2018 No. 040

Reporting Achievement Level Descriptors for the National Assessment of Educational Progress

Final Report

Prepared for:	National Assessment Governing Board 800 North Capitol Street N.W., Suite 825 Washington DC 20002 Attn: Sharyn Rosenberg, Asst Director for Psychometrics	Prepared under:	National Assessment Governing Board 800 North Capitol Street N.W., Suite 825 Washington, DC 20002 Attn: Munira Mwalimu, Contracting Officer ED-NAGB-17-C-0002
Authors:	Hillary Michaels, HumRRO Karla Egan, EdMetric Art Thacker, HumRRO Sheila Schultz, HumRRO	Date:	July 2, 2018

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314 | Phone: 703.549.3611 | humrro.org

Reporting Achievement Level Descriptors for the National Assessment of Educational Progress

Table of Contents

Table of Contents (Continued)

List of Tables

List of Figures

Reporting Achievement Level Descriptors for the National Assessment of Educational Progress

Section I. Introduction

Achievement level descriptors (ALDs) are a way to communicate expectations for student performance on an assessment. Until recently, the term ALD was loosely defined to encompass a variety of situations ranging from item development to standard setting guidance to stakeholder communication. Most testing programs and all states' programs have developed achievement level (also known as performance level) descriptors as required by the 2015 Every Student Succeeds Act (ESSA) and the previous No Child Left Behind Act (NCLB). Current peer review guidance specifies that states need to "Report on the student's achievement in terms of grade-level achievement using the State's grade-level academic achievement standards and corresponding performance level descriptors" (p.52).

Description of Types of ALDs

A common debate involves whether the purpose of ALDs is to communicate what students *should be able to do* versus what students *can demonstrate* on an assessment (Perie, 2018; Lewis & Green, 1997; Mercado & Egan, 2005). In 2012, Egan, Schneider, and Ferrara introduced an ALD framework in which they delineated four types of ALDs (policy, range, target, and reporting) to differentiate ALDs by their intended purpose. Table 1 summarizes the different types of ALDs along with their intended purpose and primary audience.

Table 1. Types of ALDs with Intended Purpose and Primary Audience

ALD Type	Intended Purpose	Primary Audience
Policy	High-level description of expected performance in each achievement level.	Policy makers
Range	Details the knowledge and skills *expected of and/or demonstrated by* students across the range of achievement within a performance level. These descriptors are typically written for each content strand. Range ALDs can be developed early in the process for item development and then either revised or developed after cut score workshops have been conducted.	Educators, item writers
Target (or Threshold)	Describes the knowledge and skills *expected of* students right at the cut score.	Standard setting panels
Reporting	Summarizes the knowledge and skills *demonstrated by* students. KSAs may be summarized for students right at the cut score, in the middle of the achievement level range, or at the very top of the achievement level range.	Stakeholders

According to Egan, Schneider, and Ferrara (2012), policy ALDs guide development of the other three ALD types. Thus, if range ALDs are used to develop items, they should be written at the

same level at which the items will be written; in other words, if items are written for each content strand, the range ALDs should be developed at the content strand level. Range ALDs are often used as the starting point for developing target ALDs, as target ALDs aggregate the knowledge, skills, and abilities (KSAs) *expected* of students, but they do so to best discriminate performance near the cut scores. The target ALDs become operationalized through the standard setting process. Finally, after cut scores are approved, range ALDs are revised to align to cut scores and reporting ALDs are crafted.

In general, ALDs help stakeholders sensibly interpret test scores. This "linked system of (A)LDs … serve to define the construct that is being measured and describe what students should know and be able to do in relation to the construct. When a clear definition of the target of measurement exists, a more fully aligned assessment system is created" (Egan, Schneider, & Ferrara, 2012, p. 80). As described above, when building a system of ALDs, there should be strong alignment among the policy, range, target, and reporting ALDs.

Purpose of Reporting ALDs

Reporting ALDs are a primary tool of communication with stakeholders. Therefore, they summarize the KSAs students demonstrate on an assessment. Reporting ALDs are developed once cut scores have been established such that the KSAs articulated in reporting ALDs are based on student test performance.

When reporting ALDs are not based on student performance, they remain descriptors of expectations for student performance rather than descriptors of KSAs that students have demonstrated. Reporting ALDs may be found on individual student reports or on an assessment-related website. By their nature, they are not as broad as policy descriptors; they are specific to a grade/content area because they report specific KSAs that students should (or can) demonstrate.

Based on a review of state technical reports and score interpretation manuals, several states such as Utah and Indiana, are developing the four types of ALDs listed.[1] However, based on the states' websites, their current reporting ALDs are short summaries that look similar to the policy descriptors on the NAEP reports.

The purposes of this paper are to:

a. Summarize how reporting ALDs are developed for states and for NAEP
b. Provide a process NAEP could use to develop reporting ALDs.

We first summarize the literature for reporting ALDs and how reporting ALDs are used by states and NAEP. Next, we examine current practices in developing reporting ALDs. Finally, we examine how reporting ALDs may be used in the context of the NAEP assessments.

[1] The Indiana and Utah reports and manuals can be found on their state websites.

Section II. Review of Literature

There is little literature on ALDs themselves, because ALDs are typically discussed as part of the standard setting process. This review focuses on the development of ALDs from assessment items/forms. The interested reader is referred to Egan, Schneider, and Ferrara (2012) for a detailed review of the ALD literature.

Prior to creation of a framework by Egan, Schneider, and Ferrara (2012), much of the discussion regarding ALDs had to do with when ALDs should be written—before or after cut scores are set. In brief, one side argued the KSAs articulated in ALDs should be used to guide standard setting and operationalized through implementation of the cut scores (Perie, 2008). If the ALDs were altered after the cut scores were set, this was seen as moving the bar on what was expected. The other side argued that ALDs should be written after the cut scores were set so they reflected the KSAs that students demonstrated on the test (Green & Lewis, 1997). Those in this camp did not write ALDs prior to standard setting, but rather they used general policy ALDs to guide development of reporting ALDs.

Mercado and Egan (2005) found that ALDs created prior to standard setting do not align with ALDs based on the cut scores. They recommended a middle ground where ALDs describing the expectations for student performance (what students *should* be able to do) guide the work of setting standard. These ALDs are then adjusted once cut scores are set to reflect the work that students *can* do.

