison of ADs based on the 2009 assessment to the 1992 achievement level descriptions, the alignment was less strong. Across all grades and levels, 25% of the alignments were rated as *weak*, 73% as *moderate* and 2% as *strong*. Only Grade 12 advanced was rated as having a strong alignment between the 1992 ALDs and the 2009 ADs. All five panelists rated the alignment of the Grade 4 Basic ALD and ADs as weak; 1 rated the Proficient alignment for Grade 4 as weak, and 3 rated the Advanced alignment for Grade 4 as weak. All of the Grade 8 alignments were rated as moderate by all 5 panelists except for 1 weak rating at the Advanced level (Donahue, et al., 2010, p. 16). In comparison to the preliminary achievement level descriptions included in the 2009 framework, the alignment of anchor descriptions was slightly less than for the policy definitions. Across the three grades and levels, 15% of the alignments were rated as weak, 48% as moderate, and 38% strong. The only weak ratings were for Grade 4 Proficient and Advanced. Grade 8 had 14 of the 18 strong ratings: 4 at the Basic level, and 5 for both Proficient and Advanced (Donahue et al., 2010, p. 17)

The panelists then began work on drafting achievement level descriptions for each grade and level. After writing achievement level descriptions for their grade group, they next

worked on the alignment of the descriptions across grades. It became apparent to the panelists that the relationship between the difficulty or complexity of the text and performance required of students seemed to be lost when the descriptions focused only on reading skills. The skills would be the same at each grade, but the difficulty and complexity of reading texts differed significantly. The group discussed how to convey this important relationship, and eventually decided that a preamble would be necessary to convey the inseparability of reading skills from the interaction of text difficulty and the item (Donahue et al., 2010, p. 17). A preamble was drafted for inclusion with the ALDs, and the panel requested that this always be presented with the ALDs.

The achievement levels descriptions were evaluated for similarity in terminology and appropriate progression of skills across levels within grades and across grades within levels. Panelists were asked to evaluate their satisfaction with the products of the study: the item-level anchor descriptions, the summary anchor descriptions, and the draft ALDs for reporting results for reading assessments with the 2009 framework. Grade 8 panelists were very satisfied with all three products. Grade 4 panelists had less positive ratings, and one person was dissatisfied with the draft ALDs. Grade 12 had a split with 1 very satisfied, 1 neutral, and 4 satisfied for both the individual item ADs and the ALDs, and all panelists were satisfied with the summary ADs developed (Donahue, et al., 2010, p. 18).

Public comment was requested via the Governing Board's website and through direct communication with stakeholders in reading education and assessment. Eight comments were submitted via the website and 23 responded to the direct communication for feedback (Donahue et al., 2010, pp. 19-20). The comments were reviewed by the panel to reach

agreement regarding the comments to be addressed in the revisions and modifications. Several iterations of revisions were exchanged between those working on editing drafts and the rest of the panel who reviewed. The final version was presented to the Governing Board and approved.

Issues for Consideration in a Study Design for Analyses of Achievement Levels

The review of previous studies has provided information about several different purposes for which anchor study designs have been implemented. Anchor studies have been used by the Governing Board for evaluating the alignment of achievement level descriptions with performance at each achievement level as a validation study, and the studies have been implemented to accomplish several specific study goals. In general, the studies were to answer questions such as: Is there evidence that the knowledge, skills, and abilities demonstrated by performance of students within achievement level score ranges match that described in the achievement levels descriptions? Do students know and can they do what the achievement levels descriptions say they should know and be able to do?

The achievement level descriptions are developed by extracting key knowledge, skills and abilities from the assessment framework that match the policy definitions: the knowledge skills and abilities needed to demonstrate the level of achievement called for in the policy definitions of achievement levels. The anchor descriptions are a back translation of the ALDs; they describe actual performance within the score range of each level. The anchor studies aim to determine the extent to which performance within the ranges of the achievement levels demonstrate the knowledge, skills and abilities called for in the achievement level descriptions. Indeed, the studies have provided the essential information used in developing new ALDs for

reporting performance relative to existing achievement level cut scores when a new framework and new item pools have been implemented. An anchor study design seems the best choice for accomplishing the research called for in recommendation 1 of the most recent evaluation of NAEP achievement levels.

While the recommendation specifies the need for alignment evaluations for mathematics at Grades 4 and 8 only, the evaluation noted the need for "additional work to verify alignment for Grade 4 reading and Grade 12 mathematics. Implementing anchor studies for all three grades in mathematics and reading seems advisable.

