NAEP ACHIEVEMENT LEVELS VALIDITY ARGUMENT REPORT

National Assessment Governing Board

Anne H. Davidson, Ed.D.

Developed for the National Assessment Governing Board under contract number 91995921F0002 by
The Manhattan Strategy Group LLC, with a subcontract to EdMetric, LLC

Table of Contents

Executive Summary

The purpose of this National Assessment of Educational Progress (NAEP) Achievement Levels Validity Argument Report is to synthesize evidence currently available to address the validity of the interpretations and uses of the NAEP Achievement Levels. Validity is the extent to which theory and evidence supports or refutes proposed and enacted test score interpretations and uses. This report was conceptualized as one of many activities outlined in an Achievement Levels Work Plan in response to recommendations made by the National Academies of Sciences, Engineering, and Medicine (NASEM) in their independent evaluation of the NAEP Achievement Levels (NASEM, 2017). This report touches on all NAEP subject areas but is primarily focused on the achievement levels associated with reading and mathematics because these subjects were the focus of NASEM's evaluation and are the most frequently assessed subject areas for NAEP.

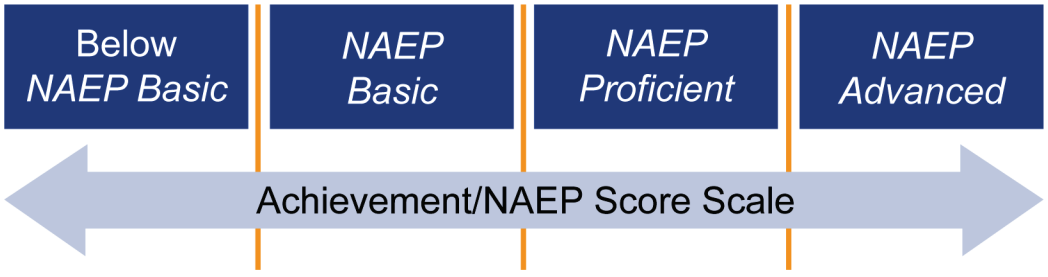
This report outlines accumulated validity evidence that can be used to clarify the appropriate uses and interpretations supported by that evidence, noting common inappropriate uses and interpretations not supported by evidence. Though this report is technical in nature and may be of interest primarily to testing and measurement professionals and others who work closely with large-scale assessment programs, this executive summary seeks to summarize key points accessible by a wide audience interested in understanding how to interpret and use the NAEP Achievement Levels.

This report is not an independent evaluation of the NAEP Achievement Levels overall or for any specific NAEP assessment. Further, we do not draw conclusions about whether there is sufficient evidence to support the removal of the trial status of the achievement levels, as indicated by legislation (Improving America's Schools Act, 1994), a determination to be made by the commissioner of the National Center for Education Statistics (NCES) based on an external evaluation of the NAEP Achievement Levels. This report may, however, help these endeavors by synthesizing available evidence to clarify what has and has not been accomplished to support validity.

Before getting into the validity evidence, it is important to understand what constitutes a NAEP Achievement Level. The NAEP Achievement Levels provide classifications of levels of skills and knowledge. NAEP defines three levels—*NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced.* The achievement levels are based on the NAEP score scale, and the cut scores for reaching each level are developed through rigorous standard-setting processes that are described in Chapter II of this report.

Figure 1 illustrates how the achievement levels relate to the scores on a test scale.

Figure 1
NAEP Score Scale and Achievement Levels



The National Assessment Governing Board (Governing Board) is legislatively responsible for the NAEP Achievement Levels. Since the first NAEP Achievement Levels policy was set by the Governing Board in 1990, there have been numerous standard settings, evaluations, policy developments, and modifications, as well as studies that have sought to ensure the three NAEP Achievement Levels provide an accurate and appropriate reflection of what students know and can do for each subject area and grade level assessed by NAEP. This compilation of information from 1990 through the present can be considered the body of evidence for the validity of the interpretations and uses of NAEP Achievement Levels. The purpose of this NAEP Achievement Levels Validity Argument Report, therefore, is to summarize the available evidence collected to date to inform the validation of the achievement levels.

This report examines validity evidence to support three major claims: (1) NAEP Achievement Levels are established based on defensible standard-setting methods that are implemented with fidelity, (2) NAEP Achievement Level Descriptions (ALDs) are defensible definitions of what students know and can do at each level, and (3) NAEP Achievement Levels meaningfully relate to other measures of student achievement and other indicators of educational outcomes for all students.

It uses a validity framework described by Kane (2006, 2011, 2013) that classifies existing evidence as (1) procedural evidence, (2) internal evidence, and (3) external evidence (Kane, 1994).

- *Procedural evidence* demonstrates the appropriateness of the procedures used and the quality of those procedures' implementation. This type of evidence is particularly important because it is relatively concrete and widely accepted as a basis for policy decisions.
- *Internal evidence* shows the consistency of the various results of a standard-setting or other evaluation study. This type of evidence is important because it provides support for the overall validity argument by checking the presumed relationship between the performance standard (i.e., achievement level) and the cut score on the test scale.
- *External evidence* is based on comparison with external sources of information related in a meaningful way to the expectations captured in the performance standards. These comparisons are rarely exact, but they are rough indications of whether a performance level is "too high, too low, or about right" (Kane, 1994, p. 448). All types of evidence are relevant in validity argumentation for achievement levels.

Claims regarding the content assessed are specific to each content area and describe what students at a given level know and can do. These claims should be represented in the ALDs and undergird the validity argument to support the claim that NAEP ALDs are defensible definitions of what students know and can do at each level. These ALDs should be consulted when seeking to understand the knowledge and skills to define student performance by NAEP Achievement Level at each grade level and content area. The accuracy and meaningfulness of the ALDs can be evaluated based on the procedural, internal, and external evidence presented throughout this report.

The NAEP Achievement Levels include a system of aligned statements that capture greater or lesser detail and are used for various purposes: policy, threshold, reporting, and range ALDs (described in this report). The ALDs collectively provide the content basis for alignment between the NAEP scores and their interpretation and use. One approach to evaluating the NAEP Achievement Levels is to examine the alignment between the policy that guides their

development and ongoing evaluation. The Governing Board's (2018) policy for developing student achievement levels for NAEP and best-practice standards in the field of educational measurement (American Educational Research Association [AERA] et al., 2014; Brennan, 2006; Joint Committee on Testing Practices, 2004; National Committee on Educational Statistics [NCES], 2012) guided the development of the NAEP ALDs.

Internal evidence has been collected to examine whether the achievement levels measure what we claim they measure—in other words, do students performing at each NAEP Achievement Level demonstrate the knowledge and skills described in the ALDs? The ALD review studies for NAEP Reading and Math at grades 4, 8, and 12 and NAEP science, U.S. history, and civics at grade 8 represent the most recently collected measure of internal evidence. These studies help ensure that the following example statements for what grade 8 students who score at the *NAEP Proficient* level know and can do are credible.

- In **mathematics** in grade 8, students performing at the *NAEP Proficient* achievement level likely can identify appropriate units or tools of measurement within the same system, convert measurements within the same system, measure lengths of objects to the nearest whole or half unit, and solve or estimate problems involving area.
- In **reading**, when reading literary texts such as fiction, poetry, and literary nonfiction, eighth-grade students performing at the *NAEP Proficient* level likely can use context explicitly and implicitly across the entire text to determine the meaning of words and nonliteral phrases; make inferences and draw conclusions about varied literary elements such as character interactions, comparison of characters, plot features, and theme; support ideas with relevant examples from the text and provide some explanation about the connection between the ideas and evidence; and provide a reasonable opinion supported by some evidence from the text.

The results from standard-setting, anchor, and alignment studies, including review studies, provide a degree of confidence in the ALDs for all grades and content areas, with the exception of *NAEP Advanced* for math at grade 12 findings during a 2022 ALD review study (Moyer & Galindo, 2022; see Chapters IV and VI).

It is important to note not only appropriate interpretations and uses for the NAEP Achievement Levels, but also inappropriate ones when considering their validity. For example, evidence does not support using NAEP Achievement Level data to make statements about the percentage of students at grade level. This common misinterpretation is inappropriate because there is no common definition of grade level in the United States. Rather, grade-level expectations are set as state policy and described in content standards and curriculum. These can change over time as lawmakers craft education laws that drive education leaders to set state policies. Further discussion of inappropriate uses and interpretations are included in Chapter V.

External evidence has also been collected for the NAEP Achievement Levels through studies that tie them to other academic measures of achievement and outcomes. NAEP does not intend to replicate any other measure, so we do not expect the NAEP Achievement Levels to align perfectly with any single external measure. Rather, external evidence can help corroborate what it means to perform at each level. Several studies have linked NAEP to other national, longitudinal surveys that offer external evidence relevant to this validity argument. For example, a study was conducted in which a common set of students took the grade 12 NAEP Math assessment and participated in the National High School Longitudinal Study (HSLS). The HSLS provided information about student characteristics and academic achievement. This linkage

allowed researchers to examine student achievement by NAEP Achievement Level. In one examination, researchers found that NAEP Achievement Levels were correlated with college and university attendance—with those who performed at higher levels more likely to attend. Further examples of such research are highlighted in Chapter IV of this report.

While there have been some rigorous studies conducted to examine the link between NAEP and external measures, the Governing Board acknowledged that more should be done. In August 2023, the Governing Board adopted a Linking Studies Resolution affirming the importance of these studies and encouraged the National Center for Education Statistics to link NAEP assessment data with data from other data sources in collaboration with the Governing Board, share linked datasets with researchers in adherence with privacy and confidentiality protections, and further disseminate information learned from linking studies to the public. (National Assessment Governing Board [Governing Board], 2023, p. 1).

Evidence from external studies supports the claim that NAEP Achievement Levels meaningfully relate to other measures of student achievement and other indicators of educational outcomes for students, allowing for some degree of comparison with these external sources of information that align with the expectations captured in the achievement levels.

All types of evidence are relevant in the validity argumentation for NAEP Achievement Levels. It is important to note that, while this report centers on the NAEP Achievement Levels, it necessarily includes discussion of scale scores, since achievement levels are dependent on the cut scores that define the range of the levels on each test scale. This evidence therefore contributes to the discussion of how validity related to scale scores contributes to the validity argument for the achievement levels.

This report does not intend to address whether the cut scores are correct. There are various arguments whether the NAEP Achievement Levels are set at the right level or are too high. On one hand, NAEP Achievement Levels may be more rigorous than those from other large-scale assessment programs. As highlighted by studies that map NAEP Achievement Levels to state achievement levels described in Section IV, NAEP Achievement Levels are higher than the majority of state achievement levels. That said, NAEP Achievement Levels and assessment frameworks were developed through a rigorous and collaborative process with educators and other content experts from across the country who have used their expertise to identify the knowledge and skills expected from students performing at each level. NAEP has the advantage of remaining consistent with these expectations over the years since it is not directly influenced by current politics or by the demands of high-stakes decisions for students or schools.

The Governing Board acknowledges that validity is not a definitive quality of any assessment, nor is validation a singular activity. More work must be done as the assessment frameworks and student populations shift and as new technologies and methodologies are identified. For example, at the time of this report there are rapid developments in generative artificial intelligence (AI) that may impact how students learn and how they are assessed, which may have implications for NAEP. In addition, AI advancements could potentially impact standard-setting and achievement-level review methodologies. Furthermore, as the educational landscape changes over time, more validation efforts will be needed to address new or emerging research questions. The need to continually consider the validity of interpretations and uses of the NAEP Achievement Levels is addressed in Governing Board achievement level policy.

Glossary

Term	Definition
Achievement level description (ALD)	Descriptions of the related student performance intended to provide interpretive guidance for the achievement levels. These descriptions can also be called achievement level *descriptors*.
Achievement levels	Descriptions of students' levels of competency in a particular area of knowledge and/or skill, defined in terms of categories ordered on a continuum (for NAEP, *NAEP Basic* to *NAEP Advanced*). The categories constitute broad ranges for classifying performance. In other programs, similar levels can be called *performance levels* or *proficiency levels*.
Alignment study	As applied in this report, a study of the degree to which the content and cognitive demands of test questions match targeted content and cognitive demands described in the achievement levels
Anchor study	A study that evaluates whether there is evidence of a problem with the location of a given cut score on a test scale (Loomis, 2018). They are called anchor studies because they use an anchoring, or item-mapping, methodology that anchors test items to the achievement levels on the test scale.
Assessment framework	A test development document that details ALDs along with the test specifications (i.e., test blueprints) for the content and design of an assessment. The NAEP assessment frameworks are designed to remain stable for as long as possible. At the same time, all frameworks are responsive to changes in national and international standards and curricula.
Claim	An affirmative statement of the proposed interpretation and use of a test score within the "network of inferences and assumptions inherent in the proposed interpretation and use" (Kane, 2013, p. 2). Claims drive the validation process.
Criterion-referenced	The characteristic of a test score for an individual or, in the case of NAEP, an average score for a defined group, indicating the individual's or group's level of performance in relationship to some defined criterion domain. In the case of NAEP, this criterion is defined through the NAEP assessment frameworks. Examples of criterion-referenced interpretations include domain-referenced score interpretations. Criterion-referenced scores contrast with norm-referenced score interpretation (see *norm-referenced*).
Governing Board	National Assessment Governing Board
Interpretation and use argument	The argument that specifies what assessment scores are intended to mean and how they can be used
Item map	A presentation of a list of items in order of their difficulty across the range of difficulty of the test. It can be mapped to an ordered item booklet with items presented in the same order as the item map. For NAEP, it is also used for score interpretation (see Figure 5-1).
Linking study	A study that establishes a statistical relationship between two test scales so they can be expressed on the same scale. This allows for meaningful comparison between two different scales. For the purpose of this report, linking studies relate NAEP Achievement Levels to external measures of academic success, and outcomes constitute external evidence for the validity of interpretations and uses.

Term	Definition
Mapping study	A study that establishes where each state's performance standards (i.e., achievement levels) fall on the NAEP scales and in relation to the NAEP Achievement Levels. Given patterns of states' performance standards adoptions, these results show a degree of reasonableness.
Norm-referenced	The characteristics of a test score based on a comparison of a student's performance with the distribution of performance in a specified reference population. NAEP is not a norm-referenced testing program. Norm-referenced scores contrast with criterion-referenced score interpretation (see *criterion-referenced*).
Ordered item booklet	An ordered item booklet presents items in order of their difficulty across the range of difficulty of the test, as indicated by an item map.
Performance descriptor	General term referring to descriptions of what test takers know and can do at specific performance levels. NAEP ALDs are performance descriptors.
Performance standards	Descriptions of levels of knowledge and skill acquisition contained in content standards, as articulated through performance-level labels (e.g., *NAEP Basic, NAEP Proficient, NAEP Advanced*). They state what test takers at different achievement levels know and can do. Cut scores or ranges of scores on the scale of an assessment differentiate performance standards.
Policy ALD	A broad (general) ALD used to communicate with broad audiences and across grade levels and content areas
Range ALD	An ALD developed at the most detailed level for the purposes of developing test items at a particular achievement level and for transparent insight into the specific knowledge and skills students can do if performing at the given achievement level
Reporting ALD	An ALD used for reporting purposes and succinctly identifying key knowledge and skills demonstrated by students scoring within the achievement level
Response probability (RP)criterion	A statistical specification for the likelihood that a student would get the item correct at a given point on the scale
Scale score	A score obtained by transforming raw scores from a test onto a scale. Scale scores are typically used to facilitate interpretation.
Score	Any specific number resulting from the testing of an individual. Achievement levels represent a range of test scores.
Standard setting	The process of setting cut scores using a structured procedure that seeks to map test scores into discrete achievement levels that are specified by ALDs
Threshold ALD	An ALD used for standard-setting purposes, which is both consistent with policy ALDs but is more detailed and focused on the knowledge and skills of the borderline or threshold student. This student is one whose performance is just barely within the specific ALD.
Validation	A process of constructing a logical argument that clarifies the meaning of test scores. It "requires a clear statement of the claims inherent in the proposed interpretations and uses of the test scores" (AERA et al., 2014, p. 1). Engaging in the validation process, the intended interpretations and uses for test results are made transparent and accessible. Note that achievement levels are essentially extensions of the test scores themselves, and therefore they factor in the validation process.

Term	Definition
Validity	The degree to which accumulated evidence and theory support a specific interpretation and use of test scores. If a test score is used or interpreted differently than originally intended, validity evidence for each use and interpretation is needed.
Validity argument	An explicit justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) and uses of test scores. A validity argument evaluates the relative success of the evidence to support those intended interpretations and uses.
Validity framework	The structure of a validity argument. The framework establishes what evidence must be collected to support the interpretation and use of given test scores, including achievement levels, and the criteria for judging the evidence.

I. Introduction

The National Assessment Governing Board (Governing Board) first identified the need to develop an achievement levels validity argument following recommendations made in an evaluation of the National Assessment of Educational Progress (NAEP) Achievement Levels by the National Academies of Sciences, Engineering, and Medicine (NASEM) (NASEM, 2017). In their evaluation, NASEM offered seven recommendations to help strengthen validity evidence and better articulate the intended interpretations and uses of the NAEP Achievement Levels. The Governing Board responded by adopting an Achievement Levels Work Plan in 2020, with planned activities to address the recommendations. One of these activities was to develop a report to summarize validity evidence to support intended interpretations and uses of the NAEP Achievement Levels. In 2023, the Governing Board's Committee on Standards, Design and Methodology (COSDAM) developed an outline of a validity argument to guide development of this document.