In K–12 assessment, state departments of education often do not have a pool of items on which to base ALDs; instead, they often base the ALDs on a single test form. When ALDs are based on a single test form, they may not generalize to student performance on future test forms (Schneider, Egan, Kim, & Brandstrom, 2008). For this reason, Crane and Winter (2006) recommended updating ALDs over time so they continue to reflect student performance. However, this approach may be suboptimal because stakeholders may not understand why the KSAs reported in the ALDs are changing.

Overview of NAEP ALDs

In 1988, P.L. 100-297 established the National Assessment Governing Board (Governing Board), which was charged with developing NAEP achievement goals [Sections (6)(A)(ii) and (6)(E)]. The Governing Board decided the NAEP scale could support three achievement levels: Basic, Proficient, and Advanced. As outlined by Bourque (2009) and the National Academies of Sciences, Engineering, and Medicine (2017) report, the Governing Board developed policy definitions that provided expectations of what students should know and do. The policy definitions were operationalized into ALDs for each grade and content area.

The NAEP program popularized the use of ALDs in standard settings (Hambleton & Pitoniak, 2006). Interestingly, throughout NAEP's history, the timing for development of ALDs has changed. In the early years, NAEP ALDs have been developed before standard setting, during standard setting, after standard setting. There was one occasion in 1996 when the Governing Board adopted cut scores very different from the recommended cut scores resulting in revisions to the ALDs (Bourque, 2009). However, since 1998, the ALDs have been finalized before standard setting, and they have been revised after cut score changes, but typically not after cut score adoption (Bourque, 2009).

The policy definitions that appear in the current Governing Board policy on Developing Student Performance Levels for NAEP have remained basically the same since 1993. However, the

ALDs have evolved since the 1990s. The most recent reading ALDs were developed in 2009 to reflect the revised reading frameworks. The grade 12 mathematics ALDs were developed in 2005 and revised in 2009. The grades 4 and 8 mathematics ALDs have been reviewed over time, but have not been revised since 1993.

Because they help interpret what students *should* know and do, the Governing Board uses ALDs to report NAEP assessment results. The current fourth grade ALDs are presented in Figure 1.

Basic **(214)**	**Fourth-grade students performing at the *Basic* level should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content areas.** Fourth-graders performing at the *Basic* level should be able to estimate and use basic facts to perform simple computations with whole numbers, show some understanding of fractions and decimals, and solve some simple real-world problems in all NAEP content areas. Students at this level should be able to use—though not always accurately—four-function calculators, rulers, and geometric shapes. Their written responses will often be minimal and presented without supporting information.
Proficient **(249)**	**Fourth-grade students performing at the *Proficient* level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas.** Fourth-graders performing at the *Proficient* level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the *Proficient* level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how they were achieved.
Advanced **(282)**	**Fourth-grade students performing at the *Advanced* level should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP content areas.** Fourth-graders performing at the *Advanced* level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. The students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

Downloaded from https://nces.ed.gov/nationsreportcard/mathematics/achieve.aspx#grade4.

Figure 1. Fourth Grade Mathematics ALDs

As presented in NAEP reports, the bold statements in Figure 1 serve as the reporting ALDs. For example, refer to https://www.nationsreportcard.gov/math_2017/#/nation/scores?grade=4. When you click on the question mark by the word Advanced, the Advanced ALD appears.

Based on the four types of ALDs described by Egan et al (2012), reporting ALDs are summary statements of what students know and *can* do; however, the NAEP ALDs provide expectations of what students *should* know and do. As currently written, the NAEP ALDs are high level policy statements of *expected* student performance; they do not summarize what students can do at

the cut point. To be true reporting ALDs, the NAEP ALDs must be consistent with the final cut scores and guide stakeholders to make valid inference about student knowledge based on test scores (Schneider et. al., 2010).

Description of Current NAEP ALDs

Periodically since the early 1990s, the Governing Board has conducted standard setting workshops for the NAEP content area assessments. ALDs have been reported since the 1990s. In the first evaluation of the NAEP standard setting, experts did not believe the current ALDs reflected the performance levels (National Academies of Sciences, Engineering, and Medicine, 2017).

For the 2017 NAEP results, ALDs can be found on the NAEP website[2], and/or by drilling down into some of the reports. The achievement level policy statements are presented on some of the NAEP reports (e.g., state average scores report and item level report, by clicking on the question marks next to the achievement level names). Examining the ALDs included in the NAEP reports, the ALDs focus on general descriptions, such as:

> *Proficient (249)*
>
> Fourth-grade students performing at the *Proficient* level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas. Learn more about the mathematics achievement level descriptions.[3]

Current NAEP ALD Development

The NAEP ALDs are created via a two-phase process. First, preliminary ALDs are created as the assessment frameworks are developed. The preliminary ALDs are intended "to guide item development and initial stages of standard setting" (National Assessment Governing Board, 2012, p. 44). These preliminary ALDs are developed by content area and mostly used to support item development. The Governing Board policy further clarifies that, "(t)he preliminary descriptions are *working descriptions* for the panels while doing the ratings. These may be expanded and revised accordingly as these panels conduct the ratings, examine empirical performance data, and work to develop their final recommendations on the levels" (NAGB, 1995, p. 8). Therefore, the final ALDs are developed prior to the standard setting (Bourque, 2009).

A second phase occurs in which a small group of experts (e.g., 5–10) participate in a workshop to create summary descriptors from the preliminary descriptors. These experts are generally a combination of committee members who developed the assessment frameworks and people who are new to the process. All panelists have expertise in the content and grade level. The NAEP ALDs are vetted through a public review process as well as a review by the Committee on Standards, Design, and Methodology (COSDAM). The expert committee and the Governing Board staff finalize the ALDs based on feedback from the two reviews. It is up to the Governing Board to adopt the ALDs. These ALDs are approved before any standard setting activities.

Figure 2 shows the final ALDs for the NAEP Technology and Engineering Literary (TEL) assessment that were adopted in 2014 by the Governing Board before the achievement level setting in 2015. Once adopted, the final ALDs replace the preliminary ALDs in the content

[2] Obtained from: https://www.nationsreportcard.gov/math_2017/#/nation/scores?grade=4).
[3] Obtained from: https://www.nationsreportcard.gov/math_2017/#/nation/scores?grade=4).

framework documents. Compared to the ALDs in Figure 1, there is more information on what students should be able to do in each of the achievement levels.