Key factors in the anchor studies will be discussed next, and suggestions for consideration in the design of the anchor studies with respect to these factors will be offered.

Panelists

The criteria for participating in the anchor study process are clearly important. The panelists make judgments throughout the process, and the criteria behind their selection are important for evaluating the results of the process. Key factors that should be considered in the selection and organization of panelists follow.

1. Qualifications of panelists

   Previous studies have included persons with extensive experience with the NAEP subject assessment: members of the framework development panel, members of the NAEP Standing Committee, persons who worked on development of achievement level descriptions, and so forth. One study, however, included persons with no previous NAEP experience to serve on one panel in a two-panel design. The study description makes no comment regarding their performance, but there is discussion of the fact that

the more novice panelists were not given additional training, nor were they given materials to better inform them regarding key aspects of the assessment (Weiss, et al., 2003).

Rather than selecting persons with NAEP content expertise, panelists could be selected to represent the same characteristics as achievement level-setting panelists. Persons with previous NAEP experience in the subject area are not selected for achievement levels-setting panels because that would give them unequal knowledge of the assessment and unequal "status" among other panelists. Panelists for the Reading Revisit in 1992 were selected to represent the qualifications of ALS panelists (ACT, Inc., 1995); otherwise, persons with NAEP content expertise have been selected for this work. Training in the framework and familiarization with the assessment and scoring rubrics would increase the time required for the anchor study if persons with no prior NAEP experience were selected as anchor panelists. Whether general public persons should participate in the study would be an important question.

Overall, it seems that persons with prior knowledge and experience in the assessment are likely to be the best choice to serve on anchor panels. The process is intense and requires communication regarding specific aspects of the subject matter than can best be accomplished by a panel that shares the common content jargon. Persons who served on one of the NAEP teams for framework development, item development, achievement level description development, or as an achievement level setting content facilitator would all be excellent candidates for anchor study panelists. There is high

overlap in the persons who serve in these roles, and most of these persons have a broad array of NAEP expertise.

2. Number of panelists for each grade group

Perhaps a smaller number of panelists is needed when three grade levels are involved in the study than when there is only one grade level. Anchor studies included in this review have ranged from three to six panelists per grade. A larger number of panelists is likely to provide more assurance for the resulting judgments and recommendations, but the task of working across the item-by-item anchor descriptions to develop the summary descriptions increases in time required and complexity of negotiating agreement as the number of panelists producing the descriptions increases. A minimum of three and a maximum of five panelists for each grade group seems appropriate. Each grade and subject panel should have the same number of panelists.

3. Number of panels

A recommendation from the 2002 Geography NAEP anchor study (Weiss et al., 2003) was that two panels be used in the study design whenever possible. Given the judgmental nature of anchor studies and the relatively high stakes associated with these studies to be conducted, involvement of a relatively large number of persons in the process seems advisable. This process will include evaluation of the alignment via the anchor study procedures, as well as reaching agreement on whether and how achievement level descriptions should be modified. Finally, there will be collection and broad-based vetting of the results.

If resources permit having two panels, a replicate panel design is a positive option to provide the opportunity to evaluate the reliability of results. To the extent that two independent groups produce similar anchor summaries and provide similar judgments regarding the alignment comparisons, confidence in study results is strengthened. The study design would be strengthened with a replicate panel design. Adjudicating differences and reaching agreement would be facilitated by having replicate panels, as opposed to two panels representing different sources of content expertise.

Given the importance of this evaluation, it is likely that the study will require a lengthy vetting process. Another panel to evaluate and vet the outcomes of the anchor process might also be beneficial. This panel would be the sort of "jury" to evaluate the results and develop the recommendations based on their interpretation of the results prior to submission of the study findings for public comment. This would have the benefit of removing this burden from staff and involve additional persons in the process. This would potentially increase the procedural validity of the vetting process.

Materials

Materials typically provided for the anchor studies should, of course, be provided.

- Frameworks
- Policy definitions
- Achievement level descriptions
- Preliminary achievement level descriptions from frameworks
- Items and scoring guides with anchor data
- Rating forms

- Evaluations

One suggested change would be to include items from an additional assessment administration in the anchor set. The 2017 NAEP was the most recent administration for Grades 4 and 8 in both reading and mathematics, and 2015 was the most recent for the subjects at Grade 12. Since all three grades were assessed in 2015, that would seem the best choice for the anchor study except for the fact that the 2017 administration was delivered as a digitally based assessment (DBA). It would be advantageous for the anchor studies to evaluate alignment using the new, DBAs for NAEP. Otherwise, the basic question of alignment will be answered as well with the 2015 item pool as the 2017. Given the fact that the most recent assessments of Grade 12 students in both reading and mathematics were in 2015 in a paper and pencil administration of NAEP, the choice of assessment year for the anchor studies is a difficult one. Using the DBA assessments is highly desirable, having the same administration format for all three grade levels in both subjects is highly desirable, and having the study results soon is highly desirable. Clearly, choices will have to be made.