Priorities for this NAEP Achievement Levels Validity Argument Report were to build it upon highly regarded validity research in the field of educational measurement; to encompass procedural, internal, and external evidence to support claims regarding the achievement levels; and to present the evidence in a manner that allows a reader to easily identify what has been done for each of these categories and thus be able to identify the strengths and weaknesses of the validity argument. In addition, this validity argument is intended to illustrate appropriate uses and interpretations of the NAEP Achievement Levels based on the available evidence and present common inappropriate uses and interpretations that are not supported by the evidence.

This report examines validity evidence to support three major claims: (1) NAEP Achievement Levels are established based on defensible standard-setting methods that are implemented with fidelity, (2) NAEP Achievement Level Descriptions (ALDs) are defensible definitions of what students know and can do at each level, and (3) NAEP Achievement Levels meaningfully relate to other measures of student achievement and other indicators of educational outcomes for all students.

Background

In 1990, the Governing Board set the first policy on achievement levels for NAEP. As a federally authorized assessment program, NAEP provides policymakers, educators, and the public with reports on the academic performance and progress of the country's students. These achievement levels, also called *performance standards*, are defined by cut scores on the test scales, and they have established ranges of scores intended to be interpreted and used by stakeholders for various purposes. Descriptions of the related student performance, or *achievement level descriptions* (ALDs), are intended to provide interpretive guidance for the achievement levels. In short, the inclusion of achievement levels provides guidance for the interpretation and use of the NAEP scores, serving as standards of academic performance.

The adoption of the NAEP Achievement Levels reflected evolving best practices from the field of educational measurement and signaled the overall program's commitment to ongoing improvement through research and innovation. This report presents claims made about the NAEP Achievement Levels along with empirical evidence, relevant literature, and logical analyses that support these claims. The concept of validity has evolved in the research literature since the first NAEP standard settings in the early 1990s. Therefore, professional standards and expectations for validity evidence have also evolved. Furthermore, data sources continue to

expand. This report will serve as a foundational document to describe efforts taken from the initial development of the NAEP Achievement Levels up through the release of this report to support the valid interpretation and use of the NAEP Achievement Levels.

This report is intended to provide a current and comprehensive discussion of validity evidence related to the NAEP Achievement Levels. The report first defines and describes the development and evolution of the NAEP Achievement Levels, including historical and theoretical context. Second, the report provides the validity framework for evaluating the NAEP Achievement Levels and documents a comprehensive compilation of existing, credible evidence. The discussion is organized around the validity argument (Kane, 2006, 2011, 2013), providing a line of logic that is supported by evidence. The validity argument makes claims about how test scores are intended to be interpreted and used, as well as detailing the assumptions underlying those claims. It follows that the validity framework also allows for identification of any inappropriate interpretations and uses and therefore reduces the risk that results are used inappropriately or for unintended uses or purposes.

This report's primary readers will have some familiarity with educational measurement, assessment development and evaluation, and achievement levels in general. However, this report is also intended to be accessible to a broader audience interested in NAEP and its validation, and the executive summary was written with a wide audience in mind.

Framing Validity Evaluation

In service to these goals, this report begins with a brief discussion of the theoretical foundation of validity argumentation with a focus on achievement levels. Best-practice standards for educational testing describe *validity* as a unitary, or central, concept that focuses on "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (American Educational Research Association [AERA] et al., 2014, p. 1). In other words, validity is "the extent to which theory and evidence supports or refutes proposed and actual or enacted test score interpretations, uses, and consequences" (Lane & Marion, in press). Thus, the degree to which the NAEP Achievement Levels and associated performance descriptors convey meaningful and credible information to various NAEP audiences is a critical component of the validity of the interpretations and uses of the achievement levels for the various NAEP assessments. A *validity framework* establishes what evidence must be collected to support the interpretation and use of the achievement levels and the criteria for judging the evidence. A validity framework starts with the intended interpretations and uses for scores, what Kane (2013) termed the "interpretation and use argument." This argument then drives decisions on the types of evidence required and criteria by which the evidence can be judged so that *claims* made for each score or achievement level interpretation or use are defensible. The claims, or "the network of inferences and assumptions inherent in the proposed interpretation and use" (Kane, 2013, p. 2), drive the design of the validity framework. *Validation* involves implementing the validity framework to construct a logical argument that clarifies the meaning of test scores. It "requires a clear statement of the claims inherent in the proposed interpretations and uses of the test scores" (AERA et al., 2014, p. 1). Engaging in the validation process, the intended interpretations and uses for test results are made transparent and accessible. Since achievement levels are essentially extensions of the test scores themselves, they factor significantly in the validation process.

Kane (2013) argued that test scores are of particular interest because they can support claims that have meaning to the public users of those scores. These meanings, or interpretations, are

important because they both justify and provide a basis for evaluating the results of testing. Together, the interconnection among claims, test scores, and their interpretations and uses constitutes the substance of the validity framework. This ensures that scores possess clear meaning and users can use the scores with justification. The claims are important to the evaluation process, creating the foundation on which decisions can be made and delineating the applicable conditions.

Kane (2013) described the interdependent nature of the *interpretation and use argument*, which specifies what the scores are intended to mean, and the *validity argument*, which evaluates the relative success of the evidence to support those intended interpretations and uses. Several assumptions surface in this dynamic interrelationship and are relevant to the validity framework for the NAEP Achievement Levels. First, there must be a plan to collect the necessary validity evidence to support claims made about the NAEP Achievement Levels. This means defining the types and sources of evidence that are relevant and necessary, collecting the evidence over time, and evaluating the evidence considering the framework. Second, validation is an ongoing process (Messick, 1989) since important public claims require ongoing justification, transparency for all stakeholders, and consideration of the consequences of the scores. All these factors can affect the interpretation and use argument and can change over time. Therefore, the validity discussion must be a living document, open to public scrutiny with processes for revision or amendment.

Note that the report compiles and presents accumulated evidence that supports claims about the valid use and interpretation of the NAEP Achievement Levels, with particular emphasis on mathematics and reading given the importance of these two subject areas in American education, the focus on them in NASEM's evaluation of the achievement levels (NASEM, 2017), and the availability of related extant research. It is beyond the scope of this report to build a singular validity argument for each of the NAEP assessment program's achievement levels. Rather, the report presents evidence organized around the concepts described by Kane (2006, 2011, 2013) and the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

Sections of the Report

Section I of the NAEP Achievement Levels Validity Argument Report presents a brief introduction to set the stage for the remainder of the report by defining terms and highlighting literature to establish best-practice expectations for validity argumentation and validation processes.

Section II describes the overall purpose of NAEP and the NAEP Achievement Levels, outlining the history of NAEP itself, when NAEP first adopted the achievement levels, and how they have evolved over time. Section II identifies the subject areas and grade levels assessed and their frequency of assessment, with the policy definitions. Finally, Section II states the major claims that can be made using the achievement levels.

Section III details the development of the NAEP Achievement Level policy guidance and the processes used to implement that policy. Section III demonstrates how the policy and processes adhere to best practices in the field of educational measurement, including the *Standards for Educational and Psychological Testing* (AERA et al., 1999, 2014), and presents the NAEP ALDs within the validity discussion at varying levels of granularity.

Section IV presents evidence accumulated to shore up the validity argument for the NAEP Achievement Levels. This evidence comes from three primary sources: standard-setting

processes, content reviews with a focus on alignment evaluation, and external validity studies of the achievement levels, including linking studies and state mapping studies.

Section V looks at the appropriateness of known interpretations and uses of the NAEP Achievement Levels. Section V also examines the claims, considering how the NAEP Achievement Levels indicate student academic performance and how they differ from existing state achievement levels. It then examines the relationship to external measures of achievement and college preparedness, including what the findings contribute to the validity argument. Finally, Section V uses the NAEP Achievement Levels to inform the understanding of differences in state achievement levels.

II. Purpose of NAEP and NAEP Achievement Levels

History and Purpose of NAEP

The broadest purpose of NAEP is to provide ongoing evaluation of education in the United States. More specifically, it should "(1) monitor continuously the knowledge, skills, and performance of the nation's children; and (2) provide objective data about student performance at the national and regional levels, the state level (since 1990), and for large urban school districts (since 2002)" (National Center for Education Statistics [NCES], 2011, p. 227). Now more than five decades old, the assessment program measures the academic performance of students in various subject areas, including reading and mathematics. It uses a large, nationally representative sample in grades 4, 8, and 12 from 53 states and jurisdictions and 26 urban school districts.

NAEP has two components: a main component and a long-term trend component. NAEP conducts the long-term trend at the national level only; the primary data collected is related to student performance and educational experience. The subjects assessed in long-term trend NAEP are mathematics and reading; for the main NAEP, subjects include reading, writing, mathematics, and science assessed at the national, state, and district levels. At grades 4 and 8, mathematics and reading are assessed every two years, and at grade 12, they are assessed every four years. NAEP assessments in other subject areas are assessed at various intervals specified in the Governing Board's Assessment Schedule based on Governing Board priorities. Long-term trend NAEP does not include achievement levels; therefore, the focus of this report is the main component of NAEP.

It is important to note that NAEP is a *criterion-referenced test* and therefore produces *criterion-referenced scores*. This means that the test scores, in this case for a defined sample of students, indicate the group's level of performance in relation to a specific criterion domain as defined by the NAEP assessment frameworks. For NAEP, this means that the cut scores used to define the NAEP Achievement Levels are based on reaching a defined minimum level of understanding of the content being assessed. There is no limit to the percentage of students that can perform at any level. This is not the same as *norm-referenced tests* that produce *norm-referenced score interpretations*, which base comparison of student performance on a distribution of performance in a specified reference population. On a norm-referenced test, a student's performance is impacted by the performance of students for which the normed scale was set.

Historical Context

In the 19th century, Congress passed legislation authorizing the United States Office of Education to provide an annual report to Congress on the state of American education. A century later, concerns about whether American students were learning what they needed to be globally competitive in the 21st century drove changes to policy, implementation, and governance of the National Report Card's metrics. Therefore, in 1969, NAEP became a federally funded program to fulfill this new mission with a focus on gathering information on the outcomes of education, or what students know and can do. The Education Commission of the States was charged with planning and administering the initial NAEP assessments. Within two years, the United States Office of Education transferred administrative responsibility to NCES, where it remains.

Intent for the Achievement Levels

In the mid-1980s, the National Governors Association (Alexander, 1986) called for more accurate comparisons of states' performances to each other and to the nation, resulting in new legislation (PL. 107-279) that in 1988 authorized a Governing Board that would be responsible for the program policy. It authorized NCES to oversee program administration and contractors to provide technical expertise.

The Governing Board was immediately tasked with developing performance standards for NAEP assessments. The law required that the Governing Board "identify appropriate achievement goals for each age and grade in each subject area to be tested" (Sec. 3403, (6)(A). In another section, it noted that "each learning area assessment shall have goal statements devised through a national consensus approach" (Sec. 3403, (6)(E). With room for interpretation and the need for definition (Bourque, 2009), the Governing Board developed policy that included calling the required achievement goals *achievement levels* and set policy guided by expert advice and best practices in the field of educational measurement (AERA et al., 1999). Further, board policy emphasized engaging a broad spectrum of stakeholders to develop and report student achievement levels.

According to current NAEP legislation, the Governing Board is charged with developing achievement levels for all NAEP assessments. Under provisions of the National Assessment of Educational Progress Authorization Act of 2002, Congress authorized the Governing Board to develop "achievement levels that are consistent with relevant widely accepted professional assessment standards and based on the appropriate level of subject matter knowledge" (Section 303[e][2][A][i][II]). Given this mandate, the Governing Board must ensure that all achievement level-setting processes and procedures align with current best practices for standard setting and that appropriate validity evidence is collected and documented to support the intended uses and interpretations of the NAEP Achievement Levels. To fulfill this purpose, the Governing Board established the policy definitions for the NAEP Achievement Levels, detailed in "Current NAEP Achievement Levels" later in this chapter. These levels outline what students should know and be able to do and maintain consistency across all assessments in which achievement levels are set (Governing Board, 2018).

Importance of Achievement Level Descriptions

The achievement goals for each grade and subject that were called for by the legislation are referred to as achievement level descriptions (ALDs). The general concept and practical application of ALDs have evolved since the first NAEP Achievement Levels were adopted (Egan et al., 2011). Most states use ALDs to develop performance standards on their high-stakes testing programs even though ALDs were almost unheard of prior to the NAEP Achievement Levels (Bourque, 2009). Now, all currently accepted standard-setting methods use ALDs (Hambleton & Pitoniak, 2006), and experts in standard setting agree that ALDs are critical to interpreting and using achievement levels appropriately (Hambleton, 2001; Mills & Jaeger, 1998). Over time, ALDs have been shown to help panelists conceptualize and internalize how scores represent student performance across the score range, supporting a common, criterion-based understanding of the expectations of student performance. The diversification and frequency of stakeholder involvement in both the development of ALDs and in their use for standard setting have also increased over time.

The use of ALDs has touched all aspects of education, including curriculum, instruction, and assessment (Bejar et al., 2007; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Hansche,

1998). Egan et al. (2011) described how ALDs can be developed at four levels of granularity to address different purposes within a coordinated and aligned criterion-based system. These levels included *policy descriptors* used to communicate with broad audiences and across grade levels and content areas. Another type of ALD, termed *threshold ALDs*, is used for standard setting, which is both consistent with policy descriptors but more detailed and focused on the knowledge and skills of the borderline or threshold student. A third type of ALD, called *reporting ALDs*, is used for reporting purposes and succinctly identify key knowledge and skills that students scoring within the achievement level demonstrate. Finally, a fourth type of ALD, *range ALDs*, is developed at the most detailed level for the purposes of developing test items at a particular achievement level to provide transparent insight into more specific knowledge and skills that students typically possess if performing at the given achievement level.

For example, the *NAEP Proficient* range ALD outlines the content expectations for students whose performance is in the *NAEP Proficient* range, which is the *NAEP Proficient* cut score up to the *NAEP Advanced* cut score on the test scale. The ALDs used for training standard-setting panelists to assess item content and recommend cut scores are necessarily focused on the student expectations right around the cut score. Panelists consider the questions “What does this item measure? Why is this item more difficult than the items that precede it?” The descriptions at the borderline or threshold are a subset of the ALDs and include the content necessary to have just reached that level, such as *NAEP Proficient*. By coordinating these ALDs within a system of policy, reporting, and range ALDs, the program can show procedural evidence of alignment and coherence.

In addition, Egan and colleagues (2011) described the intended uses for ALDs within three broad categories: standard setting, test development, and score interpretation. Since ALDs serve multiple purposes, they play a critical role in supplying alignment and coherence across the broader system of curriculum, instruction, and assessment. In other words, ALDs are inputs to various interrelated processes in a criterion-referenced system by outlining the content that students should know in their designated score range. For example, Ferrara et al. (2011) argued that students who perform at a given achievement level on a test are expected to demonstrate that they have mastered most of the knowledge and skills represented by the items at and below the cut score. Therefore, “[it] is important that these items define intended knowledge and skills, especially increasing levels of knowledge and skills, on tests that are intended to portray achievement growth across grade levels” (p. 3). To ensure coherent inferences about what it means to achieve at a given performance level, item writers, or content experts tasked to write the test items, can be trained to hit assigned achievement level targets using ALDs.

Over time, the Governing Board has established consistency in applying a process for developing the ALDs (NASEM, 2017). Per their policy, the Governing Board (2018) sets the policy definitions (described in the next subsection), which are given to consensus panels or independent content panels. Next, the draft definitions are widely vetted with the variety of NAEP audiences, including additional content specialists, stakeholders, NAEP score users, and policymakers. Range, reporting, and threshold ALDs can then be developed. Results of these processes are incorporated into revised ALDs that the Governing Board then approves before the standard-setting panels use them to develop their recommendations on cut scores and exemplar items. Finally, this collection (i.e., recommended cut scores, descriptions, and exemplars) is brought to the Governing Board for review and final approval of the ALDs.

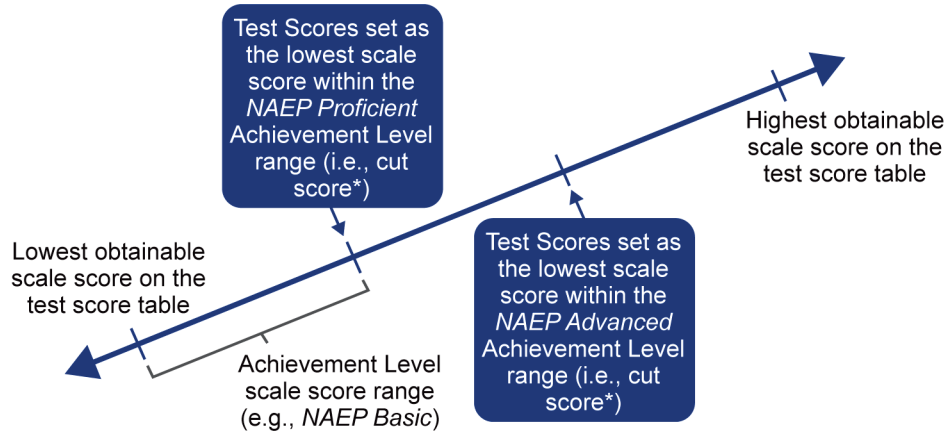
Since that time and with all variations in the process of standard setting, the ALDs have served a critical role. First, their development provides a description of what students should know and be able to do to qualify for performance at each of the three NAEP Achievement Levels. When panelists then judge whether students meet those requirements, they can better understand the specific knowledge or skill required of students to answer items correctly. The cut score is therefore more defensibly set than without such a description and panelists' resulting understanding of appropriate expectations. In this way, cut scores represent the minimal performance required for each achievement level. For example, the *NAEP Basic* cut score represents the minimal performance to meet the requirements described for that level, and the *NAEP Basic* achievement level extends to the cut score for the *NAEP Proficient* achievement level.