Basic:	Eighth grade students performing at the Basic level should be able to use common tools and media to achieve specified goals and identify major impacts. They should demonstrate an understanding that humans can develop solutions by creating and using technologies. They should be able to identify major positive and negative effects that technology can have on the natural and designed world. Students should be able to use systematic engineering design processes to solve a simple problem that responsibly addresses a human need or want. Students should distinguish components in selected technological systems and recognize that technologies require maintenance. They should select common information and communications technology tools and media for specified purposes, tasks, and audiences. Students should be able to find and evaluate sources, organize and display data and other information to address simple research tasks, give appropriate acknowledgement for use of the work of others, and use feedback from team members (assessed virtually).
Proficient:	Eighth grade students performing at the Proficient level should be able to understand the interactions among parts within systems, systematically develop solutions, and contribute to teams (assessed virtually) using common and specialized tools to achieve goals. They should be able to explain how technology and society influence each other by comparing the benefits and limitations of the technologies' impacts. Students should be able to analyze the interactions among components in technological systems and consider how the behavior of a single part affects the whole. They should be able to diagnose the cause of a simple technological problem. They should be able to use a variety of technologies and work with others using systematic engineering design processes in which they iteratively plan, analyze, generate, and communicate solutions. Students should be able to select and use an appropriate range of tools and media for a variety of purposes, tasks, and audiences. They should be able to contribute to work of team collaborators (assessed virtually) and provide constructive feedback. Students should be able to find, evaluate, organize, and display data and information to answer research questions, solve problems, and achieve goals, appropriately citing use of the ideas, words, and images of others.

Advanced:	Eighth grade students performing at the Advanced level should be able to draw upon multiple tools and media to address complex problems and goals and demonstrate their understanding of the potential impacts on society. They should be able to explain the complex relationships between technologies and society and the potential implications of technological decisions on society and the natural world. Given criteria and constraints, students should be able to use systematic engineering design processes to plan, design, and use evidence to evaluate and refine multiple possible solutions to a need or problem and justify their solutions. Students should be able to explain the relationships among components in technological systems, anticipate maintenance issues, identify root causes, and repair faults. They should be able to use a variety of common and specialized information technologies to achieve goals, and to produce and communicate solutions to complex problems. Students should be able to integrate the use of multiple tools and media, evaluate and use data and information, communicate with a range of audiences, and accomplish complex tasks. They should be able to use and explain the ethical and appropriate methods for citing use of multimedia sources and the ideas and work of others. Students should be able to contribute to collaborative tasks on a team (assessed virtually) and organize, monitor, and refine team processes.

Figure 2. Final ALDs for Grade 8 Technology and Engineering Literacy

The current documentation is vague as to when the Governing Board completes the second phase. Guideline 3 of the Governing Board standard setting policy states that, in part, "expanded descriptions of the content expected at each level is based on the preliminary descriptions provided through the national consensus process" (NAGB, 1995, p. 7). Further explanation of this guideline asserts, "(the ALDs) will reference performance within the three regions created by the cut scores" (NAGB, 1995, p. 8). This text implies that the final ALDs are created after the cut scores are set; however, Bourque (2009) states that, since 1998, practice has been to adopt the ALDs before the final cut scores are set.

There are some practices that are specific to the Governing Board. Per Governing Board policy, descriptors are not created for the *Below Basic* category[4]. In addition, ALDs are written in terms of what students *should know and be able to do*, not what they *can do*. In the case of the NAEP TEL assessment, the committee developed the ALDs by achievement level; that is, one group created the *Basic* ALD, another created the *Proficient* ALD, and a third created the *Advanced* ALD. The composition of the groups changed throughout the workshop so that each panelist worked on each ALD before the end of the workshop (WestEd, 2014). In contrast many states have one panel develop all of a grade's ALDs to enhance content articulation. In addition, panelists begin with the *Proficient* category because it is the most important anchor point. As assessment development is based on principled-centered designs, e.g., Smarter Balanced Assessment Consortia tests, ALDs must be consistent with overall program's claims. Additional information is outlined by Plake, Huff, and Resheter (2010).

The Governing Board is in the process of revising its 1995 policy and is considering the development of reporting ALDs following the approval of achievement levels after the achievement level setting meeting using empirical data on student performance. Reporting ALDs would describe what students at each achievement level *do* know and *can* do rather than

[4] States often include positive descriptors of what students in their lowest achievement level can do.

what they *should* know and *should* be able to do. The next section describes considerations for how the Governing Board could approach the development of reporting ALDs for NAEP.

Section III. Best Practices for Developing Reporting ALDs

This section focuses on reporting ALDs that are created after standard setting. It is worth noting that reporting ALDs should be based on the range and target ALDs that precede them. It is expected there is a strong relationship among the four ALD types. This section examines the ways that reporting ALDs are developed. The reporting ALDs explored here are those that would be used by parents, general public, and others who do not seek detailed explanations of student performance.

There are different ways in which reporting ALDs may be developed, all of which depend on student response data being available. There is not an agreed upon set of best practices for developing reporting ALDs. This section describes various timing issues, panelist selection, and methodologies that may be used to develop reporting ALDs.

Timing Issues

By definition, reporting ALDs should be developed after standard setting to reflect the KSAs students *can* demonstrate, as based on actual student performance. Timing issues stem from whether the reporting ALDs should be created immediately after standard setting and be based (perhaps) on a single form, or whether they should be created only after results from several operational forms are available. For NAEP, the item pools are typically deeper than they are for state K–12 assessments; thus, reporting ALDs for NAEP can be created immediately following the adoption of cut scores.

Panelist Selection

Selecting panelists is key to developing achievement level descriptors. The Governing Board must decide how the panelists represent the stakeholder diversity as well as how well panelists need to understand the content standards. For state assessments, panelists are usually educators who have deep understanding of their state's standards and of the expectation of student knowledge. As Loomis (2012) discusses, panelists involved in NAEP standard setting should include educators, policy makers, and the public.

When developing any form of ALDs, the writers must be able to identify the content, skill, and/or process demands required for student success at different intervals along the achievement level scale. In other words, panelists should have deep understanding of the content and learning progression at the grade level or band. This suggests that panelists for ALD workshops include educators and content experts. Policy makers and members of the general public most likely do not have the necessary expertise to write ALDs.