The organization of items used for the most recent anchor studies for the two subjects should be used. The rationale for coding passage complexity for reading should be reviewed and discussed with reading experts, however. If used, this will need to be completed in advance of the panel meeting. And, if the anchor studies use computer-based versions of the assessments, modifications to the organization of items and other features of the study design will be needed.

Anchoring Criteria

The studies reviewed above demonstrate that different criteria have been used for anchor studies. Most of the studies implemented for the Governing Board have featured the criteria used in selecting exemplar items in the achievement levels-setting process. The general rule to use the procedures that match those used in the ALS process should continue. There will be some instances for which this rule cannot or should not be used, however. The most important consideration is standardization of criteria across future studies.

Response Probability Criterion. Several different response probability criteria have been used in anchor studies over the years, ranging from .50 to .80. The goal is to have items mapped to the score scale using a response probability that would lend assurance that the performance is based on actual achievement without representing especially difficult or easy performance. Having this response probability be the same across studies would seem to be the most important consideration. Unless the Governing Board prefers another RP criterion, use of .67 seems reasonable.

Discrimination Criterion. Again, the Governing Board has typically instructed that the discrimination criterion used in the ALS process for selection of exemplar items be used in anchoring items. A discrimination criterion has not been used in the more recent achievement level-setting procedures, so that would create an inconsistency. The rather lenient criterion first suggested by the 1992 ALS TACSS was to calculate the value of the 40^{th} percentile of difference in response probability at a given level and that of the level below. This criterion seems reasonable. If a discrimination criterion is used for future studies, this one seems a good choice.

Correction for Guessing. Although a correction for guessing is used in the NAEP program, the achievement levels setting procedures have not used this convention. There seems no real reason for using a correction for guessing in the anchor studies, so the recommendation would be to continue with the practice of omitting this in the technical criteria for anchoring items.

Items that Do Not Anchor. As noted in the review of items for some studies, the proportion of items that do not anchor can be relatively high, and this means that a significant number of items are omitted from the anchor study. Information regarding the difficulty of the assessment and anchoring of items should be collected early in the planning stages for the studies. A finding that a large proportion of items for one grade and or level do not anchor signals the need for further research. An assessment that contains a high proportion of very difficult items, particularly in one score range, will likely have an impact on the judgments of panelists—both in an achievement levels-setting process and in the anchor study.

Rather than completely omitting them, the items that do not anchor for either low discrimination or low RP should be discussed by the anchor panel. Perhaps more information about the items should be provided to help anchor panelists determine whether the omitted items represent specific knowledge, skill(s), or ability(ies) in the assessment framework. Clearly the omission of a sizable set of items from the anchor study could impact the evaluation of alignment if there is any systematic pattern to the items that are not anchored.

Decision Rules. Anchor study panelists first write descriptions of the knowledge, skills, and abilities required for correctly responding to a question or for scoring at a specific rubric level. After writing these descriptions for all items within a content area or for the entire grade-

level item pool, panelists then summarize the descriptions for individual items to describe performance demonstrated at each achievement level. Study descriptions sometimes referred to the "weight of evidence" as the guide to determining whether to include specific performances in the summary description. A decision rule should be developed in advance to standardize the procedure across the grades, panels, and studies. The summary anchor descriptions are the key to evaluating the alignments in the study, and it is important that their development be standardized to the extent possible.

Since the number of items that anchor at each achievement level varies considerably, perhaps the rule should be an approximate percentage of the items anchored at the level, e.g. 10% of the items anchored at an achievement level. The decision should not be left to the panel members; it must be uniform across studies. Having a standard decision rule will help to assure comparability of the summary anchor descriptions developed across the grade and panel groups for evaluating alignment with NAEP achievement level policy definitions and achievement level descriptions.

Features of the Anchor Study

The anchor studies to be implemented for reading and mathematics in response to recommendation 1 from the evaluation of achievement levels—as well as anchor studies to be implemented in the future for other NAEP subjects—should follow a common set of procedures. Some differences will be necessary to accommodate the organization of the reading NAEP by passages, but the alignments to be evaluated and the procedures for evaluation should be the same. Further, the anchor studies for reading and mathematics should

be especially thorough and rigorous. The study design should be developed and shared for review and comment before it is finalized for implementation.