All the variations in the process of standard setting have also included panelists, or judges, who are experts in the specific subject matter and student populations, including classroom teachers in the subject areas and grades being assessed by NAEP, other educators (such as college faculty and curriculum directors), as well as representatives of the general public who are trained in the content area and have valuable knowledge of the skills and educational requirements for students at the grade levels.

The result of each standard-setting process is the identification of the cut scores on each NAEP scale that correspond to the lower boundary of each achievement level. Along with the cut scores, panelist judges also select exemplar items that can be used as good examples of the kinds of knowledge and skills that students in each achievement level can likely answer or demonstrate correctly. The collection of cut scores, all ALDs (including policy definitions), and exemplar items constitute the NAEP Achievement Levels.

It is important to note that, while this report centers on these achievement levels, it necessarily includes discussion of scale scores, since achievement levels include the cut scores that define the scale score range of the levels on each test scale. Evidence that inspects how scale scores relate within and across tests can contribute to the validity argument for the achievement levels. Figure 2-1 illustrates the relationship between scale scores, including cut scores, and the achievement levels.

Figure 2-1
Illustration of Achievement Level Scale Score Ranges on a Generic Test Scale



* Cut scores are set during the standard-setting process.

Current NAEP Achievement Levels

The achievement levels are composed of (1) policy definitions for the *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced* levels; (2) specific ALDs for each assessment; (3) cut scores that demarcate adjacent levels; and (4) exemplar items or tasks that illustrate performance at each level. The validity argument supports this system of performance standards, which were developed in accordance with widely accepted professional standards, to ensure score results are reasonable, useful, and informative to the public.

The policy definitions used across all subject areas and grade levels for the NAEP Achievement Levels are as follows:

- *NAEP Basic*. This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for performance at the *NAEP Proficient* level.
- *NAEP Proficient*. This level represents solid academic performance for each NAEP assessment. Students reaching this level have demonstrated competency over challenging subject matter, including subject matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
- *NAEP Advanced*. This level signifies superior performance beyond *NAEP Proficient*.

NAEP Achievement Levels are intended to be cumulative, such that student performance at the *NAEP Proficient* level includes the competencies associated with the *NAEP Basic* level, and the *NAEP Advanced* level includes the skills and knowledge associated with both the *NAEP Basic* and the *NAEP Proficient* levels.

The range-level descriptions of what students at grades 4, 8, and 12 should know and be able to do have been developed at the three achievement levels for each NAEP assessment and are included in the NAEP assessment frameworks. In 2022, reporting ALDs that describe student knowledge and skills based on student performance on items associated with each grade level were established for all grades for NAEP Reading and Math and for grade 8 for NAEP science, U.S. history, and civics (Moyer & Galindo, 2022).

The range of performance that does not reach the cut score for *NAEP Basic* is referred to as below *NAEP Basic*. This low end of the achievement scale has received attention in recent years, including the 2022 release of NAEP Reading and Math data when student achievement levels showed significant decreases due to the COVID-19 pandemic. The Committee on Standards, Design and Methodology (COSDAM) has had discussions during Governing Board quarterly meetings on whether the increased focus on the low end of the scale warranted reconsideration of policy to include three achievement levels; however, it was decided that the focus should be on increasing the number of items at the low end of the scale to provide more information about the knowledge and skills for a greater range of student performance.

The next section will describe the approaches the Governing Board has taken to set the cut scores for the three achievement levels.

Overview of the Evaluation of NAEP Achievement Levels

Since this early charge, the story of the NAEP Achievement Levels has been a dynamic one (Bourque, 2009) in which NAEP has both led and been informed by the states' assessment program designs. Over time, the NAEP Achievement Level processes have reflected current best practices from the field of educational measurement as they themselves have developed in

sophistication through the years (AERA et al., 1999, 2014). On the other hand, the NAEP Achievement Levels have influenced states' assessment programs, especially under the No Child Left Behind Act of 2001, where expectations for statewide assessment systems included setting performance standards. Many states attended to the rigor of the NAEP's performance standards as a model when setting standards for their own state assessments. In this way, policy, technical, practical, and contextual issues have all played into the development and application of the NAEP Achievement Levels over time.

When the Governing Board developed its initial policy for developing NAEP Achievement Levels in 1990, the field of standard setting was much less developed than it is today and most performance standards had been established around professional licensure and certification (Bourque, 2009). Also at that time, little standard setting had been conducted in K–12 education, particularly at the national level (Bourque, 2009). Therefore, the Governing Board's initial efforts to develop achievement levels for NAEP were groundbreaking, and the Governing Board's research and practices in this area over the past four decades have influenced how standard setting has been conducted in K–12 education, especially at the state level. However, early evaluations of the NAEP Achievement Levels conducted in the 1990s criticized the Governing Board's procedures and reflected distrust of standard-setting approaches in general. When the initial NAEP legislation was reauthorized in 1994, it included language that the NAEP Achievement Levels should be "used on a developmental basis until the [NCES] Commissioner determines, as a result of an evaluation ... that such levels are reasonable, valid, and informative to the public" (Improving America's Schools Act, 1994). The term "developmental basis" was replaced with "trial basis." The NCES commissioner has the legislative authority to determine when there is sufficient evidence the levels are reasonable, valid, and informative to the public, as the result of an evaluation, to drop the trial designation.

In 1999, a report of an evaluation conducted by the National Research Council (NRC) stated that the process for setting NAEP Achievement Levels was "fundamentally flawed" (p. 7). Many technical experts disputed this conclusion, and several prominent researchers issued a response criticizing the NRC evaluation. Note that, although the Governing Board had adopted innovative approaches to developing the NAEP Achievement Levels since the 1990s, the original methodology (i.e., modified Angoff; Angoff, 1971) is still considered a reputable approach to standard setting and remains in wide use today. This chapter delves deeper into the Angoff method, providing additional clarification within the subsection titled "Methodologies Used to Set NAEP Achievement Levels." More recent evaluations of the NAEP Achievement Levels have recognized their value and have been more positive about the standard-setting methodology used by NAEP.

By the early 2000s, item-mapping approaches (Hambleton & Pitoniak, 2006; Hambleton, 2001; Lewis et al., 1996) were becoming widely used across the country for state-level standard setting. These approaches were particularly appropriate for criterion-based assessments, such as NAEP and other content standards-based state assessment programs typical under the No Child Left Behind Act of 2001. Results of the 2005 grade 12 NAEP mathematics standard-setting process concluded that current validation processes reflected strong evidence to support the validity of the process of setting achievement level standards: "the procedure was sound, followed recommendations for best practices in the area of standard setting, and involved multiple quality control checks to support the defensibility of the process" (Sireci et al., 2009, pp. 339). In particular, the report focused on the importance of the intended purposes and uses of the NAEP ALDs.

By 2016, there was evidence of broad acceptance of NAEP Achievement Levels and their descriptions. An evaluation report completed by NASEM (2017) concluded that the NAEP Achievement Levels had utility and offered recommendations to enhance the validity of the achievement levels, including as a first recommendation to examine the alignment of assessment content with the ALDs to provide sufficient evidence to inform the removal of the trial status. With a focus on the NAEP Mathematics and Reading achievement levels for grades 4, 8, and 12, the report stated that

> during their 24 years [the achievement levels] have acquired meaning for NAEP's various audiences and stakeholders; they serve as stable benchmarks for monitoring achievement trends, and they are widely used to inform public discourse and policy decisions. Users regard them as a regular, permanent feature of the NAEP reports. (p. ix).

However, the National Academies' report documented reasons to endorse conclusions from Buckendahl and colleagues (2009) that

> given the importance of a highly visible national assessment program, it is essential that a validity framework be created to coordinate a program of validity research on NAEP, aimed at informing the validity of score interpretation and use. This should be a highlighted component of NAEP. (NASEM, 2017, p. 1-96)

A primary focus of the report's recommendations was the need for improvement in the guidance provided to users in interpreting and using NAEP Achievement Level reports.

While the committee of authors found that the achievement levels were widely disseminated to and used by many audiences, more interpretive guidance about the meaning and appropriate uses of those levels needed to be provided to users. "Without appropriate guidance, misuses are likely" (NASEM, 2017, p. 214). More information was needed about the intended interpretations and uses of the achievement levels and about the validity evidence that supported these interpretations and uses. Further, the report called out the need for "information on the actual interpretations and uses commonly made by the [NAEP's] various audiences" (p. 214), which could include appropriate and inappropriate interpretations and uses. Finally, the committee of authors concluded that NAEP needed to provide evidence to evaluate the validity of the interpretation and uses of any of the achievement levels, including better descriptors of what students at a given level know and can do with "accurate and specific information about the things that students at the cut score for each level know and can do" (p. 214).

In response to the evaluation, the Governing Board (2018) updated their policy on developing student achievement levels for NAEP. The Governing Board's COSDAM led the effort to update the policy and incorporated NASEM recommendations to conduct additional periodic reviews of the ALDs and to address potential misunderstandings and misuses of NAEP Achievement Level labels.

Also, reported in 2022, the Governing Board contracted with Pearson to design and implement a review of the NAEP ALDs in reading and mathematics assessments for grades 4, 8, and 12 (Moyer & Galindo, 2022). Based on earlier pilot studies, the study addressed the Governing Board's updated achievement level policy by developing reporting ALDs to state how the assessment content aligns with the existing content ALDs and achievement level policy definitions (i.e., the broad definitions used to communicate performance standards with broad audiences and across grade levels and content areas).

Methodologies Used to Set NAEP Achievement Levels

Providing information about the standard-setting processes used to develop the cut scores associated with NAEP Achievement Levels is important for informing the first major claim of this report: NAEP Achievement Levels are established based on defensible standard-setting methods that are implemented with fidelity.

Technical developments for NAEP Achievement Level setting have been one key aspect of NAEP's history (Bourque, 2009; Egan et al., 2011). In the early 1990s when the Governing Board first envisioned how to set performance standards, the choice of standard-setting methods was limited (Berk, 1986; Cizek & Bunch, 2007). The early procedures, including the Nedelsky (1954) method, the Ebel (1972) procedure, and the Angoff method (1971), had been used in testing programs, including professional licensure and certifications, and were appropriate to tests with multiple-choice items such as NAEP. The Angoff method, however, was the most researched in the literature at the time (Hambleton & Pitoniak, 2006). In the method, panelists review multiple-choice items, deciding item by item an estimated likelihood that a "minimally competent" student would answer the item correctly. This question is more relevant to licensure and certification exams but is also used in K–12 assessments. The method had additional advantages making it suited to NAEP, including its straightforward procedure and the ability to adapt it to multiple levels and item formats. It was therefore chosen as a methodology for NAEP standard setting with experts' support. Panels of educators and other stakeholders examined items and the specific item content to make judgments about the probability of the students who are just reaching proficiency or advanced levels answering an item correctly.

Modified Angoff

In the 1990s, the Governing Board contracted with ACT in Iowa City to implement the standard settings (Loomis & Bourque, 2001; Reckase, 2000), first with mathematics in 1992. Consistent with the Governing Board's policy, ACT was responsible for developing ALDs; convening national samples of participants for grade 4, 8, and 12 panels; implementing pilot and research studies; conducting the standard-setting meetings; providing recommendations to the Governing Board; and producing all process and technical reports. Then, between the 1992 and 1998 cycles, ACT developed standards for seven NAEP subjects: mathematics, reading, science, writing, civics, U.S. history, and geography. All seven used a modified Angoff (1971) procedure to develop the achievement levels that was eventually modified to the extent that it became known as the ACT/NAEP method.

During this period, ACT technical staff, with the advice of their Technical Advisory Committee on Standard Setting, expanded and refined the standard-setting process in many ways, including training approaches and principles. A critical part of the standard-setting training was the fact that panelists were trained and encouraged to internalize every aspect of the NAEP assessments, including the NAEP assessment framework, performance descriptors, ALDs, and item formats being used in a particular assessment, before moving on to the standard-setting task. These training activities aligned the procedural evidence with the criterion-referenced purpose of NAEP. In addition, panelists received systematic feedback during the standard-setting process so they could see the consequences of recommended cut scores in terms of the impact, including the percent of students performing at each achievement level based on where they placed the cut scores. "The standard-setting process was viewed not as providing simply a professional opinion about the standards, but rather one's professional judgment about the appropriate standards" (Bourque, 2009, p. 22). At the time, the notion of feedback to panelists

was novel in most standard settings, and NAEP used inter- and intra-rater location data, rater consistency feedback, empirical data such as p-values, whole booklet feedback, and other useful reality checks for panelists to ground themselves. In addition, panelists participated in a series of self-reported questionnaires to monitor the process through their individual experiences, to improve future processes, and to determine whether the panelists felt confident in the work they had completed. These process evaluations added to the evidence of procedural validity (Kane, 1994). Standardized and comprehensive training, extensive use of feedback, and formal process evaluations were all modifications ACT/NAEP incorporated in the original Angoff method.

Bookmark Item Mapping Methods

Starting with standard setting for grade 12 mathematics in 2005, a new method known as the Mapmark standard-setting procedure (Schulz & Mitzel, 2011) was developed and implemented, using a bookmark item mapping method (Hambleton & Pitoniak, 2006). The method's elements had been tried in states under the No Child Left Behind Act of 2001 and further developed for NAEP. Bookmark methods use spatially representative *item map* displays and holistic feedback to inform panelists as they make judgments about where cut scores should be placed. The item map presents a list of items in order of their difficulty across the range of difficulty of the test. It can be mapped to an *ordered item booklet* with items presented in the same order (Figure 2-2).

Figure 2-2
Illustration of an Ordered Item Booklet Reflecting an Item Map

The use of difficulty-ordered items and holistic feedback to panelists were novel features of the method at the time. The method leveraged advantages of item response theory used for developing the test scales and the most current developments in domain score theory and technology. Item response theory allowed for the creation of item maps based on empirical difficulty and for the use of a response probability criterion that also corresponded to the content domains, areas of knowledge, and skills.

A primary advantage of the Mapmark approach (Schulz & Mitzel, 2011) was how it was theoretically founded on panelists' understanding of performance standards. The process was based on building panelists' understanding of how student achievement increases as a sequential mastery of knowledge and skills linking performance standards to the test content.

As panelists made judgments, they focused on what the item measured and why the item was more difficult than the items that preceded it.

Panelists' process evaluation questionnaires showed that they understood the concepts and tasks of the method, that they were confident in their cut score recommendations, and that they believed that the process allowed them to use their best judgment. These results suggested that the cut scores and the achievement level percentages associated with them may be more generally perceived as reasonable. In general, experts agreed that the "Mapmark component of the standard-setting process conducted for [the National Assessment Governing Board] contributed positively to the overall procedural validity of the process" (Schulz & Mitzel, 2011, p. 60).

A particularly important aspect of the advantages of bookmark approaches for setting cut scores is the consistency with criterion-referenced meaning and the minimizing of norm-referenced meaning of panelists' individual cut scores. Panelists are trained to focus first on the content domain, emphasizing the knowledge and skills expected around the cut scores (at the borderline or threshold) before they entertain feedback about other panelists' decisions and the impact of their judgments on the cut scores using range and median cuts. This keeps the value on the criterion-referenced nature of performance standards.

Body of Work

In 2011, an innovative approach was employed for the writing assessments in grades 8 and 12: the Body of Work methodology (Bay et al., 2012). This method belongs to the holistic family of standard-setting methods in which panelists review a series of examinee work samples and assign each sample to one of several performance categories (Hambleton & Pitoniak, 2006). Kingston and colleagues (2001) had argued that the method was the most appropriate for writing assessments because it was developed specifically for use with performance assessments that measure student achievement using open-response items (Kahl et al., 1995).

The method is typically implemented in two stages: rangefinding and pinpointing. During the rangefinding stage, panelists review a set of scored work performances that span the entire range of performance. Panelists do not see the scores themselves, so they do not bias their classifications during the rating phase. Panelists then classify those work performances into achievement level categories. These classifications are then used to compute cut scores. Then, during the pinpointing stage, panelists are provided a set of work performances that have scores in the vicinity of the cut scores as determined during the rangefinding stage. For each cut score, panelists rate each work performance as below, at, or above the achievement level.

Achievement Levels by Content Area

The achievement levels currently in use and specific to each content area and NAEP grade levels are summarized here (NCES, 2012a), as demonstrated in Table 2-1.

Table 2-1
Achievement Levels

Year	Subject Area	Purpose	Method	Reference
1992	Mathematics	Set achievement levels	Modified Angoff	Johnson (1992)
1992	Reading	Set achievement levels	Modified Angoff	Mullis (1993)
1994	Reading	Validate achievement levels*	Modified Angoff	Allen, Carlson, & Zeklenak (1999)

Year	Subject Area	Purpose	Method	Reference
1994	U.S. History	Set achievement levels	Modified Angoff	Beatty et al. (1994)
1994	Geography	Set achievement levels	Modified Angoff	Williams et al. (1995)
1996	Mathematics	Validate achievement levels*	Modified Angoff	Allen, Carlson, & Zeklenak (1999)
1998	Civics	Set achievement levels	Modified Angoff	Lutkus et al. (1999)
1998	Writing, Grade 4	Set achievement levels	Modified Angoff	Loomis & Hanick (2000)
2005	Mathematics, Grade 12	Set achievement levels	Modified Bookmark	Pitoniak et al. (2010)
2007	Economics, Grade 12	Set achievement levels	Modified Bookmark	NCES (2012b)
2009	Science	Set achievement levels**	Modified Bookmark	NCES (2012b)
2011	Writing, Grade 8 & 12	Set achievement levels	Body of Work	Bay et al. (2012)
2014	Technology and Engineering Literacy, Grade 8	Set achievement levels	Modified Bookmark	Nebelsick-Gullet & Fitzpatrick (2016)

Notes. *The validation study also corrected a weighting error. **The Governing Board first developed science ALDs in 1996.