ALD development is a consensus process. The group needs to represent stakeholder diversity while being of a size that is not unwieldy to reach a shared agreement. The ALD panelists usually have access to policy statements, range ALDs, and/or target ALDs to assist their work.

The number of panelists needed to develop or review reporting ALDs is not dependent on the number of cut scores. We suggest between four to six experts per grade/content to revise ALDs. These numbers are consistent with the numbers used for the recently formed assessment consortia.

Methodologies

Several methods have been used to develop reporting ALDs. Egan, Schneider, and Ferrara (2012) used an item mapping technique to create reporting ALDs, while others used portfolios of student work or patterns of student responses on an assessment. This section discusses these three possible methods for developing reporting ALDs: item mapping, construct mapping, and response patterns. The method selected depends on the type of test administered. Item mapping works best for assessments comprised of multiple-choice, constructed-response, and technology-enhanced items. Construct mapping works best for writing or other test types based on one or two items. Response patterns work best for tests that use some sort of branching rule.

Item Mapping

Item mapping is part of a family of standard setting techniques that includes the Bookmark standard setting procedure, MapMark, and ID matching. With item mapping, the items are ordered from easiest to most difficult using empirical data from student performance. Experts analyze the items in terms of what the items measure and how the items increase in difficulty across the scale. When using item mapping techniques to create reporting ALDs, it is important to consider for whom the reporting ALDs will be written. Decisions must be made regarding which (a) item pool to use to create the ALDs and (b) response probability criterion to use to describe the items' locations on the scale, given item ordering differs based on item response theory models and response probability (RP) values (Brevetvas, 2004).

Figure 3 shows an example of an item map. The columns at the far right allow the panelists to tell a story about the types of KSAs that students in each achievement level can demonstrate. Taken together, these items are used to describe performance within the achievement level.

Sample OIB Item Map
Grade 4

Order of Difficulty (OIB Page Number)	Location	Score Point	Test Item #	Item Type	What does this item or score point measure? That is, what do you know about a student who responds successfully to this item or score point?	Why is this item or score point more difficult than the items that precede it?
1	348	1	1	SC		
2	368	1	2	SC		
3	373	1	6	SC		
4	401	1	12	SC		
5	406	1	5	SC		
6	408	1 of 2	14.1	ER		
7	421	2 of 3	14.2	ER		
8	429	3 of 3	14.3	ER		

Figure 3. Sample of Item Map

The framework created by Egan, Schneider, and Ferrara (2012) presents target or threshold ALDs, which are written to describe students who perform right at the cut scores. For reporting ALDs, it is necessary to decide which student is being described. Figure 4 shows an example of a test scale where students are ordered by their scale score and items are ordered by their difficulty value. In Figure 4, the cut score is set at 212 so students with scale scores at or above

212 are considered *Proficient* and students below the cut are *Basic*. The question becomes which items should be used to describe what it means to be *Proficient*?

Ideally, we would have reporting ALDs that are targeted to each scale score to differentiate performance between students who are close together on a scale. This would require extraordinarily deep item pools and an ability to distinguish fine-grained differences between the performances of students who have similar scale scores. Typically, a single point describes student performance on reports; however, it is possible to describe multiple points throughout the range (e.g., Beginning Proficient student, Mid-Proficient Student, etc…). The point, such as just barely over the cut score or in the middle of the achievement level, is based on policy recommendations that best support the assessment purpose and interpretation use.

Entering Proficient Students. If the desire is to report the KSAs held in common by all *Proficient* students, then it is necessary to compile the KSAs of the items in the previous achievement level plus any items right at the cut score. In Figure 4, this means the KSAs found in items 1, 2, and 3 are held in common by all *Proficient* students; therefore, items 1, 2, and 3 would be used to describe *Proficient* performance. In Figure 4, all Proficient students have mastered items 1, 2, and 3. It is worth nothing that the skills of the just-Proficient student and very high-Basic student will have similar descriptions—the distinction between the two are represented by items at the cut score.

Mid-Proficient Student. In some cases, the end-user may wish to describe the KSAs of the student who is at the midpoint of the achievement level. In Figure 4 students at the midpoint would have a scale score of 236. To describe these students, we would compile the KSAs of items at or above the midpoint of the *Basic* category to the midpoint of the *Proficient* category.

Highly-Proficient Student. If the desire is to report the KSAs held by the highly *Proficient* student, then it would be necessary to compile the KSAs of all the items within the *Proficient* category. In Figure 1, this means that the KSAs in items 3 through 9 would be compiled to create the reporting ALD. For a just Advanced student, items at the cut score need to be included. Keep in mind that many of the KSAs mastered by the highly-Proficient student have not been mastered by the entering-Proficient or mid-Proficient student.

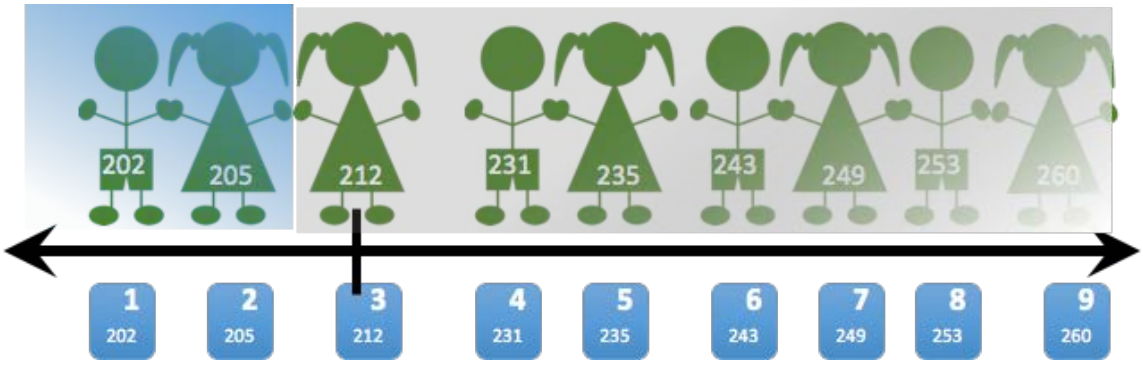

Figure 4. Students and Items Mapped on the Same Test Scale (Egan, 2017).