Timing. The studies should be implemented as soon as possible, but not with any sense of urgency. The agendas for more recent anchor studies have scheduled four days for the process. That is the minimum time that should be scheduled for these studies. Running out of time or being hurried at the end must be avoided.

The studies for the two subject areas can be implemented at the same time, or they can be staggered.

Logistics. Studies for the two content areas can be conducted in the same location or in different locations. There would appear to be no compelling reason to opt for one arrangement over the other, assuming the accommodations are the same. Securing the necessary number of separate meeting rooms for the panels may be a challenge, especially if the design calls for replicate panels for each of the three grades.

Facilitators. Facilitation of the study should be led by persons with extensive experience in the anchor process and thorough knowledge of the NAEP assessment. A lead facilitator should provide the training and instruction to all panel members of the subject study. This person will be responsible for monitoring the process to assure that procedures in each panel workroom are being followed according to the study design and schedule. A separate facilitator will be needed for each panelists' work room to implement the study with each panel group. These facilitators should be experienced in the anchor methodology and the NAEP subject assessment for the grade level. Securing the necessary number of experienced facilitators may be a challenge if the design calls for replicate panels at each grade.

A staff person with strong skills in manipulating data files will be needed to present the anchor descriptions to the group to facilitate the group discussions and development of anchor summaries. This role may be accomplished by the room facilitator or an additional person may be required to perform this role. Observation leads to the conclusion that it is essential to have someone fulfill this role for each panel.

Ratings. In more recent anchor studies, panelists have been asked to rate the alignment of summary anchor descriptions to policy definitions, existing achievement level descriptions, preliminary achievement level descriptions in new assessment framework documents, and so forth. The ratings are simply *weak, moderate, or strong.* These ratings were added to the study designs in 2009, and they clearly add value to the studies by having more objective criteria for evaluating the study results. These ratings may suffice, but perhaps a clearer definition of them is needed. Having a definition of the rating factors provided to panelists would, at a minimum, increase the likelihood that the ratings can be interpreted the same.

An even more detailed version of the rating system would require that panelists specify the statements from the two sets of descriptions that do and do not align. For those that do align, panelists would identify the alignment and rate the level (e.g. weak, moderate, strong) of their alignment. This would clearly take longer, but it would provide much more detail and more clarity regarding the alignment. After completing ratings independently, panelists would be asked to discuss their ratings before providing a final holistic rating of the alignment of the two descriptions for each level.

Panelists should be asked a straightforward question regarding their evaluation of the alignment and their recommendations regarding the need for modifications to the achievement

level descriptions. Panelists should provide an independent response to this question and then discuss their responses in order to reach agreement on a final recommendation.

Evaluations. Panelists have been asked to evaluate their satisfaction with the products resulting from the anchor studies. This is important information to collect. In addition, however, having evaluations collected throughout the study could produce additional information of value to understanding and interpreting the results of the study and of value to future studies. Having panelists' feedback on the timing, organization of materials, instructions and training, and so forth would be of benefit to the study and future studies. Evaluations throughout the study will also provide a basis for comparing across panels.

Other Documentation. Having a more complete record of observations shared in panelists' discussion of the alignment of summary anchor descriptions to the policy definitions, achievement levels descriptions, and any other materials will be helpful. The reports for studies previously conducted do include several examples of comments, and those are very helpful. If recordings are made of the panelists' group discussions, they would need to remain classified as confidential, but the comments could be shared without attribution.

The Vetting Process

The considerations suggested for the study design generally aim to increase standardization across studies and to provide more evidence of procedural validity. The review of study designs over time clearly reveals modifications and changes with this same goal. The vetting process is not a part of the anchor study design, but anchor study contractors and panelists have generally been key participants.

Once the anchor panels have completed their work and made their recommendations, the results of the anchor studies should be shared for external review. This is typically done by the Governing Board, and extensive efforts are made to collect comments from key groups and stakeholders in the content area of the assessment. Having a panel for the vetting process would be helpful. This panel would monitor the vetting process and help to coordinate the finalization of achievement levels descriptions based on the feedback received through collection of public comments. The vetting panel could be a subset of the Anchor Study panel, a new set of panelists, or a mix. Depending upon the specific recommendations provided as feedback through the vetting process, it might be necessary to convene the anchor panel, or a subset thereof, to review the recommendations and implement recommended modifications that have agreement among the vetting panel. This part of the vetting process will require extensive interaction and exchange of ideas. Several weeks should be scheduled to complete the process.