The current scale scores required to reach each NAEP Achievement Level for NAEP Reading and Mathematics are presented in Table 2-2. Note that NAEP is not on a vertical scale and so each set of cut points should be interpreted independently for each grade.

Table 2-2
NAEP Achievement Level Cut Scores, by Grade and Content Area

Assessment	NAEP Basic	NAEP Proficient	NAEP Advanced
Math Grade 4	214	249	282
Math Grade 8	262	299	333
Math Grade 12	141	176	216
Reading Grade 4	208	238	268
Reading Grade 8	243	281	323
Reading Grade 12	265	302	346

Major Claims

Considering the Governing Board's (2018) revised policy and program goals, the remainder of this report summarizes validity evidence in support of the validity argument with accumulated evidence in support of the NAEP Achievement Levels. Claims made about the assessment results are specific to each content area and describe what students at a given level know and can do with "accurate and specific information about the things that students at the cut score for each level know and can do" (NASEM, 2017, p. 214).

Broadly, the mathematics assessment claims describe knowledge of basic mathematical facts, ability to carry out computations using paper and pencil, knowledge of basic measurement formulas in geometric settings, and application of mathematics to daily living skills involving time and money. Reading assessment claims describe reading comprehension, vocabulary knowledge, literary experience, and text types, reading strategies, and purposes (NCES, 2024d).

The knowledge and skills are further defined in the ALDs. These should be referred to when seeking to understand the knowledge and skills to define student performance by NAEP Achievement Level at each grade level and content area. The accuracy and importance of the ALDs can be evaluated based on the procedural, internal, and external evidence presented throughout this report. The achievement levels are sometimes used or interpreted incorrectly, in ways not supported by validity evidence. An example of common inappropriate uses include references to the *NAEP Proficient* level as being equivalent to "grade level." Another example is blanket statements, such as "Students who do not reach the *NAEP Basic* level or *NAEP Proficient* level cannot read" or "cannot do math," which reflect an erroneous and inappropriate interpretation of the achievement levels.

The remainder of this report seeks to describe the evidence to support or not support appropriate and inappropriate uses of the NAEP Achievement Levels. By synthesizing this information in one report, greater clarity will be provided regarding what has and has not been done.

III. Achievement Level Development Policy and Process

Having described the history and purpose of NAEP and the NAEP Achievement Levels in Chapter II, Chapter III describes the Governing Board's Achievement Level Development Policy and demonstrates the policy's adherence to best practices in the field of educational measurement. The chapter goes on to describe the system of NAEP Achievement Levels, focusing on the NAEP content ALDs and reporting ALDs, which are based on this policy and are intended to provide a basis for the alignment between the NAEP scores and their interpretation and use. This information is intended to provide further evidence in support the first major claim, that NAEP Achievement Levels are established based on defensible standard-setting methods and are implemented with fidelity. Information in this section also addresses the second major claim, that NAEP ALDs are defensible definitions of what students know and can do at each level.

The Governing Board sets policy for developing student achievement levels for NAEP. This policy sets expectations for "comprehensive, inclusive, and deliberative processes" (Governing Board, 2018, p. 1) and serves to guide decisions related to the development and review of the NAEP Achievement Levels. As described in Chapter II, the Governing Board first established the NAEP Achievement Levels policy in 1990. At that time, reporting included the percentage of test takers at each defined level and those falling below the *NAEP Basic* level. The NAEP congressional reauthorization in 1994 made new stipulation that the NAEP Achievement Levels be designated as on a trial basis until the NCES commissioner could consider evaluation results and determine that the NAEP Achievement Levels were reasonable, reliable, valid, and informative to the public. In 2001, the Governing Board set revised policy that included the establishment of policy descriptions of each performance level.

The 2017, the NASEM report included in its recommendations that alignment "among the frameworks, the item pools, the achievement level descriptors, and the cut scores is fundamental to the validity of inferences about student achievement" (p. 245). Based on alignment evaluations from 2009, the Governing Board acknowledged the need for additional work to ensure adequate alignment verification of specific NAEP assessments. These assessments included grade 4 and grade 8 mathematics assessments, as well as grade 4 reading and grade 12 mathematics. In response to the report, the most current policy (Governing Board, 2018) established guidelines to ensure that NAEP Achievement Levels balance consistency with ongoing evaluation and that they establish an effective and defensible working definition of student performance. The policy was intended to guide development and validation of the NAEP Achievement Levels in support of an ultimate goal of reporting results that are reasonable, useful, and informative to the public.

The Governing Board, through its COSDAM, is charged with monitoring the development and review of the NAEP Achievement Levels to ensure that all adopted ALDs, cut scores, and exemplars comply with all principles of this policy. The principles of the policy are as follows:

- Principle 1 – Elements of Achievement Levels
- Principle 2 – Development of Achievement Level Recommendations
- Principle 3 – Validation and Reporting of Achievement Level Results
- Principle 4 – Periodic Review of Achievement Levels
- Principle 5 – Stakeholder Input

- Principle 6 – Role of the Governing Board

The policy includes the expectation that eligible contractors would be selected through a competitive bidding process. These contractors carry out the development and validation of the NAEP Achievement Levels' elements and are managed in a technically sound, efficient, cost-effective, and timely manner. In the next section, we demonstrate how the policy links to standards for best practice.

Adherence to Best Practices for Testing and Measurement

One aspect of the validation process is to inspect how the policy used to guide the NAEP Achievement Levels' development aligns with best practice in the field of educational measurement. The Governing Board's (2018) policy is organized around the six principles listed above that relate directly to standards for best practice in educational measurement (AERA et al., 2014; Brennan, 2006; Joint Committee on Testing Practices, 2004; NCES, 2012a). In general, these best-practice, industry standards were developed through peer-reviewed research and broad consensus in the field of educational measurement. While all are used by test publishers broadly, the *Standards for Psychological and Educational Testing* (hereafter called the *Standards*; AERA et al., 2014) provide the most comprehensive set of criteria for sound testing practices. The *Standards* aim "to provide a basis for evaluating the quality of those practices" (AERA et al., 2014, p. 1). The criteria detailed in the *Standards* are used by the professionals who specify, develop, or select tests as well as by those who interpret or evaluate test results, including state and federal governments and test development organizations. The *Standards* make clear that a "critical step in the development and use of some tests is to establish one or more cut scores dividing the score range to partition the distribution of scores into categories" (p. 100).

The six principles of the Governing Board's (2018) policy relate to the *Standards* (AERA et al., 2014) broadly since the NAEP Achievement Levels are the basis for the intended interpretations of NAEP scores. The NAEP Achievement Levels relate to foundational topics (i.e., validity, reliability, fairness), each constituting a chapter of the *Standards*. Each topic presents overarching standards, which serve as broad guiding principles, as well as more specific standards that address thematic clusters of related topics, including all operational, procedural, and evaluative guidance. Since validity is dependent on the assessment being defensibly reliable and fair, all evidence related to the reliability or internal consistency (Chapter 2) as well as all evidence of fairness for all examinees in the tested population (Chapter 3) are ostensibly included in the validity argument.

Standard 2.0 of the *Standards* relates to the topic of reliability, especially as it focuses on precision "when the consequences of decision and interpretations grow in importance" (AERA et al., 2014, p. 33). Standard 2.0 states that "appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use" (p. 42). Consistent with Kane (1994), internal validity evidence requires evidence of the reliability of scores, including the consistency of the interpretation of those scores by different groups of people. Though the focus of this report is not on the reliability of the scale scores that underlie the achievement levels, it is an important aspect of ensuring validity, and more information about how NAEP ensures reliability can be found in its technical documentation.[1]

[1] https://nces.ed.gov/nationsreportcard/tdw/sitemap.aspx

Standard 3.0 states that all steps should "promote valid score interpretations for the intended uses for all examinees in the intended population" (AERA et al., 2014, p. 63). Since the interpretation and use of the NAEP Achievement Levels falls within the testing process, evidence of fairness is important to an overall assessment of validity and must demonstrate valid interpretations for all subgroups, including socioeconomic, cultural, ethnic, racial, and gender subgroups. This fairness argumentation therefore requires evidence that demonstrates how the experience with the test is fair for all subgroups and that all known sources of undue barrier or lack of access have been removed.

Also of importance for the discussion of the validity of NAEP Achievement Levels are those standards that more specifically guide the operational tasks of setting cut scores and the interpretation and use of reported scores, which is discussed here. The sections that follow demonstrate how each of the Governing Board's policy principles that guide the NAEP Achievement Level practices relate to specific references from the *Standards*.

Principle 1 – Elements of Achievement Levels

The first principle described in the Governing Board's (2018) policy describes the three elements of the NAEP Achievement Levels, consisting "of content achievement level descriptions (ALDs), cut scores that demarcate adjacent levels, and exemplar items or tasks that illustrate performance at each level" (p. 5). In addition to aspects of the overarching foundational standards in Chapters 1–3, key criteria that address the importance of the elements of NAEP Achievement Levels are emphasized in Standards 5.1, 5.5, and 5.21 (AERA et al., 2014). Together, these standards call for clear expectations and rationale for intended interpretations of scores.

Standard 5.1 addresses expectations, stating that "test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations" (AERA et al., 2014, p. 102). Criterion-referenced assessments such as NAEP evaluate how well a student meets established criteria or standards. Speaking specifically to criterion-referenced interpretations, Standard 5.5 emphasizes that, when scale scores are "designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be explained clearly" (p. 103). Finally, and related specifically to cut scores, Standard 5.21 states that "when proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly" (p. 107). Together these standards serve to justify the Governing Board's definition and organization of the achievement levels.

Principle 2 – Development of Achievement Level Recommendations

The second principle described in the Governing Board's (2018) policy addresses the numerous technical requirements of the development of NAEP Achievement Levels, "consistent with relevant widely accepted professional assessment standards, [and] based on the appropriate level of subject matter knowledge" (p. 5). Again, broadly, all evidence contributing to the validity argument, including reliability and fairness clusters, are relevant to Principle 2, including Standard 3.6, which clarifies that "test developers are responsible for examining evidence for validity of score interpretations for intended uses for individuals from those subgroups" (p. 65). For example, Principle 2 includes subsection (c), which requires that standard-setting panels "reflect diversity in terms of gender, race/ethnicity, region of the country, urbanicity, and experience with students with disabilities and English language learners" (p. 6).

Standard 5.2 and 5.22 emphasize the need for defensible procedures, especially given the judgmental nature of the standard-setting process. Along with describing the procedures used for constructing scales that are used for reporting scores, "the rationale for these procedures should be described clearly" (Standard 5.2; AERA et al., 2014, p. 102). Standard 5.22 underscores the need for procedural evidence when cut scores defining proficiency levels "are based on direct judgments about the adequacy of item or test performances" (p. 108); "the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way" (p. 108). Both of these standards align with the details of Principle 2, including a design document, panel involvement, process and evaluation steps, technical advisory committee, and documentation.

Principle 3 – Validation and Reporting of Achievement Level Results

The third principle described in the Governing Board's (2018) policy addresses the validation and reporting of the NAEP Achievement Levels. "The achievement level setting process shall produce results that have validity evidence for the intended uses and interpretations and are informative to policy makers, educators, and the public" (p. 8). Validity is the key concept in educational measurement, unifying all foundational topics and the development, administration, and reporting activities in testing. The validation process, including building and evaluating a validity argument, as described in Chapter 1 of the Standards, is at the heart of Principle 3.

Standard 1.0 states that "clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (AERA et al., 2014, p. 23). The NAEP Achievement Levels provide this clear articulation, especially as a comprehensive system of ALDs that are related and aligned but used for differing purposes. "A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation" (Standard 1.2; p. 23). In addition to Standard 1.0, Standards 2.0 (reliability) and 3.0 (fairness) are critically linked to the validation and reporting of the NAEP Achievement Levels. The third principle also addresses the potential for inconsistencies in score interpretation and the need for transparency:

If validity for some common or likely interpretation for a given use has not been evaluated, or if such an interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations. (Standard 1.3; p. 23)

In addition, Chapters 2 (reliability) and 3 (fairness) relate to Principle 3. Together, these foundational topics set forth guiding principles for the evidentiary basis of the validity argument that are complemented by more specific standards. The reliability of score interpretations across different groups of people relates to Standard 2.16, requiring evidence that test takers would be classified in the same way in replication. Cluster 4 of Chapter 3 (fairness) requires "safeguards against inappropriate score interpretations for intended uses" (p. 70), including the demand that test developers and publishers provide information "to support appropriate test score interpretations for their intended uses for individuals from these subgroups" (Standard 3.15; p. 70).

In addition, Standard 5.3 is relevant to the reporting of scores with published results and the potential for misinterpretation of scores. "If there is sound reason to believe that specific misinterpretations of a score scale are likely, test users should be explicitly cautioned" (p. 102).

Finally, Standard 12.1 addresses the specific duty of educational testing programs to be responsible for clear guidance on how scores should be interpreted and used:

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described by those who mandate the tests. It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test developer and/or user. (p. 195)

Overall, Principle 3 demonstrates the Governing Board's responsibility to develop and oversee validity argumentation for the use and interpretation of the NAEP Achievement Levels.

Principle 4 – Periodic Review of Achievement Levels

The fourth principle described in the Governing Board's (2018) policy establishes the importance of ongoing monitoring of the alignment of NAEP Achievement Levels and current conditions of education and student achievement over time. "Periodic reviews of existing achievement levels shall determine whether new achievement level descriptions and/or cut scores are needed to continue valid and reliable measurement of current student performance and trends over time" (p. 9).

Standards 5.6, 7.14, and 9.7 (AERA et al., 2014) address different aspects of ongoing review and willingness to revise NAEP Achievement Levels based on documented evidence. Standard 5.6 requires testing programs that maintain a common scale "conduct periodic checks of the stability of the scale on which scores are reported" (p. 103). In this way, the achievement levels serve as a content-based anchor for the meaning of the scores. Also, the *Standards* address situations when substantial changes are made to a test, as has happened in the history of NAEP. In such cases, "the test's documentation should be amended, supplemented, or revised to keep information for users current and to provide useful additional information or cautions" (Standard 7.14; p. 129). Finally, test users themselves have a responsibility to "verify periodically that their interpretations of test data continue to be appropriate given any significant changes in the population of test takers, the mode(s) of test administration, or the purposes in testing" (Standard 9.7; p. 144).

Principle 5 – Stakeholder Input

Over time, the Governing Board has demonstrated strong commitment to the inclusion of stakeholders in the developmental processes and evaluative activities of NAEP assessments (Bourque, 2009; Bourque & Byrd, 2000; Moyer & Galindo, 2022). "The process of developing student achievement levels is a widely inclusive activity. The Governing Board shall provide opportunities to engage multiple stakeholders throughout the achievement level setting process and shall strive to maximize transparency of the process" (Governing Board, 2018, p. 10).

Principle 5 includes subsection (c), which requires that all stakeholders are represented in the development of achievement levels:

Achievement level setting panelists shall include teachers, non-teacher educators, and other interested members of the general public with relevant educational background and experience, including parents, researchers, and employers. Each panel shall reflect diversity in terms of gender, race/ethnicity, region of the country, urbanicity, and experience with students with disabilities and English language learners. (Governing Board, 2018, p. 10)

The commitment to the inclusion of stakeholders in the development of the NAEP Achievement Levels is also consistent with Standard 5.22 (AERA et al., 2014), with emphasis on how participants in standard-setting processes bring their range of knowledge and experience to bear when making judgments about cut scores and ALDs. Also, the policy's Principle 5 emphasizes the value of inclusion of stakeholders, consistent with the *Standards*' (AERA et al., 2014) broad emphasis on fairness (Standard 3.0). In sum, the Governing Board's Principle 5 is consistent with the *Standards'* explicit expectations for stakeholder involvement.

Principle 6 – Role of the Governing Board

The Governing Board demonstrates its commitment to best practices by maintaining and upholding policy consistent with the overall intent of the *Standards*:

Educational . . . testing and assessment are among the most important contributions of cognitive and behavioral sciences to our society, providing fundamental and significant sources of information about individuals and groups. . . . Well-constructed tests that are valid for their intended purposes have the potential to provide substantial benefits for test takers and test users. Their proper use can result in better decisions about individuals and programs than would result without their use and can also provide a route to broader and more equitable access to education and employment. The improper use of tests, on the other hand, can cause considerable harm to test takers and other parties affected by test-based decisions. (AERA et al., 2014, p. 1)

The Governing Board monitors "the development and review of student achievement levels to ensure that the final ALDs, cut scores, and exemplars recommended to the Governing Board" (Governing Board, 2018, p. 1). In sum, this process is ongoing, iterative, and reflective of standards for best practice in the field of educational measurement.

Achievement Level Descriptions

The NAEP Achievement Levels, as extensions of scores, are communicated to users in various ways to guide their intended uses and interpretations. Aligned with the NAEP policy descriptions discussed in Chapter II, the NAEP ALDs create a system of descriptors (Egan et al., 2011) that provide this interpretive guidance to content developers, users, and stakeholders. Along with the policy descriptions, this system includes content ALDs and reporting ALDs discussed in this chapter.

The NAEP frameworks (NCES, 2024c) report all elements of the NAEP Achievement Levels along with the blueprints for the content and design of each NAEP assessment. In accordance with policy, the Governing Board (2018) works with a committee of subject matter experts, practitioners, and members of the general public to develop the NAEP Achievement Levels.