Content Strands

Item maps are ordered and summarized within their content strand when using the item mapping method to create reporting ALDs. This method provides a well-defined structure in which panelists can organize their work. It also provides a deeper level of information for the stakeholders of the KSAs demonstrated by students in each achievement level.
In examining ordered content strands, items sometimes appear to be ordered inappropriately. Outliers can be noted in all mapping procedures. The item KSA or the KSAs represented by the strand will be examined by content specialists. The specialists will consider how the students interact with the item content, item presentation, and scoring processes.

Item Pool

When using an item-mapping method, the item pool provides the source of KSAs in the reporting ALDs. The item pool may consist of items from the current operational pool or the released item pool. The item pool should be as deep as possible when constructing reporting ALDs. Consider a typical occurrence at an ALD workshop in which panelists find the same KSA in different achievement levels. This KSA is measured by an easy item in the *Basic* level but by a more difficult item in the *Proficient* level. Panelists will use the surrounding items within each achievement level to make sense of the repeated skills. Consider Figure 4 where a single item represents each scale score and imagine that 10 or 20 items are available to measure each scale score. This type of depth allows the panelists to look for common themes across all the items.

The item pools associated with each NAEP assessment is quite large. To develop reporting ALDs, the administration's operational items can be used. These reflect the current constructs and how they are measured. Previously released items may not reflect the current content constructs or item types included in the pool. We recommend that current operational items be used to develop the NAEP ALDs.

New item types that optimally measure the desired content constructs are always being considered. Once these item types are included on operational forms, reporting ALDs should be revised to include the KSAs they add to each of the ALDs. If the new item types do not offer any additional insight into what students know and can do, they do not need to be revised.

Response Probability Criterion

The response probability (RP) criterion indicates mastery because it specifies the probability with which students at a specific ability level will answer an item correctly. The choice of RP criterion will affect how panelists think about mastery of content. Several researchers, including Williams and Schultz (2005) and Zwick, Senturk, Wang, and Loomis (2001) found the cognitive complexity for understanding mastery using item mapping was reduced when using (RPs) between 0.65–0.74.

Different RP values are used for item types because RP74 corrects for guessing inherent in multiple-choice items (NCES, 2008). In state-level assessments, RP values of p=.50 and p=.67 are typically used in item mapping methods. For NAEP achievement level setting, items are ordered at RP67 during achievement level setting activities. However, the adopted RP values

for the exemplars reported on the NAEP website item maps[5] includes *p*=.74 for multiple-choice items and *p*=.65 for constructed-response items (NCES, 2008).

Whatever RP value is used, it provides parameters for how the panelists should think of mastery. For example, when an RP67 is used to order items, this means each item has a scale location where 67 percent of students with that scale score will answer the item correctly. When panelists are describing the items mastered by the students, they know the students at or above the cut score have *at least* a two-thirds chance of answering the item correctly.

The use of different RP values for the same item set can result in different ordering of items, and items may change achievement level. The reader is referred to S.N. Brevetas (2004) for more information.

Creating the Reporting ALD

When using item mapping, the reporting ALD is based on panelists' description of items. These may be compiled in the form of a bulleted list or a short paragraph. The reporting ALDs should be written in a language that is appropriate for the stakeholders, and they should affirmatively describe KSAs of students in each achievement level.

Review of Reporting ALDs

After reporting ALDs are created, independent experts should review them for consistency of language across grades and content areas. In addition, an outside expert should review the ALDs for the progression of concepts across the grades and content areas. Another crosswalk between the ALDs used to set standards and the final ALDs should be completed. The two sets of ALDs should be similar. It is well known that if the ALDs are different, it is a validity issue (refer to Egan & Davidson, 2018). These findings become part of the validity evidence for the achievement levels set and the reporting ALDs.

Construct Mapping

Item mapping methods are appropriate for tests with varied item types while construct mapping is appropriate for assessments with a limited number of performance events (e.g., writing tests).[6] With construct mapping (Wyse, Bunch, Deville, & Viger, 2013), portfolios of student work are ordered along the test scale. Panelists study the portfolios (including rubrics) in terms of student performance and how it improves across the test scale. With this method, one needs to consider the number and range of portfolios that should be used to create reporting ALDs. This is a similar method to the Body of Work Method used to recommend the 2011 NAEP Writing cut scores for grades 8 and 12.

Breadth and Depth of the Portfolio Pool

Like the item pool, it is easy to say that a large number of portfolios should be included in the pool to be studied. Unlike a Body-of-Work standard setting where panelists categorize student work into achievement levels, writers of reporting ALDs are attempting to synthesize elements of student performance across multiple portfolios. This is a more complex task than categorizing

[5] Such as can be found at https://www.nationsreportcard.gov/itemmaps/?subj=MAT&grade=4&year=2017

[6] Please note, discrete items can be included in portfolio methods (Kingston & Tiemann, 2012), but our experience is that assessments composed of mostly discrete items use item mapping methods.

student work. To create a manageable task, the portfolio pool must be limited in its breadth and depth.

With the item mapping techniques, panelists describe all items within a specific range. With construct mapping, panelists describe performance at or near the cut scores. There is no specific guidance on how narrow or broad the pool of portfolios should be: Should the pool include portfolios at the cut score and within one standard error of measurement? Should the pool be limited to portfolios at and within a single scale point of the cut scores? Should only the cut score be considered?

At the same time, one must consider how many portfolios should be considered at each score point. How many portfolios are needed to get a sense of student performance at each scale score point? Again, there is no specific guidance on how many portfolios to use. We suggest conducting a study to determine the optimal number of portfolios and explore the work in the context of the literature on working memory. For a short test, panelists might consider three score points—the cut score and two points above the cut score. Four portfolios would be provided for each score point. This practice assumes the panelists participated in a standard setting workshop where they studied the range of portfolios ahead of time.

Handling Conflicting Rubric Results

As with outliers in item mapping methods, scoring rubrics may seem inconsistent with portfolio sets. In developing the ALDs, content experts must discuss how to make sense of any dissonance. However, the ALDs do not need to refer directly to the rubric. The reporting ALDs need to describe what students (whose work was scored with the rubrics) know and can do.

Student Response Patterns

Reporting ALDs indicate what students know and can do. It may be possible to investigate student response patterns to better understand the learning progression. Consider a simulation where the student must make a series of choices. Each series of choices could result in a different outcome. Student scores would be based on the pattern of responses.