Final Thoughts

The evaluators called for research studies similar to those conducted for the 2009 reading NAEP at all three grades and for the 2009 Grade 12 mathematics NAEP to evaluate the achievement levels for Grades 4 and 8 in mathematics. The suggestion in this paper is to implement an anchor study for all three grades in each subject, mathematics and reading. The study design implemented for the 2009 Grade 12 mathematics NAEP and the 2009 reading NAEP at all three grade levels was apparently satisfactory. The suggestions for specifying more criteria for making and evaluating the anchor study judgments should provide even more compelling evidence.

The evaluators also recommended more research for Grade 4 reading and Grade 12 mathematics. The anchor studies with the modified design features will provide more concrete information to use with additional studies. For example, it will be helpful to know if the more recent grade 4 Reading NAEP contains a similarly high proportion of very difficult items. The impact of this needs further examination. The studies implemented to collect external evidence to use for reaching agreement on cut scores for the Science NAEP in 2009 can serve as a guide to designing additional research to clarify results for grade 4 reading and grade 12 mathematics.

References

ACT, Inc. (1995). *NAEP Reading Revisit: An Evaluation of the 1992 Achievement Levels Descriptions.* Iowa City, IA: The American College Testing Program.

ACT, Inc. (2005). Developing Achievement Levels on the 2005 National Assessment of Educational *Progress in Grade Twelve Mathematics: Process Report.* Iowa City, IA: ACT, Inc.

ACT, Inc. (2010). *Developing achievement levels on the 2009 National Assessment of Educational Progress in Science for Grades Four, Eight, and Twelve: Process Report.* Iowa City, IA: ACT, Inc.

Beatty, A.S., Reese, C.M., Persky, H.R.; Carr, P. (1996). *NAEP 1994 U.S. History Report Card. Findings from the National Assessment of Educational Progress.* Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from https://nces.ed.gov/nationsreportcard/pubs/main1994/96085.aspx.

Bourque, M.L. (1999). *Report on Developing Achievement Level Descriptions for the 1996 NAEP Science Assessment.* Appendix G of *The NAEP 1996 Technical Report*. Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main1996/1999452.pdf.

Campbell, J.R., Donahue, P.L., Reese, C.M., Phillips, G.W. (1996). *NAEP 1994 Reading Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress and Trial State Assessments.* Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from https://eric.ed.gov/?id=ED388962.

Donahue, P., Beaulieu, N., Freund, D., & Pitoniak, M. (2009). *Report on the 2009 Reading Achievement Level and Scale-Anchoring Study (Draft).* Princeton, NJ: Educational Testing Service.

Donahue, P., Beaulieu, N., & Pitoniak, M. (2010). *Final Report on the Study to Draft Achievement-Level Descriptions for Reporting Results of the 2009 National Assessment of Educational Progress in Reading for Grades 4, 8, and 12.* Princeton, NJ: Educational Testing Service.

National Assessment Governing Board. (2016). *Response to the National Academies of Sciences, Engineering, and Medicine 2016 Evaluation of NAEP Achievement Levels*.

National Academies of Sciences, Engineering, and Medicine. (2017). *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress.* Washington, DC: The National Academies Press. doi: 10.17226/23409.

National Center for Education Statistics, *The Nation's report Card for Reading 1992*. Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from https://archive.org/details/ERIC_ED369067.

Persky, H.R., Reese, C.M., O'Sullivan, C.Y., Lazer, S., Moore, J., & Shakrani, S. (1996). *NAEP 1994 Geography Report Card: Findings from the National Assessment of Educational Progress.* Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main1994/96087.pdf.

Phillips, G., Mullis, I.V.S., Bourque, M.L., Williams, P.L., Hambleton, R.K., Owen, E.H., & Barton, P.E. (1993). *Interpreting NAEP Scales.* Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from https://eric.ed.gov/?id=ED36139.

Pitoniak, M., Chen, S-K., Holler, A., & Lauko, M. (2010). *Final Report on the Study to Examine the Grade 4 Achievement Levels for the National Assessment of Educational Progress in Science.* Princeton, NJ: Educational Testing Service.

Pitoniak, M., Dion, G., & Garber, D. (2010). *Final Report on the Study to Draft Achievement-Level Descriptions for Reporting Results of the 2009 National Assessment of Educational Progress in Mathematics for Grade 12*. Princeton, NJ: Educational Testing Service.

Weiss, A. (2003). *Report on 2002 Geography Scale-Anchoring Study* (Draft). Princeton, NJ: Educational Testing Service.