In order to measure trends in student performance, NAEP frameworks are designed to remain stable for as long as possible. At the same time, all frameworks are responsive to changes in national and international standards and curricula. Without advocating any particular approach to instruction, NAEP frameworks provide a starting point for constructive conversations about high-quality educational standards and assessments. (NCES, 2024c)

In the framework documents, appendices include the content ALDs, including the label, cut scores associated with the level, and more detailed reporting ALDs. The 2022 and 2024 NAEP frameworks have been relatively consistent over time.

- The Governing Board (2022a) noted that the 2022 and 2024 NAEP Mathematics Assessment Framework is the same framework as that developed for the 1992 NAEP Mathematics assessment for grades 4 and 8, with minor modifications to clarify assessment objectives. It is the same framework developed for the 2005 grade 12 assessment and includes 2009 modifications to support NAEP reporting on academic preparedness for postsecondary endeavors.
- The 2022 and 2024 NAEP Reading Assessment Framework (Governing Board, 2022b) is the same framework first developed for the 2009 NAEP Reading assessment, which includes 2009 modifications for grade 12 to support NAEP reporting on academic preparedness for postsecondary endeavors.

Content Achievement Level Descriptions

The first principle of the Governing Board's (2018) policy, Elements of Achievement Levels, includes the call to develop content ALDs that "translate the policy definitions into specific expectations about student knowledge and skills in a particular content area, at each achievement level, for each subject and grade" (p. 5). These content ALDs should provide descriptions of specific expected knowledge and skills of students performing at each achievement level and "reflect the range of performance that items and tasks should measure" (p. 5). The policy clarifies that the ALDs aim to ensure consistency and specificity in the interpretations of policy definitions for each assessment for standard-setting panelists.

To illustrate, Table 3-1 provides examples of the content ALDs at the fourth-grade level for mathematics and reading. The cut score demarcating the lower end of the score range is noted in parentheses. Each description illustrates the knowledge and skills students must demonstrate for the specified NAEP Achievement Level.

Table 3-1
Example Content Achievement Level Descriptions for NAEP

Content Area	NAEP Achievement Level (Cut Score)	Content ALD
Mathematics	NAEP Basic (214) *	Fourth-grade students performing at the *NAEP Basic* level should show evidence of understanding the mathematical concepts and procedures. 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—although not always accurately—four-function calculators, rulers, and geometric shapes. Their written responses are often minimal and presented without supporting information.

Content Area	NAEP Achievement Level (Cut Score)	Content ALD
Reading	*NAEP Basic* (208) **	Fourth-grade students performing at the *NAEP Basic* level should be able to locate relevant information, make simple inferences, and use their understanding of the text to identify details that support a given interpretation or conclusion. Students should be able to interpret the meaning of a word as it is used in the text. When reading literary texts such as fiction, poetry, and literary nonfiction, fourth-grade students performing at the *Basic* level should be able to make simple inferences about characters, events, plot, and setting. They should be able to identify a problem in a story and relevant information that supports an interpretation of a text. When reading informational texts such as articles and excerpts from books, fourth-grade students performing at the *Basic* level should be able to identify the main purpose and an explicitly stated main idea, as well as gather information from various parts of a text to provide supporting information.

Notes. * Governing Board, 2022a, p. 71; ** Governing Board, 2022b, p. 64

Reporting Achievement Level Descriptions

Distinct from but aligned with the content ALDs, reporting ALDs explain what students show they can likely do based on actual NAEP performance (Governing Board, 2022a, 2022b). These ALDs were developed in response to the first NASEM recommendation that recommended evaluating the alignment of the assessments back to the content ALDs included in the NAEP frameworks. They are intended for use with NAEP score results, and they go into detail on the specific knowledge and skills students can likely demonstrate when they perform at a given level. The reporting ALDs are intended to be updated over time to maintain accurate representations of student performance.

The NAEP reporting ALDs have been developed by educators and content experts based on a review of assessment items assigned to each NAEP Achievement Level. The reporting ALDs intended for first use with the 2022 reporting cycle were developed using a methodology (Moyer & Galindo, 2022) similar to the approach used for alignment evaluation and revision of the 2009 NAEP reading ALDs for grades 4, 8, and 12 (Donahue et al., 2010) and the 2009 NAEP mathematics ALDs for grade 12 (Pitoniak et al., 2010). With oversight from COSDAM from the outset, a technical advisory committee with six experts in achievement levels and ALDs provided in-depth technical guidance on all phases of the work. Panelist recruitment involved multiple steps designed to obtain broadly representative, well-qualified panelists familiar with the knowledge and skills needed by students in grades 4, 8, or 12 in mathematics or reading as well as the full diversity and educational needs of the student population. In the development of these ALDs, “items are assigned to an achievement level based on the likelihood that most students (67 percent) within a level can answer the item correctly” (Governing Board, 2022a, p. 77), thus maximizing alignment of score interpretations, test content, and student performance.

The ALDs, as a system, function to align content specifications, assessment items, score results, and guidance for the interpretation and use of scores (Egan et al., 2011). Their use to assess achievement level description validity is described in Chapter IV. A comparison of the reporting ALDs to the content ALDs found that qualified panelists observed “alignment between the knowledge and skills students demonstrated in an achievement level, described by the

summary statements, and the expected knowledge and skills for an achievement level, described by the content ALDs in the framework" (Moyer & Galindo, 2022, p. 2). Table 3-2 illustrates the NAEP reporting ALDs for grade 4 *NAEP Basic* level for both mathematics and reading.

Table 3-2
Example Reporting Achievement Level Descriptions for NAEP

Content Area	Achievement Level	Reporting ALDs
Mathematics	*NAEP Basic**	Students performing at the *NAEP Basic* achievement level can likely • determine place value of whole numbers up to hundred thousands; • locate whole numbers on a number line; • read, write, compose, and decompose multi-digit whole numbers in a variety of forms based on place value; • identify even and odd numbers and understand factors; • add and subtract multi-digit whole numbers with single-step and/or regrouping; • add and subtract decimals to the hundredths place; and • understand inverse operations and their properties and apply concepts of multiplication. Students performing at the *NAEP Basic* achievement level can likely • identify appropriate measurement tools in real-world scenarios; • measure or estimate lengths of objects in standard and non-standard units; and • find the perimeter of polygons given a visual aid. Students performing at the *NAEP Basic* achievement level can likely • identify lines of symmetry; • identify attributes of polygons as well as 3D shapes; and • compare these attributes with the support of visual aids. Students performing at the *NAEP Basic* achievement level can likely • correlate information between tables and data displays; and • read and interpret tables and scaled graphs.

Content Area	Achievement Level	Reporting ALDs
Reading	*NAEP Basic***	When reading literary texts such as fiction, poetry, and literary nonfiction, fourth-grade students performing at the *NAEP Basic* level can likely • determine the relevant meaning of familiar words using context within the same sentence or paragraph; • identify a specific detail to make a simple inference about the characters' actions, motivations, or feelings using a single point or multiple points in the text if they are in close proximity; • sequence or categorize events from the story; • make a general reference to an appropriate section of the text or provide some support for ideas related to the plot or characters; • find meaning or provide evidence from one of the texts when making a comparison across texts; • identify explicit details from the text; and • state an opinion with general support from one section of the text. When reading informational texts such as articles and excerpts from books, fourth-grade students performing at the *NAEP Basic* level can likely • determine the relevant meaning of familiar words using context from a single section of the text; • locate a specific detail from the text and make simple inferences from one section of the text; • restate a problem or solution presented in a single section of the text; • provide a description of a text feature or the author's craft using a general reference to the text; and • provide an opinion using a general reference to the text.

Notes. * Governing Board, 2022a, p. 77; ** Governing Board, 2022b, p. 69

IV. Validity Research

Chapter III described how the Governing Board's Achievement Level Development Policy and the system of NAEP Achievement Levels directs the development and evaluation of the NAEP content and reporting ALDs. Chapter IV focuses on evidence of the validity argument for the NAEP Achievement Levels. NASEM (2017) recommended the need for continued research that examines the alignment of the *actual* knowledge and skills of students at each NAEP Achievement Level to the assessments themselves, with particular focus on mathematics and reading. In addition, validity arguments require evidence of the quality of the procedures used to develop the achievement levels and how the achievement levels relate in meaningful ways to evidence external to the assessment. Therefore, this chapter examines evidence related to the claims that 1) NAEP Achievement Levels are established based on defensible standard-setting methods that are implemented with fidelity, 2) NAEP Achievement Levels are definitions of what students know and can do at each level, and 3) NAEP Achievement Levels meaningfully relate to other measures of student achievement and other indicators of educational outcomes for all students. The chapter examines different sources of validity evidence to support these claims, including anchoring and alignment studies and external linking studies.

Various types of validity evidence are necessary to support the interpretation of achievement levels and make a convincing case for the interpretation and use argument (Kane, 2013). Kane (1994) provided guidance for establishing the validity of performance standards for high-stakes achievement tests, addressing the question of how to determine whether a score's interpretation is arbitrary. He described the performance standard as the conceptual version of a desired level of competence and the score itself as the operational version of that score. In this sense, the NAEP Achievement Levels are related to the ALDs, which establish that desired level of performance. The process of validating the meaning of the ALDs and the accuracy of the scores involves various types of evidence that collectively shows that the NAEP Achievement Levels reflect what they intended to mean. Kane described three types of evidence to contribute to the validation process: procedural evidence, internal evidence, and external evidence.

- *Procedural evidence* demonstrates the appropriateness of the procedures used and the quality of those procedures' implementation. This type of evidence is particularly important because it is relatively concrete and widely accepted as a basis for policy decisions. While procedural evidence cannot establish the appropriateness of a score or its performance standard, it can invalidate a score or performance standard. For example, if a method for setting performance standards is implemented without fidelity to the important procedural steps, the resulting performance standards would be in question.
- *Internal evidence* shows the consistency of the various results of a standard-setting or evaluation study. This type of evidence is important because it provides support for the overall validity argument by checking the presumed relationship between the performance standard and the cut score. While internal evidence cannot show that the study was appropriately selected or implemented, it can show that results are replicable, even when judgmental processes are prone to error. It can also show indirect evidence of the integrity of the process of establishing performance standards, including ALDs.
- *External evidence* is based on comparison with external sources of information related in a meaningful way to the expectations captured in the performance standards. These comparisons are rarely exact but are rather rough indications of whether a performance level is "too high, too low, or about right" (Kane, 1994, p. 448). All types of evidence are relevant in validity argumentation for achievement levels.

The research presented in this chapter provides all three types of evidence as described by Kane (1994), each serving the overall validity argument (Kane, 2013) in support of the claims. Because standard-setting procedures and results were discussed in Chapters II and III, this chapter only briefly discusses how standard-setting studies have contributed to the overall validity argument for the NAEP Achievement Levels. Internal evidence, especially in relation to the consistency of score interpretation and use across various stakeholders and score users, is primarily from the *anchor studies* that compare ALDs to actual student performance and *alignment studies* that compare ALDs to frameworks and policy definitions. External evidence of validity is focused on studies that empirically link NAEP scores and achievement levels to different but related *external measures*. These studies look at how the performance of students on the NAEP assessments compares to other assessments and the interpretation of their scores.

Evidence from Standard-Setting Studies

The standard-setting process—or the process of setting cut scores on test scales—in and of itself provides procedural evidence (Kane, 1994) for the validity argument for the NAEP Achievement Levels to primarily support the claim that NAEP achievement levels are established based on defensible standard-setting methods that are implemented with fidelity. Chapter II described the various methodologies used to establish the NAEP Achievement Levels over time, with emphasis on the process of determining cut scores content-based approaches used to set NAEP Achievement Levels have reflected (1) the most appropriate of the available options accepted in the field of educational measurement; (2) the most appropriate fit of each approach to the assessment itself; and (3) the inclusion of recommendations from industry standards (e.g., AERA et al., 2014), technical advisors (i.e., Technical Advisory Committees), external reviewers, and other stakeholders (i.e., public comment opportunities). In each case, adherence to the selected method, implementation of the method conducted with fidelity, and thorough technical documentation provide both procedural and internal evidence of the validity argument.

Evidence from Anchor and Alignment Studies

The validity of the interpretations and uses of the NAEP Achievement Levels depends on the alignment between the *intended* knowledge and skills, as captured in the ALDs, and the *actual* knowledge and skills students demonstrate at each NAEP Achievement Level. Kane (1994) described this as the alignment between "the conceptual version of the desired performance" (p. 426) and the minimally adequate performance for a given purpose. This alignment can be evaluated by comparing the ALDs within the content framework and assessment score data. To examine these comparisons, the Governing Board has called for anchor studies and alignment studies. Both types of studies offer both procedural and internal evidence in support of the validity argument and the claims: (1) NAEP Achievement Levels are established based on defensible standard-setting methods that are implemented with fidelity, (2) NAEP ALDs are defensible definitions of what students know and can do at each level.

Anchor studies evaluate whether there is evidence of a problem with the location of a given cut score on a test scale (Loomis, 2018). They are called anchor studies because they use an anchoring, or item-mapping, methodology that anchors test items to the achievement levels on the test scale. As with standard setting, these approaches use statistical techniques and expert judgments to determine the extent to which there is alignment of ALDs with other important factors related to the validity of the achievement levels.

The first anchor studies were conducted on NAEP in 1984 (NCES, 1993). By applying statistical criteria to locate, or anchor, items to achievement levels, anchor descriptions are developed to describe the knowledge and skills related to scale-score intervals. In other words, "anchor descriptions are a back translation of the ALDs; they describe actual performance within the score range of each level" (Loomis, 2018, p. 27). Integrating statistical information available through item-response theory scaling techniques, the procedure uses a *response probability criterion* (i.e., the likelihood that a student would get the item correct at a given point on the scale) along with other criteria. These may include the expected interval scale score between score points, discrimination criteria to ensure that the difficulty between two items would be large enough to be useful or detectable, and, in some cases, a correction for guessing to modify the response probability based on the number of response options available. The anchor descriptions report *what* students know and can do as opposed to the NAEP content ALDs included in the assessment frameworks, which report what students *should* know and be able to do.

In addition to anchor studies, research has evaluated the alignment of the achievement levels with the content ALDs, the frameworks themselves, and anchor descriptions, all of which have contributed to the validity evidence. Expert panels make judgments regarding the degree of alignment of the ALDs (i.e., strong, moderate, weak, none) with the achievement level policy definitions and the content ALDs (Moyer et al., 2022, 2023). These alignment methods also use an item-mapping approach combined with expert review and deliberation. In addition, public comment and replications of reviews have been used to determine alignment.

Table 4-1 provides a summary of key anchor and alignment studies. The evidence the studies have provided is described and categorized in Kane's (1994) categories of validity evidence. In all cases, these studies provide procedural and internal evidence of validity. In some cases, the study was valuable for how it revealed lack of the validity evidence or contributed to improving the methodology for developing and reporting achievement levels. They are presented in chronological order to illustrate the development of both approaches and quality of outcomes.

Table 4-1
A Summary of Anchor and Alignment Studies

Evidence Source	Brief Description	Contribution to Validity Evidence for the NAEP Achievement Levels
Interpreting NAEP Scales (Phillips et al., 1993)	The NAEP 1992 assessments were administered to nationally representative samples of fourth-, eighth-, and twelfth-grade students attending public and private schools and to state representative public-school samples of fourth graders. The NAEP results were reported using three achievement levels at each grade: *Basic*, *Proficient*, and *Advanced*. Using these results along with 1985 literacy results, Phillips and colleagues (1993) conducted an anchor study. For reading, the criteria applied used a response probability (RP)[2] within a 12.5-point range around the cut (either above or below) to select anchor items. For mathematics, the criteria applied used an RP value of .65, discrimination criteria applied based on the achievement level, and guessing correction adjusted depending on the number of answer options. It pointed to the "need for validity evidence to support the interpretations presented by use of the achievement levels" (p. 84) and supported Congress through NCES to provide for an independent evaluation to start the process and a mechanism to examine evidence from several studies that will help assess the validity of interpretations.	The anchor study established the need for internal evidence for the NAEP Achievement Levels using the anchor study methodology. The method mapped items to the NAEP scales and therefore to NAEP Achievement Levels. It established a precedent for use of the types of criteria used to judge alignment between student performance and the interpretation of achievement levels, including RP, discrimination, and correction for guessing.

[2] The RP is the likelihood that a student would get the item correct at a given point on the scale.

Evidence Source	Brief Description	Contribution to Validity Evidence for the NAEP Achievement Levels
NAEP 1994 Reading Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress and Trial State Assessment (Campbell et al., 1996) *NAEP 1994 Geography Report Card: Findings from the National Assessment of Educational Progress* (Persky et al., 1996) *NAEP 1994 U.S. History Report Card: Findings from the National Assessment of Educational Progress* (Beatty et al., 1996)	The NAEP 1994 assessments were administered to nationally representative samples of fourth-, eighth-, and twelfth-grade students attending public and private schools and to state representative public-school samples of fourth graders. The NAEP results were reported using three achievement levels at each grade: *Basic*, *Proficient*, and *Advanced*. Anchor studies applied criteria including anchoring to 25th, 50th, and 90th percentile scale scores, with an anchor range of ±5 percentile points from these values. Anchor descriptions were developed for use with reporting *what* students know and can do, rather than what students *should* know and be able to do.	The 1994 anchor studies contributed additional internal evidence. They refined anchor criteria and applied the methodology to multiple content areas. The studies demonstrate how NCES continued to develop anchor descriptions for reporting performance in the Nation's Report Card. This collection of studies served to move the method and, by reporting achievement levels without a strong validity basis, provided a basis for the work that followed.