Possible Inclusion of Process Data

As research and analysis of the student process data becomes routine, student process data could be included in the reporting ALDs. Process data is analyzed using all student keystrokes. Feng (2018) has described different patterns of student process data at different NAEP achievement levels. Though this work is still in its infancy, it holds promise for understanding what students know and can do.

Section IV. Developing Reporting ALDs for NAEP

The goal of reporting ALDs is to clearly communicate to stakeholders the KSAs demonstrated by students at each achievement level. When considering the use of reporting ALDs for NAEP, it is important to consider who the end users will be and how the reporting ALDs will be used. Without a clearly stated purpose and a targeted end user, it is probable that new NAEP ALDs will be less effective and less useful than they might be. This is particularly important in NAEP because there are no individual scores. Since reporting ALDs are summary statements, they have different requirements than individual reporting results to be meaningful to the users.

Within the Egan, Schneider, and Ferrara (2012) framework, reporting ALDs are developed after cut scores are set. The form that reporting ALDs will take depends on the intended audience and intended use. Reporting ALDs may take the form of the range ALD that has been updated based on cut scores to a short summary of demonstrated KSAs. The important point for reporting ALDs is that they summarize information based on cut score placement and on skills that can be demonstrated by students.

In this section, we explore how the intended audience and intended purpose intersects with the form that the reporting ALDs will take. To do this, we examine potential stakeholder groups and forms that reporting ALDs have taken in the past.

Stakeholders

In K–12 assessment, reporting ALDs are often written for individual student reports. This type of reporting ALD is primarily intended for parents, students, and teachers. Other stakeholders, such as reporters, may reference reporting ALDs to understand the knowledge and skills demonstrated by students. For NAEP, individual student reports do not exist. Instead, stakeholders are interested in a more global view of the types of KSAs being demonstrated by students at the national or state level. For NAEP, the stakeholder groups are state department of education staff at the national, state, and district levels, politicians, reporters, psychometricians, content experts, and educators.

With this in mind, we must ask how each stakeholder group may use the information released by NAEP through reporting ALDs. Politicians and other policy makers are probably less interested in the demonstrated knowledge and skills than they are in the global look and feel of students at each achievement level. The current NAEP policy descriptors are probably most referenced by this group.

Other groups of stakeholders—Department of Education staff, content experts, educators, and psychometricians—are most likely interested in a finer grain of detail than is currently provided by NAEP ALDs. These stakeholders are likely interested in such things as how students acquire and demonstrate knowledge across the scale, learning progressions as demonstrated through student performance on items, how content and complexity interact with student performance, as well as differences between state and local standards. Range ALDs that have been updated following standard setting are probably the best fit for this group.

Forms of Reporting ALDs

In Table 2 we examine the various ways that entities report descriptors of student performance.

Table 2. Form of ALD for Target Audiences

Target Audience	Target Information	Form of ALD
Policymakers, politicians	Look and feel of student performance	Policy ALDs
Teachers, Content & Curriculum Experts, Psychometricians	Details of student performance across the scale	Range ALDs updated following standard setting
General Public, Reporters	Summary of student performance for an achievement level	Reporting ALD

Policy Descriptors as Reporting ALDs

Table 3 summarizes the mathematics reporting ALDs for Smarter Balanced Assessment Consortium[7]. These descriptors do not discuss knowledge and skills, but rather they are high-level summaries of student performance. These descriptors are different from Smarter Balanced Assessment Consortium's policy descriptors. Even so, the descriptors in Table 3 are firmly in the realm of policy descriptors.

Table 3. Smarter Balanced Mathematics Policy ALDs

High School	Grades 6–8	Grades 3–5
Level 4 The student has exceeded the achievement standard and demonstrates the knowledge and skills in mathematics needed for likely success in entry-level credit-bearing college coursework after high school.	Level 4 The student has exceeded the achievement standard and demonstrates advanced progress toward mastery of the knowledge and skills in mathematics needed for likely success in future coursework.	Level 4 The student has exceeded the achievement standard and demonstrates advanced progress toward mastery of the knowledge and skills in mathematics needed for likely success in future coursework.
Level 3 The student has met the achievement standard and demonstrates progress toward mastery of the knowledge and skills in mathematics needed for likely success in entry-level credit-bearing college coursework after completing high school coursework.	Level 3 The student has met the achievement standard and demonstrates progress toward mastery of the knowledge and skills in mathematics needed for likely success in future coursework.	Level 3 The student has met the achievement standard and demonstrates progress toward mastery of the knowledge and skills in mathematics needed for likely success in future coursework.

[7] https://portal.smarterbalanced.org/library/en/achievement-level-descriptors.pdf

Table 3. (Continued)

High School	Grades 6–8	Grades 3–5
Level 2 The student has nearly met the achievement standard and may require further development to demonstrate the knowledge and skills in mathematics needed for likely success in entry-level credit-bearing college coursework after high school.	Level 2 The student has nearly met the achievement standard and may require further development to demonstrate the knowledge and skills in mathematics needed for likely success in future coursework.	Level 2 The student has nearly met the achievement standard and may require further development to demonstrate the knowledge and skills in mathematics needed for likely success in future coursework.
Level 1 The student has not met the achievement standard and needs substantial improvement to demonstrate the knowledge and skills in mathematics needed for likely success in entry-level credit-bearing college coursework after high school.	Level 1 The student has not met the achievement standard and needs substantial improvement to demonstrate the knowledge and skills in mathematics needed for likely success in future coursework.	Level 1 The student has not met the achievement standard and needs substantial improvement to demonstrate the knowledge and skills in mathematics needed for likely success in future coursework.

Range Descriptors as Reporting ALDs

It is not clear if any states use range ALDs as their reporting ALDs; however, several states have posted range ALDs on their websites (e.g., see the websites for Georgia, Florida, Missouri, and Utah). Range ALDs describe the boundaries among achievement levels. They increase in cognitive complexity and depict the contextual elements students need to demonstrate their knowledge and skills. After the final cut scores are approved, range ALDs should be revised to be consistent with them and the final reporting ALDs.

Reporting ALDs

Most states provide an achievement level description on their reports. This is also true for NAEP reports that present results; however, the ALDs presented are often policy ALDs rather than reporting ALDs. Reporting ALDs are necessarily developed after the cut scores have been adopted. As mentioned in Section 1, target ALDs are the initial expectation of student performance just barely into each of the performance levels while reporting ALDs reflect actual performance and convey useable information to stakeholders about student knowledge and abilities. States and testing programs that include reporting ALDs of this type include Florida, Indiana, and Australia's NAPLAN, as can be found on their websites. An example from Florida is presented in Table 4.