Evidence Source	Brief Description	Contribution to Validity Evidence for the NAEP Achievement Levels
NAEP Reading Revisit: An Evaluation of the 1992 Achievement Levels Descriptions (ACT, 1995)	This alignment study investigated the 1992 NAEP Achievement Levels for reading. Using an item-mapping approach, the study examined the question of reliability of interpretation, asking whether different people would interpret the achievement levels differently. The alignment study investigated questions related to the reliability of interpretation, asking "Would different people interpret the achievement levels differently?" (ACT, 1995). The study used a comparative approach with two methods: item difficulty classification and judgmental item classification. A criterion of RP .50 was applied at different subranges within each achievement level, providing guidance to reviewers for how to interpret "can do" and "can't do" items. Two sets of classifications were compared by computing the number of hits 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, as quoted by Loomis, 2018, p. 7). The study concluded by supporting the use of the 1992 NAEP Achievement Levels for reading.	The alignment study provided important internal evidence by demonstrating that two different methods of review resulted in similar interpretations of the reading achievement levels through item classification, contributing to continuity of score interpretation.
Report on Developing Achievement Level Descriptions for the 1996 NAEP Science Assessment (Bourque, 1999)	The Governing Board conducted a panel review study to assess the results of a 1996 standard setting. An RP value of .65 was used for item mapping. Two expert panels independently wrote "can do" descriptions and collaborated to merge and formulate a single set of anchor descriptions.	The anchor study provided procedural evidence of the Governing Board's responsiveness to findings from earlier studies. Further, it contributed to internal evidence via the additional expert review that informed the interpretation of the NAEP Achievement Levels through the "can do" anchor descriptions.

Evidence Source	Brief Description	Contribution to Validity Evidence for the NAEP Achievement Levels
Report on 2002 Geography Scale-Anchoring Study (Weiss, 2003)	This anchor study used a comparative approach to evaluate the NAEP geography assessment. It involved two panels with different eligibility requirements: the first panel had knowledge of the NAEP geography framework, while the second panel, although highly qualified in geography, lacked experience with the NAEP framework. The study applied an RP criterion of .50 along with a discrimination criterion. The two panels then merged their reviews. The findings demonstrated strong alignment between the results of the 1994 and 2001 studies.	The anchor study provided internal evidence by demonstrating consistency of interpretation across multiple administrations of the NAEP geography assessment and across different qualified panels of geography experts.
Report on the 2003 Mathematics Scale-Anchoring Study (Braswell & Haberstroh, 2004)	In this anchor study of grades 4 and 8, panels compared anchor descriptions between the 1992 item pool and the 2003 item pool to determine whether (1) minor changes to the framework and (2) the release and replacement of items in the pool had impacted anchor descriptions. The study found strong similarities across the two pools in terms of anchor descriptions, alignment between ALDs and the anchor descriptions, and consistency with policy definitions.	The anchor study provided internal evidence for mathematics in grades 4 and 8 by comparing anchor descriptions across administrations. Results showed alignment.
Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Science for Grades Four, Eight, and Twelve: Process Report (ACT, 2010)	A standard-setting study was conducted to address concerns that the grade 4 *NAEP Basic* cut score was too low and potentially inconsistent with the policy definition and the other cut scores. The standard setting used a modified bookmark approach to set standards in 2009. Results were deemed inconclusive. In response, the Governing Board compiled research, adjusted the panel recommendations, and compared the NAEP ALDs for science in 2009 with another anchor study (ACT, 2010), after which the Governing Board approved the cut scores and ALDs.	The anchor study provided procedural evidence of the Governing Board's responsiveness to findings from earlier studies. Further, it contributed to internal evidence through the additional evaluation studies and expert review.

Evidence Source	Brief Description	Contribution to Validity Evidence for the NAEP Achievement Levels
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 (Pitoniak et al., 2010)	This anchor study addressed the NAEP Mathematics assessment for grade 12. Given changes to the framework in algebra II and the elimination of some objectives, the Governing Board needed to investigate whether it could maintain reporting trends. "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" (Loomis, 2018, p. 16). The study applied an RP criterion of .67 and a discrimination criterion. Issues related to the alignment at the *NAEP Proficient* and *Advanced* levels when the anchor descriptions were compared to policy and ALDs precipitated the drafting of the new ALDs. Expert panelists identified themes from the anchor descriptions to be addressed in the ALDs. Starting with the 2005 ALDs, they modified the anchor descriptions to align with the ALDs. After multiple reviews, results were approved by the Governing Board for grade 12 mathematics.	The anchor study provided procedural evidence of the Governing Board's responsiveness to findings from earlier studies. Further, it contributed to internal evidence through the additional evaluation studies and expert review.

Evidence Source	Brief Description	Contribution to Validity Evidence for the NAEP Achievement Levels
2009 Preliminary Reading Study, Grade 4 and 8 (Donahue et al., 2009) *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* (Donahue et al., 2010)	These two studies addressed the NAEP Reading assessment. With the development of a new reading framework first administered in 2009, the Governing Board needed to investigate whether it could maintain reporting trends. The preliminary study applied an RP criterion of .50 and a discrimination criterion. It examined alignment using two presentations of items: one in which items were organized by difficulty and one in which items were presented along with passages. Panelists examined the 2008 trial items, comparing 1992 ALDs to 2002 and 2009 anchor descriptions. The study asked, "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?" (Loomis, 2018, p. 21). The study concluded that the 1992 ALDs could be revised and edited for implementation in the 2009 reading framework. In a second anchor study, item mapping used an RP criterion of .67 and a discrimination criterion. The study also addressed complexity as distinct from difficulty. It used a similar design to the preliminary study regarding item organization. The panelists wrote new ALDs, which were revised based upon public comment. Ultimately, a compilation of anchor and alignment results with additional psychometric evidence (i.e., linking study results) allowed the Governing Board to approve the use of the new ALDs with the existing score scale.	The anchor studies provided procedural evidence of the Governing Board's responsiveness to findings from earlier studies. The two studies provided internal evidence across grades, administrations, and study designs.

Evidence Source	Brief Description	Contribution to Validity Evidence for the NAEP Achievement Levels
Achievement Level Description Review for the National Assessment of Educational Progress Mathematics and Reading Assessments (Moyer & Galindo, 2022) *Achievement Level Description Review for the National Assessment of Educational Progress Grade 8 Science, U.S. History, and Civics Assessments* (Moyer & Galindo, 2023)	The studies reviewed the ALDs for NAEP Reading and Mathematics assessments for grades 4, 8, and 12 and NAEP science, U.S. history, and civics for grade 8. The results provided evidence of alignment between the knowledge and skills students demonstrated in an achievement level, as described by the summary statements, and the expected knowledge and skills for an achievement level, as described by the content ALDs within the framework. This alignment was characterized by moderate or strong alignment in most judgments, with one exception for grade 12 mathematics at the *NAEP Advanced* level.	The studies provided procedural evidence for the reporting ALDs. It also produced internal evidence for the alignment of the reporting ALDs to the NAEP content ALDs and the frameworks.

In summary, anchor and alignment studies provide both procedural and internal evidence to support the claims that (1) NAEP Achievement Levels are established based on defensible standard-setting methods that are implemented with fidelity and (2) NAEP ALDs are defensible definitions of what students know and can do at each level. From a procedural standpoint, the studies in Table 4-1 generally support the claim that the Governing Board has met the expectations of its policy related to the appropriateness of methods and the demonstration of sound implementation. Looking across studies, there is evidence of the Governing Board's responsiveness to individual study findings and outcomes. The anchor and alignment studies also provide evidence of the relative consistency of interpretation of the NAEP Achievement Levels by content area assessment, grade level, and administrations. In general, the Governing Board has been responsive to concerns regarding internal evidence of validity that have been identified through these studies over time.

Evidence from Linking and Mapping Studies

Linking studies that relate NAEP Achievement Levels to external measures of academic success and outcomes constitute external evidence for the validity of interpretations and uses (Kane, 1994). These studies support the claim that NAEP Achievement Levels meaningfully relate to other measures of student achievement and other indicators of educational outcomes for all students. While there are various methodological approaches to accomplishing a linking study (e.g., equating, calibration, projection) (Kolen & Brennan, 2014; Mislevy, 1992), the goal of a linking study is to establish a statistical connection between two test scales (or a test scale to a non-test measure) so they can be expressed on the same scale. This allows for meaningful comparison between two different scales.

As external evidence, linking studies may come from sources that are themselves open to question. However, the Governing Board looks for a convergence of data to support the validity argument (Hambleton & Pitoniak, 2006). Such evidence could come from additional sources of information, including information from schools in the form of teacher ratings or course-taking information, as well as from the results of tests that assess similar constructs. Comparing standard-setting results to external sources of information is a way to check whether the performance standards are set at approximately the right level (Kane, 2001). Evidence could also be based on the *reasonableness* of the performance standards, such as those practitioners in a field deem useful or appropriately stringent or how different subgroups within the examinee population perform. Reasonableness is most relevant if pass rates are very different from what was expected (Kane, 2001). Evidence could also be drawn from the relationships between test scores, including achievement levels, and important outcomes, thereby observing the scores' ability to predict future performance. Because NAEP Achievement Levels are dependent on the cut scores derived from the scale scores, linking studies are included that both directly describe the achievement levels as well as studies that focus on the scale scores.

The NASEM (2017) report noted the importance of linking to external measures to help add meaning to NAEP Achievement Levels. Since NAEP was not designed or intended to match any one external measure exactly, these studies do not look for an exact match in interpretation and use of scores, including performance standards. For example, we would not expect performance at *NAEP Proficient* to match exactly to a state or international assessment's proficient performance. Likewise, we would not expect a perfect correlation between NAEP performance and other indicators of educational outcomes, such as graduation rate or college preparedness measures. However, we can still find meaning when results between the assessment scores demonstrate reasonable patterns of similarity or concurrence. For example, evidence of validity could come from findings that students performing at *NAEP Proficient* are more likely to attend a college or university than those at *NAEP Basic*, or students performing at *NAEP Advanced* are more likely to major in a science, technology, engineering, or mathematics (STEM) field in college than those performing at other levels. These patterns of performance lend validity evidence to the NAEP Achievement Levels when they support a logical argument with empirical evidence. The next section contains a summary of studies describing external evidence to address how the NAEP Mathematics and Reading Achievement Levels relate to various related constructs and scales and the reasonableness of these relationships. The studies provide support for the claim that NAEP Achievement Levels meaningfully relate to other measures of student achievement and other indicators of educational outcomes for all students.

Linking NAEP Grade 12 Mathematics to the High School Longitudinal Study

A series of working papers commissioned by NCES and conducted by researchers at the American Institutes for Research used "overlap samples" of students from the national High School Longitudinal Study (HSLS) of 2009 who were also part of the 2013 grade 12 NAEP Mathematics assessment sample. The studies examined the scale scores, including the specific scale score ranges that operationalize the NAEP Achievement Levels (see Figure 2-1), to evaluate validity claims. The four studies provide a body of evidence that examines relationships between NAEP scores, including NAEP Achievement Levels, and various relevant student characteristics and performance factors. As a corpus, they examine how student characteristics early in high school correlate with subsequent NAEP performance in grade 12

mathematics, the connection between these grade 12 NAEP scores and important postsecondary outcomes, and student characteristics, such as gender and race/ethnicity. Each set of questions requires a frame to establish how the evidence supports the NAEP validity argument, with focus on levels of achievement. Collectively, they provide evidence in support of the validity argument broadly as well as relating to topics of fairness.

Evidence from the Relationship Between STEM Course-Taking in High School and Grade 12 NAEP Mathematics Performance

Yee et al. (2021) examined high school STEM course-taking in relation to end-of-high school mathematics proficiency, as measured by the NAEP 2013 grade 12 mathematics assessment. The overlap sample included 2,710 students who participated in the HSLS of 2009. The authors used a marginal maximum likelihood regression analysis and cluster analysis to examine (1) how strongly STEM course-taking in high school related to end-of-high school mathematics proficiency, (2) how the relationship between STEM course-taking and end-of-high school mathematics proficiency changed when controlling for measures of prior mathematics achievement and student background characteristics, and (3) whether there was evidence of distinct STEM course-taking patterns in high school for students who score at or above NAEP's college preparedness benchmark in mathematics.

Yee et al. (2021) found strong relationships between course-taking indicators and grade 12 NAEP Mathematics performance with notable size of the statistical relationships. For example, students who took calculus scored nearly 46 NAEP scale-score points (equivalent to 1.4 standard deviations on the grade 12 NAEP) higher on average than students with a similar number of total mathematics course credits whose most advanced mathematics course was algebra II. Also, they found that, after accounting for pre-high school characteristics (e.g., race/ethnicity, gender, socioeconomic status, prior mathematics achievement), science and engineering courses, and STEM course grade point averages, the above relationships were attenuated but still substantial:

> While these results underscore the fact that high school course-taking in mathematics is strongly related to other factors, including pre-high school factors (such as math proficiency at the beginning of grade 9), they also suggest that more advanced course-taking may lead to meaningfully higher mathematics proficiency. (pp. iii–iv)

Yee et al. (2021) also used this linkage to examine how the NAEP Achievement Levels were associated with student course-taking. Students who had reached algebra I or geometry as their highest mathematics course in high school had an average NAEP score falling below *NAEP Basic*. Those who had gone beyond algebra II on average performed at *NAEP Basic* or above, and those who had taken a calculus course by the end of grade 12 had an average score that fell in the *NAEP Proficient* range.

Yee et al.'s (2021) findings were consistent with earlier research that showed positive effects of advanced course-taking on achievement. For example, Byun et al. (2015) also used a nationally representative longitudinal dataset and an indicator for having taken mathematics courses beyond algebra I as its main predictor of interest, and their results were similar to those of Yee et al. (2021). These findings, while "suggesting that advanced coursework in mathematics and science may help to improve mathematics proficiency" (p. 29), also provide evidence that mathematics proficiency as measured by NAEP among U.S. students is meaningfully related to

course-taking patterns, providing a source of external evidence in that those at higher proficiency levels experience different outcomes than those at lower proficiency levels.

Evidence from Motivation, High School STEM Course-Taking, NAEP Mathematics Achievement, and Social Networks

Zhang, Bohrnstedt, Zheng, et al. (2021) looked at high school students' pathways to a college STEM major and achievements that begin in early schooling years and continue to develop in secondary school based on previous research on the "STEM pipeline" (Eccles, 1994; Wang, 2013; Wang & Degol, 2013). Examining simple comparisons of mathematics and science motivation between students in STEM and non-STEM majors, the authors found evidence that students in STEM majors had a higher level of mathematics and science motivation in measured constructs as compared with students in non-STEM majors, adding to an argument for the predictive validity of grade 12 NAEP Mathematics scores for college enrollment and the choice of a STEM major.

Zhang, Bohrnstedt, Zheng, et al. (2021) used an overlap sample from the full HSLS data from multiple waves (2009–2016), which included 13,433 students across all waves, and the overlap sample of approximately 3,480 students who participated in the HSLS of 2009 and who also took the 2013 grade 12 NAEP Mathematics assessment. The authors conducted a two-stage multiple imputation to compute projected NAEP Mathematics achievement scores for the full HSLS sample members. The imputation was implemented for the full HSLS sample members' background variables, and the HSLS sample members' imputed NAEP plausible values, computed by conditioning on all HSLS variables. Finally, a regression model was used to generate the NAEP plausible values for the full HSLS sample members (Ogut et al., 2015). This imputation allowed the NAEP-HSLS overlap sample to serve as an opportunity to obtain predictive validity evidence of grade 12 NAEP Mathematics scores for college enrollment and the choice of a STEM major. The authors found that NAEP scores were associated with the probability of students' entrance into a STEM major in college, as well as the direct relationships between other HSLS variables, including motivational variables and STEM course-taking variables, and the likelihood of students entering a STEM field in college. Though this study focused on scale scores, studies also examined these data focusing on achievement levels specifically.

Evidence from College Enrollment Benchmarks for the NAEP Grade 12 Mathematics Assessment

Ogut et al. (2021, 2023) also used an overlap sample of 3,470 students' data from the 2009 HSLS and the 2013 NAEP in grade 12 mathematics to examine the relationship between NAEP achievement and college enrollment and other college-related outcomes. First, they used grade 12 NAEP Mathematics achievement to model the probability of students' enrollment in postsecondary education with or without remediation by the selectivity of the colleges they enrolled in, their persistence in postsecondary education, and their majoring in a STEM field. They used ordered logistic regression for this analysis. Next, the authors examined how well grade 12 NAEP Mathematics achievement, as compared to achievement on the SAT mathematics college entrance exam, predicted postsecondary outcomes (e.g., enrollment in a four-year college) with and without controlling for high school grade point average. The authors obtained estimates from models with NAEP or the SAT as the only predictor, as well as from models including grade point average and NAEP or the SAT. Model fit statistics were used to

compare the relative performance of NAEP and SAT in the prediction of the postsecondary outcomes.

Overall results of the Ogut et al. studies (2021, 2023) showed that NAEP Mathematics achievement explained much of the variation in postsecondary outcomes. In terms of reasonableness, study findings (Hambleton & Pitoniak, 2006) could be considered to show reasonable patterns of performance on NAEP in terms of college enrollment, the need for remedial course-taking, choosing STEM majors, and persistence. About 28 percent of the variation in overall postsecondary enrollment and 34 percent of the variation in the selectivity of college enrollment was explained by NAEP Mathematics achievement. Performance for the *NAEP Basic* achievement level corresponded to a 33-percent probability of entry into a four-year college (5 percent into a highly selective college), performance at the *NAEP Proficient* level had a 64-percent probability of entry into a four-year college (18 percent into a highly selective college), and performance at the *NAEP Advanced* level had an 88-percent probability of enrollment (50 percent into a highly selective college). In addition, the majority of those who performed at *NAEP Proficient* or *NAEP Advanced* were most likely to be enrolled in postsecondary education without need for remediation, whereas two-thirds of those who fell below *NAEP Basic* did require remediation, and approximately half of those who performed at *NAEP Basic*.

Beyond enrollment and remediation, Ogut et. al (2021) examined how performance classified by NAEP Achievement Levels related to selecting a major in STEM. Performance at the *NAEP Basic* achievement level was associated with a 13-percent probability of choosing a STEM major, as compared with 28 percent of the students who performed at the *NAEP Proficient* achievement level and 52 percent at *NAEP Advanced*. Finally, performance at the *NAEP Basic* level was associated with a 72-percent probability of persisting in college until at least the junior year, while performance at the *NAEP Proficient* level corresponded to an 84-percent probability, and *NAEP Advanced* corresponded to a 92-percent probability of persistence. The *NAEP Proficient* achievement level in grade 12 mathematics was associated with higher percentages of students matriculating to four-year colleges and higher probability of related factors than *NAEP Basic,* and *NAEP Advanced* performance was associated with greater success than *NAEP Proficient*.

In terms of predictive validity, Ogut et al. (2021) also compared the SAT mathematics college admissions test with the NAEP grade 12 mathematics assessment, finding that the NAEP was about equally good at predicting students' enrollment into postsecondary education, the selectivity of college enrollment, remedial course-taking, choosing a STEM major, and persistence when controlling for overall high school grade point average. This comparison of tests of similar constructs on a common student sample provides additional external evidence of validity.

Evidence from Examining Motivation and Student Performance

Zhang, Bohrnstedt, Park, et al. (2021) undertook a study to examine the relationship between student motivation and performance on the grade 12 NAEP Mathematics assessment, finding evidence supporting a claim of fairness within the validity argument across subgroups. Using an overlap sample of approximately 3,500 students from the 2009 HSLS and the 2013 NAEP in mathematics, the authors first constructed a hypothesized conceptual model based on extensive research, which consistently demonstrated that students with high mathematics self-efficacy who also had a high subjective value about mathematics were more likely to have

higher mathematics achievement. The authors described the model as evidence-based and comprehensive in its description of the relationships among mathematics performance, mathematics motivation, educational expectations, and mathematics course-taking as students move from grade 9 to grade 12. The model included a series of sequential paths representing five interrelated, evidence-based hypotheses about how students' mathematics motivation and educational expectations in their freshman year of high school related to mathematics motivation, educational expectations, and mathematics course-taking in grade 11 and how these variables, along with school-level contextual variables, related to mathematics achievement in grade 12. In a series of factor analyses, they confirmed and refined the model.

Then, using multiple group structural equation modeling analyses (Bollen, 1989; 1993), Zhang, Bohrnstedt, Park, et al. (2021) addressed the research question: What was the relationship between grade 12 NAEP Mathematics performance and mathematics motivation (and educational expectations), taking into account grade 9 mathematics achievement, family and school background factors, and difficulty of high school mathematics courses taken (e.g., taking advanced, regular, or basic courses)? And did these relationships differ by gender and race/ethnicity? They determined whether the relationships varied by gender and race/ethnicity groups. From this multiple group structural equation modeling analysis, the authors found evidence that the mathematics motivation model fit the overlap sample data quite well regardless of gender or race/ethnicity, with a root mean square error approximation of 0.03 and a comparative fit index of 0.96 overall. Subgroup comparisons showed a similar pattern, suggesting evidence of measurement invariance across subgroups and supporting the validity argument across subgroups.

Though this study did not focus on NAEP Achievement Levels specifically, it was included because the NAEP Achievement Levels are based on scale score cut points, and so the fairness and validity of the scale scores are important to consider.

Linking NAEP Reading to the Early Childhood Longitudinal Study

Dogan et al. (2015) looked at the correspondence between NAEP Achievement Levels and the Early Childhood Longitudinal Study (ECLS-K) proficiency levels at grade 8. The ECLS-K was a longitudinal study conducted by NCES that followed a cohort of students who entered kindergarten during the 1998–1999 school year through their eighth-grade year in 2006–2007, collecting data from students, parents, teachers, and schools. Unlike NAEP Reading, the reading portion of the ECLS-K was reported at the student level and at 10 developmentally descriptive ECLS-K reading proficiency levels. The two highly correlated assessments (r=.83) were developed from the same NAEP framework, evidence of a very similar test construct.

Using data from students who took both the NAEP and the ECLS-K in 2007 (N=1,290), Dogan and colleagues (2015) established the statistical link that allowed the comparison between grade 8 NAEP Achievement Levels in reading and 10 fine-grain and developmentally descriptive ECLS-K reading proficiency levels. Using a regression procedure (Cohen, 2005), they projected the scores onto a common scale for comparison. The authors found that most of the students taking the ECLS-K who were at Level 1 (letter recognition) through Level 6 (literal inference) and roughly half of the students at Levels 7 (extrapolation) and 8 (evaluation) were at the *NAEP Basic* level. Sixty-four percent of those at Level 9 (evaluating nonfiction) and 70 percent of those at Level 10 (evaluating complex syntax) were at the *NAEP Proficient* level. In addition, 13 percent of the students at Level 10 were at the *NAEP Advanced* level. In sum, their results indicated a strong and consistent relationship between NAEP Achievement Levels and

ECLS-K proficiency levels. “Higher proficiency levels on ECLS-K corresponded to higher achievement levels in NAEP” (p. 199). The authors’ findings add to the body of evidence in support of the grade 8 reading NAEP Achievement Levels by establishing their relationship with developmental reading skills.

Linking NAEP Reading and Mathematics to College Entrance Exams and Other Postsecondary Preparedness Measures

Other relevant lines of research providing external evidence of validity for the NAEP Achievement Levels come from efforts to relate NAEP scale scores and achievement levels to measures that reflect student postsecondary preparedness (Tables 4-2 and 4-3). One of these lines comes from the effort to evaluate the possibility of reporting the preparedness of U.S. grade 12 students for postsecondary education or entry into job-training programs as part of NAEP reporting by researchers from the Educational Testing Service (Moran, Freund, & Oranje, 2012; Moran, Oranje, & Freund, 2012). The next studies to be summarized focused on both the national and state levels. These studies were conducted to statistically relate performance on NAEP with results from other assessments that serve as indicators of college readiness, course placement, and workforce entry. A second line of research comes from efforts to understand whether students in grade 8 are on track to be ready for college and career. This body of research evaluates readiness for postsecondary performance by relating NAEP to external measures previously benchmarked for readiness for postsecondary demands (Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b).

NAEP Grade 12 Academic Preparedness Research

The overarching objective of this collection of studies was to establish statistical relationships between NAEP scores and various indicators of postsecondary performance. This was done to identify reference points or ranges on the grade 12 NAEP Reading and Mathematics scales that correlate reasonably with other postsecondary benchmarks for reading and mathematics. Essentially, the goal was to see if such associations could justify including preparedness indicators into NAEP reporting, such as the percentage of grade 12 students who are academically ready for college, both nationally and at the state level. The key steps of the analyses across studies were (1) estimating the correlation between NAEP and postsecondary measure(s), (2) determining the appropriate methodology for linking based on those correlations, and (3) applying procedures effectively. Table 4-2 summarizes one national study and four state-specific studies (Florida, Massachusetts, Michigan, Tennessee) that addressed preparedness for postsecondary benchmarks on other assessments (SAT, ACT) as well as other measures in the case of Florida (Moran, Freund, & Oranje, 2012; Moran, Oranje, & Freund, 2012; Xi et al., 2016a, 2016b, 2016c).

Due to correlations between NAEP and postsecondary measures, with some exceptions in mathematics, NAEP and the other measures could not be deemed sufficient for assessing the same construct using a minimum correlation criterion of .87. (Dorans & Walker, 2007). Instead they assessed constructs by statistically relating them to see the nature of the relationships between postsecondary benchmarks and NAEP scores, including achievement levels. In the studies (Moran, Freund, & Oranje, 2012; Moran, Oranje, & Freund, 2012; Xi et al., 2016a, 2016b, 2016c), various statistical techniques, including latent regression, smoothing, and statistical projection, were used to establish the relationships and identify potential markers on the NAEP scale, providing validity evidence. Additional analyses in these studies examined the

invariance across subgroups, assessing the relative stability of the construct across subgroups and providing evidence relevant to the evaluation of fairness in the assessment of students.

The studies described in Table 4-2 exhibit several limitations, including at least three general ones. First, because analyses necessarily were conducted at the state level to account for the influence of state-specific education policies and practices, the findings of the studies may not be applicable to all states or to the nation as a whole. Second, evidence of a lack of invariance across student subpopulations raises concerns regarding the stability of test constructs for certain groups, suggesting a potential fairness issue. Finally, in most cases, correlations between the NAEP scale and the postsecondary measures were not considered "strong" (Dorans & Walker, 2007). While this is not surprising given that the tests were not designed to be exactly equivalent in meaning (i.e., measuring somewhat different constructs), it underscores the need for caution in interpretation.

When examining relationships among measures for different subgroups, there are sometimes restrictions to keep in mind (Linn, 1994). For example, if some subgroup performances are not as variable at one end of the score distribution, there could be a restriction of range for the subgroup that impacts correlation coefficients. Such statistical artifacts can require caution when interpreting. For example, one of the major findings from the SAT linking study (Moran, Oranje, & Freund, 2012; Table 4-1) was that the preparedness indicator for reading aligned with *NAEP Proficient* at grade 12, but for math it was between *NAEP Basic* and *NAEP Proficient*, which is difficult to explain. Though limitations exist, these studies (summarized in Table 4-2) provide some evidence of the relationships between NAEP score scales, including NAEP Achievement Levels, and other widely used and technically defensible measures of college entrance (Moran, Oranje, & Freund, 2012; Xi et al., 2016a, 2016b, 2016c). In addition, postsecondary readiness indicators and performance outcome measures provide further information to help add meaning to NAEP Achievement Levels at the high school level (Moran, Freund, & Oranje, 2012). Further research may be warranted to strengthen understanding of the relationship between NAEP Achievement Levels and postsecondary preparedness.

Table 4-2
NAEP Grade 12 Academic Preparedness Research

Evidence Source	Summary of Contribution to Validity Evidence for the NAEP Achievement Levels	Study-Specific Limitations
Establishing a Statistical Relationship between NAEP and SAT® (Moran, Oranje, & Freund, 2012)	The study statistically related NAEP and the SAT and used that relationship to identify a reference point or range on the grade 12 NAEP Reading and Mathematics scales associated with the College Board's preparedness benchmarks on the SAT reading and mathematics measures. The study results included reference points based on the percentages of students in the overall 2009 NAEP twelfth-grade sample (from both public and private schools). The study found the correlation between scores on the two reading scales was 0.74, and the correlation was 0.91 between the two math scales.	Results showed a lack of invariance across major population subgroups in the statistical relationships established between NAEP and the SAT for both mathematics and reading. There was also a weak relationship between NAEP and SAT reading, which called for additional investigation and evaluation to determine how or if preparedness can be reported for NAEP twelfth-grade reading.

Evidence Source	Summary of Contribution to Validity Evidence for the NAEP Achievement Levels	Study-Specific Limitations
NAEP 12th Grade Preparedness Research: Analyses Relating Florida Students' Performance on NAEP to Preparedness Indicators and Postsecondary Performance (Moran, Freund, & Oranje, 2012)	The study explored the relationships between Florida students' performance on the grade 12 NAEP assessments and various indicators of postsecondary preparedness (e.g., college enrollment status and first-year college grade point averages) to explore potential preparedness reference points on the NAEP scales. Data from Florida public school students who participated in the 2009 NAEP grade 12 reading or mathematics assessments (approximately 3,200 in math and 3,400 in reading) were used with NAEP sampling weights to appropriately represent twelfth-grade public school students in Florida in that year. The study found that patterns of results did not contradict the potential preparedness reference points on the NAEP Reading and Mathematics benchmarks identified through the national NAEP-SAT linking study.	Data was limited to students enrolled in Florida public postsecondary institutions.
NAEP 12th Grade Preparedness Research: Establishing a Statistical Relationship between the NAEP and ACT Assessments in Reading and Mathematics for Grade 12 Michigan Students (Xi et al., 2016a)	Michigan participated in the state-level statistical linking research connecting NAEP and ACT, which used data on students who were part of the NAEP grade 12 sample during the 2012–2013 school year. About 2,900 and 3,100 students in grade 12 were assessed in reading and mathematics, respectively. The results showed that, in Michigan, the ACT College Readiness Benchmarks and the *NAEP Proficient* achievement level cut scores corresponded well to each other for reading in both linking directions but slightly differed for mathematics.	The grade 12 NAEP assessment in Michigan was administered almost a year later than the statewide ACT administration.
NAEP 12th Grade Preparedness Research: Establishing a Statistical Relationship between the NAEP and ACT Assessments in Reading and Mathematics for Grade 12 Tennessee Students (Xi et al., 2016b)	Tennessee participated in the state-level statistical linking research connecting NAEP and ACT, which used data on students who were part of the NAEP grade 12 sample during the 2012–2013 school year. About 3,000 and 3,200 students in grade 12 were assessed in reading and mathematics, respectively. The results showed that, in Tennessee, the ACT College Readiness Benchmarks and the *NAEP Proficient* achievement level cut scores corresponded well to each other for reading in both linking directions (i.e., the projection results were 1 scale-score point different from the ACT benchmark/*NAEP Proficient* level) but differed more for mathematics.	The grade 12 NAEP assessment in Tennessee was administered almost a year later than the statewide ACT administration.

Evidence Source	Summary of Contribution to Validity Evidence for the NAEP Achievement Levels	Study-Specific Limitations
NAEP 12th Grade Preparedness Research: Establishing a Statistical Relationship between the NAEP and SAT Assessments in Reading and Mathematics for Grade 12 Massachusetts Students (Xi et al., 2016c)	Massachusetts participated in this study and provided the critical SAT data necessary to conduct the linking study with NAEP. Approximately 2,400 public school students in Massachusetts were sampled for each subject. The results showed that, in Massachusetts, the SAT benchmarks and the *NAEP Proficient* achievement level cut scores corresponded well to each other for reading in both linking directions but differed somewhat for mathematics.	

NAEP Grade 8 Academic Preparedness Research

In another collection of studies, researchers from Educational Testing Service looked for evidence of readiness for postsecondary performance based on grade 8 NAEP performance (Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b). The studies looked at relationships that identified reference points or ranges on the NAEP Reading and Mathematics scales that reasonably associated with the other postsecondary benchmarks for reading and mathematics measures.

ACT's EXPLORE® was used for the three studies in three different states. Each study identified a reference point or range on the NAEP grade 8 reading and mathematics scales reasonably associated with ACT's preparedness benchmarks on the EXPLORE® reading and mathematics measures. In other words, identifying such points could have justified including in NAEP reporting the percentage of students at grade 8 who are academically on track for college for the nation and for states. Three states (Kentucky, North Carolina, Tennessee) provided the EXPLORE® data necessary to calculate the relationship with NAEP. As expected, due to correlations between NAEP and postsecondary measures, with some exceptions in mathematics, NAEP and the other measures could not be considered equivalent using a criterion of a minimum correlation of .87 (Dorans & Walker, 2007). However, the scales could be related statistically to see the relationships between postsecondary benchmarks and NAEP scores, including achievement levels. Table 4-3 summarizes the three state-specific studies that address preparedness for postsecondary benchmarks on the other assessment.

In the studies, various statistical techniques, including latent regression, smoothing, and statistical projection, were used to establish the relationships and identify potential markers on the NAEP scale, providing validity evidence. Across the three states, study results showed that the College Readiness Benchmarks for EXPLORE® and the *NAEP Proficient* achievement level cut scores correspond well to each other in both linking directions, with NAEP scale-score points just above the *NAEP Proficient* achievement levels providing a reasonable basis for reporting "on track for postsecondary preparedness." In all three studies, the authors recommended additional analyses to examine measurement invariance across subgroups to establish the relative stability of tested constructs across subgroups, thereby providing evidence relevant to the evaluation of fairness.

As with the studies described in Table 4-2, the studies in Table 4-3 have some limitations, including at least three general ones. First, because analyses necessarily were conducted at the

state level given the influence of state-specific education policies and practices, the studies may not generalize to all states or to the nation overall. Second, evidence of a lack of invariance across student subpopulations raises questions as to whether the test constructs are stable for certain groups, suggesting a fairness issue. Finally, in most cases, correlations between the NAEP scale and the postsecondary measures were not considered "strong" (Dorans & Walker, 2007). While not surprising since the tests were not designed to be exactly equivalent in meaning (i.e., measuring somewhat different constructs), it does call for caution in interpretation. In addition, the authors of the three studies recommended further content alignment work be conducted independently to provide further context for these results (see Table 4-1).

Table 4-3
NAEP Grade 8 Academic Preparedness Research

Evidence Source	Summary of Contribution to Validity Evidence for the NAEP Achievement Levels
NAEP 8th Grade Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for Kentucky Students (Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016)	Results showed that NAEP scale-score points at or just above the *NAEP Proficient* achievement levels could form a reasonable basis for reporting "on track for preparedness." Approximately 32 percent of Kentucky grade 8 students met that criterion for reading, and 31 percent met the criterion for math.
NAEP 8th Grade Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for North Carolina Students (Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a)	Approximately 29 percent of North Carolina grade 8 students met that criterion for reading, and 35 percent met the criterion for math. On the other hand, the projection results of the *NAEP Proficient* cut score on the EXPLORE® scale are very close to the existing EXPLORE® benchmarks for reading and mathematics.
NAEP 8th Grade Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for Tennessee Students (Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b)	Approximately 31 percent of Tennessee grade 8 students met that criterion for reading, and 32 percent met the criterion for math. On the other hand, the projection results of the *NAEP Proficient* cut score on the EXPLORE® scale are very close to the existing EXPLORE® benchmarks for reading and mathematics.