Table 4. Example: Florida Grade 5 ELA Reporting ALDs for Levels 1, 3, and 5

Level 1	Level 3	Level 5
Performance at this level indicates an inadequate level of success with the challenging content of the Next Generation Sunshine State Standards for reading	Presented with grade-appropriate texts encompassing a range of complexity, students will **generally** be able to • use context clues to determine the meaning of an unfamiliar word; • determine the meanings of complex words by using the meaning of familiar base words and affixes; • determine the meanings of complex words by using Greek or Latin roots; • use knowledge of antonyms or synonyms to determine meanings of words; • analyze the context surrounding a word with multiple meanings to determine the correct meaning of the word; and • analyze the word or phrase to determine small or subtle differences in meanings between related words.	Presented with grade-appropriate texts encompassing a range of complexity, students will **consistently** be able to • use context clues to determine the meaning of an unfamiliar word; • determine the meanings of complex words by using the meaning of familiar base words and affixes; • determine the meanings of complex words by using Greek or Latin roots; • use knowledge of antonyms or synonyms to infer the meanings of words by using simple analysis; • analyze the context surrounding a word with multiple meanings to determine the correct meaning of the word; and • analyze the word or phrase to determine small or subtle differences in meanings between related words.

Considerations in Developing Reporting ALDs

Current NAEP ALDs are written as “should” statements. If they were written as “can” statements, the ALDs would provide specific information about KSAs students have at each performance level. We hypothesize that revised “can” ALDs that focus on student skills and knowledge would be more useful to most stakeholders.

Clear definitions of the purpose and expected use of ALDs should be explicated as the Governing Board considers developing reporting ALDs for NAEP. Because NAEP is not reported at the individual level, we suggest developing samples for different stakeholders and investigate whether they were clearly and correctly interpreted and if they have utility to specific stakeholder groups.

Reporting ALD development would rely on the cut scores being finalized (instead of tentative) and item ordering. As mentioned earlier, the items are typically ordered based on response probabilities. People have trouble understanding the response probability. Though Williams and

Schultz (2005) have suggest that panelists are comfortable working with items ordered at RP67, others have found that the response probability ordering should be a policy decision as panelists do not appreciate the impact of changing the RP values on their work (Lewis, Mitzel, Mercado & Schulz, 2012).

Review NAEP's Current ALDs to New Reporting ALDs

An example of the current mathematics grade 4 NAEP ALDs is presented in Figure 1. The National Academies of Sciences, Engineering, and Medicine report (2017) suggest no revised standards be set at this time, but they do suggest that ALDs be revised. Revised ALDs should align with content frameworks, items, and current cut scores. As discussed by Egan and Davidson (2018), when ALDs are used on score reports the statements need to distillations of what students know and can do. The reporting ALDs need to be general, but include actionable information.

Reporting ALDs should have consistent language among the grades and content areas. Articulation of the ALDs is usually completed by the content experts who review, refine, and revise the final ALDs from standard setting. Progressions of key concepts and skills are included in the ALDs. This helps make the testing system more coherent and aids in score interpretation.

Practical Challenges to Reporting NAEP ALDs

As mentioned previously, some of the challenges to creating reporting ALDs include NAEP's cut scores being defined as trial status and NAEP results not being reported at the student level. There are some other challenges unique to NAEP. NAEP's content frameworks are specific to the NAEP assessments. States have adopted their own content standards. These differences may result in students from one state scoring poorly on content not covered or operationalized in the same way by both assessments, limiting the interpretability of score differences between states or between states and NAEP. However, if the NAEP ALDs were developed with the "can" statements, state personnel could more easily determine whether there are substantive differences in content between NAEP and state assessments and make better sense of comparison data.

Implementation Procedures

The goal of reporting ALDs is to summarize the knowledge and skills *demonstrated by* students at the cut score. As noted earlier, reporting ALDs include specific information about the KSAs and they assist NAEP stakeholders in understanding what students know and can do. In addition, the Governing Board needs to articulate the goal of the reporting ALDs—why they are needed and who they are for. Until the goal is articulated, it will be difficult to design an optimal process for creating ALDs.

Disseminating ALDs

The primary audience for NAEP is the American public including everyone from policymakers and researchers to parents and media (NAGB, 2017). Results need to be concise and understandable. In their evaluation, Zenisky, Hambleton, and Sireci (2009) realized that stakeholders found the current ALDs confusing. They suggest working with state NAEP coordinators to make the reporting ALDs more understandable.

NAEP reports should continue to present the ALDs. However, the Governing must recognize these are not true summary reporting ALDs. The target and range ALDs provide more breadth. It is easier for stakeholders to see skill progression in the range ALDs. Thus, we recommend the range and target ALDs also be posted along with the reporting ALDs. We suggest the Governing Board conduct interviews and focus groups with various stakeholder types to assess the value of posting these ALDs. Usability investigations from stakeholder focus groups would provide insight to their usefulness and usability.

Collecting Validity Evidence

As outlined by Kane (2001) and Hambleton et. al. (2012), procedural evidence should be collected and reported to evaluate the standard setting and achievement level descriptor process. The procedures should indicate the process was conducted by stakeholders and content experts who understood and followed the outlined procedures. By documenting what should have occurred as well as what did happen and the results, the process becomes replicable and can be evaluated. Because a standard setting panel is unique in terms of participants, time, and experience, it is unlikely that another panel would recommend the exact same cut scores.

The ALDs are part of a larger system: content frameworks, items, rubrics, and reporting and policy ALDs. The alignment of these indicators provides evidence of validity. It is important to review the alignment of a system periodically (i.e., every 5–10 years), as well as when any significant change occurs to any part of the system.

ALD development is a consensus process. Facilitators should keep notes of content strands or items where panelists had an easier or more difficult time achieving agreement. Panelists should complete evaluations to determine if they felt everyone was heard and if they support the consensus reporting ALDs.

Obtaining external validity evidence is always challenging. Alignment to state content can be one measure. Another is an audit of the ALDs to ensure there is consistency between the ALDs developed for achievement level setting and reporting ALDs and the other parts of the NAEP assessment system, from item development through reporting. In addition, an evaluative audit of the reporting ALDs can be conducted for alignment with NAEPs reporting claims and goals. If the goal of reporting is to support district and/or state instructional strategies, stakeholders can be surveyed to determine how they are using the reporting ALDs.

Reporting ALDs need to reflect the KSAs of the reported cut scores. As changes in the cut scores or addition of item types are made, the reporting ALDs may need to be revisited.

Conclusion

This paper presented information on reporting ALDs, including best practices for creating reporting ALDs as well as information the Governing Board should consider as it enhances the current NAEP ALDs. However, there are several policy decisions that must be made before a plan for reporting ALD development can be proposed.

Decisions Points

As demonstrated in the paper, there are multiple techniques for creating reporting ALDs. All must be guided by inputs from the Governing Board. These inputs include the:

- type of student for whom the ALDs will be written,
- way in which the ALDs will be used,
- stakeholder for whom the ALDs are created,
- way in which ALDs will be written (should vs can), and
- inclusion of process information.

Student

In the paper, we discussed how the reporting ALDs may be written for different areas of an achievement level. It may be the most productive for the Governing Board if it develops short reporting ALDs for multiple areas within an achievement level. For example, the Governing Board could summarize the KSAs of the students just entering Proficient, midway through Proficient, and at the high-end of Proficient.

Use and Stakeholder

Perhaps the most important decision point is the intended use and intended audience for the reporting PLDs. It is likely that different audiences need information from the ALDs, and it would behoove the Governing Board to create different types of ALDs targeted to an intended audience. Some stakeholders, such as educators, may benefit from the range ALDs that describe in detail what students can do and provide specific examples than does a typical reporting ALD. There could be a way to combine the range ALDs and the posted item map information once they are aligned.

Should vs Can

The Governing Board must also decide if the reporting ALDs will represent the KSAs demonstrated by students (can statements) or the KSAs that are aspirational (should statements).

Process Data

The Governing Board must further decide if the KSAs included in the ALDs will include process data. Though process data are not currently used in NAEP scoring and reporting, they provide useful information about what students do.

Stakeholder Meetings

To make decisions on each of the point, we suggest the Governing Board conduct cognitive laboratories and focus groups with stakeholders to understand what is desired by each group. There is not information regarding what stakeholders want out of a system of ALDs; thus, the various stakeholders should be asked what they want and obtain their reactions to different types of ALDs.

The Governing Board is in the preliminary steps of enhancing the ALDs currently being used. Once the decision points are settled, the logistics of creating reporting ALDs will be fairly straightforward.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bourque, M.L. (2008). A history of NAEP achievement levels: Issues, implementation, and impact 1989-2009. Paper commissioned for the 20th anniversary of the National Assessment Governing Board. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/who-we-are/20-anniversary/bourque-achievement-levels-formatted.pdf

Beretvas, S. N. (2004). Comparison of bookmark difficulty locations under different response models. *Applied Psychological* Measurement, 28(1), pp. 25-47.

Egan, K. & Davidson, A. (2018). Toward coherence in assessment systems through achievement level descriptors using the NAEP example. Alexandria, VA: HumRRO.

Feng, G. (2018). *Analyzing writing process in NAEP writing assessment: Implementation and evidence extraction*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, New York.

Hambleton, R.K. and Pitoniak, M.J. (2006). Setting Performance Standards. In R.L. Brennan (Ed.), *Educational measurement*, 4th ed. Westport, CT: Praeger Publishers.

Hambleton, R., Pitoniak, M., & Copella, J. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G.J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (pp. 47-76). New York: Routledge.

Kane, M. (2001). So much remains the same: conception and status of validation in setting standards. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp.53-88). Mahwah, NJ: Lawrence Erlbaum.

Lewis, D. M., & Green, R. (1997, June). The validity of PLDs. Paper presented at the National Conference on Large Scale Assessment, Colorado Springs, CO.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, M. (2012). The Bookmark standard setting procedure. In G. Cizek *Setting Performance Standards* (pp. 225-282). New York, NY: Routledge.

Loomis, S. (2012). Selecting and training standard setting participants. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp.107-134). Mahwah, NJ: Lawrence Erlbaum.

Mercado, R. L., & Egan, K. L. (2005). Performance level descriptors. Paper presented at the National Council on Measurement in Education, Montreal, Quebec, Canada.

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. https://doi.org/10.17226/23409.

National Assessment Governing Board. (March 4, 2017). *Policy statement on reporting, release, and dissemination of NAEP results*.

National Assessment Governing Board. (2012). Reading framework for the 2013 National Assessment of Educational Progress. Washington, DC: Author.

National Center for Education Statistics. (2008). Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/describing_itemmapping.asp

Perie, M. (2008). A guide to understanding and developing PLDs. Educational Measurement: Issues and Practice, 27(4), 15–29.

Plake, B.S., Huff, K., & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement levels descriptor development and for standard setting. Applied Measurement in Education, 23, 307-309.

Public Law 100-297 (1988). National assessment of educational progress improvement act. (Article No. USC1221). Washington, DC.

Schneider, M. C., Huff, K. L., Egan, K. L., Tully, M., & Ferrara, S. (2010). Aligning achievement level descriptors to mapped item demands to enhance valid interpretations of scale scores and inform item development. In *annual meeting of the American Educational Research Association, Denver, CO*.

Schneider, M., Kitmitto, S., Muhusani, H., & Zhu, B. (2015). Using the National Assessment of Educational Progress as an Indicator for College and Career Preparedness. Washington, DC: Author. Retrieved from http://www.air.org/sites/default/files/downloads/report/Using-NAEP-as-an-Indicator-College-Career-Preparedness-Oct-2015.pdf

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). Setting performance standards for student achievement: A report of the National Academy of Education Panel on the Evaluation of the 1992 Achievement Levels. Stanford, CA: National Academy of Education

Williams, N.J., & Schulz, E.M. (April, 2005). *An investigation of response probability values used in standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Wyse, A.E., Bunch, M.B., Deville, C., & Viger, S.G. (2013). A Body of Work standard-setting method with construct maps. *Educational and Psychological Measurement, 74*(2), 236-262.

Zenisky, A., Hambleton, R.K., & Sireci, S.G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359-375.

Zwick, R., Senturk, D., Wang, J., & Loomis, S.C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20(2)*, 15-25.