Linking to International Assessments

Another source of external information that may help provide context for understanding NAEP Achievement Levels comes from linking studies of NAEP and international assessments. Early attempts to link to international assessments were conducted in the 1990s with various linking approaches (Beaton & Gonzales, 1993; Johnson & Siengondorf, 1998), establishing a set of approaches for linking NAEP with international assessments. After the current achievement levels were set, Phillips (2014) conducted a statistical linking study between the 2011 NAEP in grade 4 reading and the 2011 Progress in International Reading Literacy Study (PIRLS) in grade 4 reading. The primary purpose of the study was to obtain a statistical comparison between NAEP and PIRLS and by expressing both assessments in the same metric, produce international benchmarks for the NAEP grade 4 reading achievement levels. "At each level, the linking shows that the NAEP Grade 4 reading achievement levels are higher than the PIRLS

international benchmarks. This finding provides one piece of validity evidence that NAEP results are internationally competitive" (Phillips, 2014, p. i).

NCES (2013) conducted a linking study of NAEP and Trends in International Mathematics and Science Study (TIMSS) assessments to provide each U.S. state with a way to examine how their students compare academically with their peers around the world in mathematics and science. Grade 8 students in all 50 states, the District of Columbia, and Department of Defense schools were assessed in mathematics and science in 2011. The study used nine states' NAEP scores to predict performance on TIMSS. In addition to the TIMSS U.S. national sample, Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina participated in 2011 TIMSS at the state level. Approximately 19,600 public school students from the validation states were selected to participate in the TIMSS assessment. In addition, a total of 10,500 grade 8 students were selected from randomly sampled classrooms in 500 U.S. public and private schools to participate in the TIMSS assessment.

These nine states served as "validation states" for the linking study. Their actual TIMSS scores were used to validate the predicted results. In addition, 38 countries and nine subnational education systems from various countries assessed grade 8 students in 2011 TIMSS. Multiple samples of students were assessed during the NAEP testing window (January–March) as well as the TIMSS testing window (April–June). Results in mathematics and science were reported as average scores on the TIMSS scale (0–1,000, with an average of 500).

The study used three different approaches to linking (i.e., statistical moderation, statistical project, and calibration linking). Given the number of differences between NAEP and TIMSS content, administration, and accommodation policies, authors of the study report did not support interpreting predicted TIMSS scores and actual TIMSS scores as the same. However, the linking methods were all applied to predict likely TIMSS scores for each of the states based on their NAEP results, and findings provided strong validity evidence that this could be accomplished. Further, the "difference between predicted and actual TIMSS results was not statistically significant for any of the national gender or racial/ethnic groups across all linking methods" (NCES, 2013, p. 27). Though this study was not focused on NAEP Achievement Levels, it is included to inform how the scale scores relate to a well-known international assessment. Future research into linkages of TIMSS specifically to the NAEP Achievement Levels may be beneficial to gain further understanding.

Mapping to State Performance Standards

Since 2003, NCES has been comparing each state's standard for proficient performance in reading and mathematics at grades 4 and 8 by aligning the state standards with common scales from NAEP. Studies of state mapping result in establishing where each state's performance standards (i.e., achievement levels) fall on the NAEP scales and in relation to the NAEP Achievement Levels.[3] Ji et al. (2021) mapped the state proficiency standards onto the NAEP scales using state assessment results from the 2018–2019 school year and the 2019 NAEP assessments for public schools, focusing on the reading and mathematics standards that states set for grades 4 and 8.[3] For each state, NAEP equivalent scores with a range of 0–500 were determined for *NAEP Basic* and *NAEP Proficient*. Though these studies do not provide

[3] NCES is to release an updated mapping study at the end of 2024. The study was not completed in time for inclusion in this report.

evidence NAEP Achievement Levels are set appropriately, they do offer meaningful information for understanding how NAEP Achievement Levels relate to state achievement levels.

Overall, state standards for proficient performance in 2019 mapped at the *NAEP Basic* achievement level for most states in both grades and subjects. Though there are some cautions and limitations due to the indirect nature, this linkage permits comparison of state achievement levels, which is not possible using state achievement levels alone given that each state is permitted to define their own achievement levels and cut scores. When put into historical and current context, results serve as external evidence of validity. For example, results from the 2021 mapping study (Ji et al., 2021) found that state standards for proficiency mapped at a higher NAEP Achievement Level in 2019 than in 2009 for grades 4 and 8 in both reading and mathematics. Given patterns of states' performance standards adoptions, these results appear to be reasonable.

V. Uses of NAEP Achievement Levels

In previous chapters, we examined the body of evidence that supports the validity argument and the three major claims for the NAEP Achievement Levels. Chapter V draws conclusions from the validity evidence earlier in the report by inspecting the appropriateness of known interpretations and uses of the NAEP Achievement Levels. The chapter considers evidence of the degree to which the NAEP Achievement Levels reflect academic performance and college readiness as well as how the achievement levels relate to external measures of achievement and college preparedness with a focus on findings that contribute to the overall validity argument.

Characteristics of the NAEP distinguish it from other assessments, including state assessments. The main NAEP assessments are administered at the national, state, and selected urban district levels every two years, and results are reported on student achievement in grades 4, 8, and 12 at the national level. Because the main NAEP assessment is administered to a nationally representative sample of students, it is reported on student achievement in the aggregate and does not report on the performance of any individual student or school.

It is important to note that the development and purpose of NAEP Achievement Levels differ from the development and purpose of achievement levels used by states for their statewide assessment programs. Therefore, the NAEP Achievement Levels must always be differentiated from state achievement levels. The NAEP Achievement Levels are developed by panels of subject matter experts who identify the appropriate content that students should know and be able to do from a national perspective. These subject matter experts represent the nation's education researchers, educators, business leaders, and policymakers. They demonstrate knowledge of the specific subject matter and pedagogy and represent all regions of the country and demographic groups. These panels focus on the NAEP assessment frameworks during the development process rather than state-specific content standards, and they make decisions through deliberation across each panel group.

The goals of NAEP include provision of insight into how well the nation's students are meeting expectations for academic performance at the national level. The assessments allow for a common measure by which to evaluate education systems across states and districts. The NAEP results can thus illuminate trends in performance over time across the country. Therefore, the NAEP results differ in their appropriate interpretations and uses as compared to those of states' assessment results. States may tie performance to high-stakes decisions for schools or individuals. For example, Ohio requires students to meet a specific performance level on the state reading test in grade 3 to be promoted to grade 4 (Ohio Department of Education & Workforce, 2024). In this case, students who do not pass the threshold score may be retained in grade 3. Another example of a state-level high-stakes use of scores is the use of assessment results in high school graduation requirements. The Education Commission of the States (Erwin et al., 2023) reports that a majority of states (34 out of 50) require their state assessment results be included in a high school student's graduation requirement. In such examples, student test results factor into specific consequences for individual students or schools. There are other means for students to demonstrate their knowledge and skills if they do not show them on the state assessment (e.g., grade point average, interim or benchmark assessments). Such uses are different from those for NAEP. This chapter examines these appropriate interpretations and uses of NAEP Achievement Levels.

The validity argument captured in this report is intended to provide a thorough examination of validity evidence currently available in relation to the appropriate and inappropriate interpretations and uses of achievement levels. It is technical in nature and may be more information than is needed by all stakeholder groups. In addition to this report and based on the NASEM (2017, p. 9) recommendation to offer consistent interpretive guidance, the Governing Board has efforts underway to develop interpretive guides to accompany NAEP releases that offer straightforward guidance for interpreting NAEP Achievement Level results.

Appropriate Uses of NAEP Achievement Levels

In general, the appropriate use of NAEP Achievement Levels is centered on the broad policy definitions for *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced*, as they are interpreted across the grades and content areas. The percentages at or above achievement level cut scores indicate the percentage of students in a group who meet or exceed the knowledge and skills represented by specific content ALDs. These results describe "achievement for groups of students at a single point in time, progress in educational achievement for groups of students over time, and differential educational achievement and progress among jurisdictions and subpopulations" (Governing Board, 2020, p. 1). The following discussion relates bodies of evidence to the appropriate uses of these achievement levels.

Direct Interpretations of NAEP Achievement Levels

The NAEP Achievement Levels are designed to describe the students in a given group who meet or exceed the knowledge and skills represented by specific content ALDs (Governing Board, 2020). This chapter provides evidence to support the appropriateness of interpreting the range ALDs as descriptions of the NAEP Achievement Levels.

These specific descriptions are found in the NAEP assessment frameworks and reports and guide the appropriate use of NAEP Achievement Levels by articulating the connection between assessment claims and resulting scores (Governing Board, 2021a, 2021b, 2022a, 2022b). The ALDs provide substantive summaries of what the assessments were designed to assess each NAEP Achievement Level. The ALDs reflect the achievement levels' cumulative gain of knowledge and skills across grade levels within a content area. Therefore,

> students performing at the *NAEP Proficient* level also display the competencies associated with the *NAEP Basic* level, and students at the *NAEP Advanced* level also demonstrate the skills and knowledge associated with both the *NAEP Basic* and the *NAEP Proficient* levels. (Governing Board, 2022a, p. 71)

The achievement levels reflect, and the ALDs characterize, the cumulative gain in the content-area knowledge and skills. The ALDs provide examples of what students performing at the *NAEP Basic*, *NAEP Proficient,* and *NAEP Advanced* achievement levels should know and be able to do in terms of the content areas identified in the given framework (Governing Board, 2021a, 2021b, 2022a, 2022b). The ALDs are also intended to provide specific and unambiguous guidance to item developers and to provide explicit elaborations of the knowledge and skills students should demonstrate and the actions they should perform at each grade level and within each achievement level. The ALDs in the framework are accompanied by example items targeting each achievement level within each grade level and illustrating the knowledge and skills required at different NAEP Achievement Levels. Both range and reporting ALDs are specific to grade level and content area. The high-level descriptions of the knowledge and skills

are provided here for mathematics and reading, with a discussion of how range ALDs guide both development and interpretations of the NAEP Achievement Levels.

Assessment Content

NAEP has routinely gathered data on students' understanding of mathematical content (Governing Board, 2022a) with a consistent focus on collecting information on student performance in five key areas:

1. Number Properties and Operations (including computation and understanding of number concepts)
2. Measurement (including use of instruments, application of processes, and concepts of area and volume)
3. Geometry (including spatial reasoning and applying geometric properties)
4. Data Analysis, Statistics, and Probability (including graphical displays and statistics)
5. Algebra (including representations and relationships)

This classification approach describes the full spectrum of mathematical content assessed by NAEP and ensures that important mathematical concepts and skills are assessed in a balanced way across the grade levels, including mathematical practices.

The NAEP Reading assessment uses varying types of informational and literary text to allow the measurement of students' comprehension of the different kinds of text they encounter in their school and out-of-school reading experiences. The NAEP Reading assessment also measures students' ability to apply their knowledge of vocabulary as an aid in their comprehension processes. Many of the NAEP passages require interpretive and critical skills usually taught as part of the English curriculum. While NAEP assesses varied reading skills, it is not a comprehensive assessment of literary study (Governing Board, 2022b). Although similar reading behaviors are included at the different performance levels and grades, these skills are being described in relation to texts and assessment questions of varying difficulty.

Range ALDs

Each content framework incorporates range ALDs with each achievement level detailing observable evidence of student achievement. In many cases, range ALDs illustrate changes in skills across achievement levels demanding increasingly sophisticated understandings of the content from one achievement level (and from one grade level) to the next. Range ALDs communicate the expectations for students by answering the question "Given what we know about the development of reading, what should students be able to do at different grade and achievement levels when responding to different combinations of texts and tasks?" (Governing Board, 2021b, p. 64).

The ALD review studies (Moyer & Galindo, 2022, 2023) indicated generally accurate depictions of what students likely know and can do. The studies included evaluation of the alignment of assessment performance of students performing at each achievement level, as defined by reporting ALDs, to the range ALDs included in the assessment frameworks. Expert panels make judgments regarding the degree of alignment of the ALDs with the achievement level policy definitions and the ALDs (Moyer & Galindo, 2022, 2023). The evidence provided in this report supports the appropriateness of interpreting the range ALDs as accurate descriptions of student performance at each NAEP Achievement Level.

Reporting ALDs

While range ALDs communicate expectations for students, reporting ALDs are based on the results of actual student performance that is mapped back to test content, as described in Chapter IV. They answer the question "Given the distribution of NAEP performance, what can students at different grade and achievement levels do when responding to various combinations of texts and tasks?" (Governing Board, 2021b, p. 79). In this chapter, evidence supports the appropriateness of interpreting the reporting ALDs as descriptions of the NAEP Achievement Levels.

The description of anchor studies in Chapter IV includes evidence in support of cut scores on the test scale to the test content itself. With the statistical information available through item-response theory scaling techniques, the anchor descriptions report *what* students actually know and can do as opposed to the NAEP Achievement Levels, which report what students *should* know and be able to do. The studies provide evidence of alignment between the knowledge and skills students demonstrated in an achievement level, as described by the summary statements, and the expected knowledge and skills for an achievement level, as described by the content ALDs within the framework (Moyer & Galindo, 2022, 2023).

The reporting ALDs were developed using recent NAEP data and describe items for which students at each level were at least 67 percent likely to respond to correctly. Similarly, since the levels are cumulative, it is appropriate to note that a student performing at *NAEP Proficient* likely also knows and can do the skills at *NAEP Basic,* and a student performing at *NAEP Advanced* likely also knows and can do the skills at *NAEP Proficient* and *NAEP Basic*. This evidence supports the appropriateness of using reporting ALDs to express what students at each level likely know and can demonstrate at each achievement level.

A Common Yardstick

The NAEP assessments provide a common measurement across all regions, states, and subgroups for each grade level and content area. In this way, NAEP is unique in the United States, and therefore it affords interpretations that cannot otherwise be made: to examine student performance across the nation, including through achievement levels. Because the assessment is in common across states, one can compare the percentage of students performing at a given achievement level between one state to another. This is not possible in most cases when using state assessments because they differ in their design, development, test scales, administration processes, scoring, and reporting.

Even when state tests and NAEP are directly comparable, they can serve as external evidence of validity for each other. As described in Chapter IV, the NAEP mapping studies provide insight into the relative differences between assessment standards across states in terms of rigor. They allow the use of NAEP to compare the rigor of assessment standards in one state versus another. "State mapping" studies establish where each state's performance standards (i.e., achievement levels) fall on the NAEP scales in relation to the NAEP Achievement Levels. This is achieved by comparing each state's standard for proficient performance in reading and mathematics at grades 4 and 8 and aligning the state standards with common scales from NAEP (Ji et al., 2021). For example, state standards for proficient performance in 2019 mapped at the *NAEP Basic* achievement level for most states in both grades and subjects. The mapping studies allow for the inspection of patterns of states' performance standards and support the appropriate use of NAEP Achievement Levels to interpret student performance across the country.

External Evidence Supporting the NAEP Achievement Levels

The results of research studies can be used to understand the broad trends in student performance beyond validation processes. Drawing from the linking studies described in Chapter IV, this chapter uses the reported studies to identify some broad interpretations. It is important to note the studies' limitations when interpreting the results and to avoid causal inferences in almost all cases. However, evidence supports broad interpretations, described here and summarized in Table 5-1.

- Linking NAEP Reading and Mathematics to college entrance exams provided evidence that students who performed at higher levels on NAEP demonstrated readiness for postsecondary demands and college readiness as reflected by postsecondary benchmarks. This body of research evaluated readiness for postsecondary performance by relating NAEP to external measures previously benchmarked for readiness for postsecondary demands (Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b).
- Linking NAEP Reading and Mathematics to international assessments supported the interpretation that NAEP Achievement Levels can be used to predict TIMSS scores and make international comparisons. This study provided an international point of comparison to participating American states. The linking methods predicted TIMSS scores for each of the states based on their NAEP results. Further, the "difference between predicted and actual TIMSS results was not statistically significant for any of the national gender or racial/ethnic groups across all linking methods" (NCES, 2013, p. 27).
- NAEP grade 8 academic preparedness research provided evidence of a correspondence between related college readiness benchmark measures and the *NAEP Proficient* achievement level cut scores in grade 8, substantiating the ALDs of reading skill and knowledge development from grade 4 to 8 (Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b). Across the three participating states, results showed that the College Readiness Benchmarks for EXPLORE® and the *NAEP Proficient* achievement level cut scores correspond well to each other in both linking directions, with NAEP scale-score points just above the *NAEP Proficient* achievement levels providing a reasonable basis for reporting "on track for postsecondary preparedness." In all three studies, the authors recommended additional analyses to examine measurement invariance across subgroups to establish the relative stability of tested constructs across subgroups, thereby providing evidence relevant to the evaluation of fairness.
- NAEP grade 12 preparedness research demonstrated how higher NAEP score scales, which define the NAEP Achievement Levels, show evidence of higher postsecondary readiness. The studies (see Table 4-2) provide meaningful evidence of the relationships between NAEP and other widely used and technically defensible measures of college entrance (Moran, Oranje, & Freund, 2012; Xi et al., 2016a, 2016b, 2016c). In addition, postsecondary readiness indicators and performance outcome measures align with the NAEP Achievement Levels in high school (Moran, Freund, & Oranje, 2012).
- Linking NAEP Reading to the ECLS-K supported the interpretation that the NAEP Reading ALDs in grades 4 and 8 reflect developing reading skills (Dogan et al., 2015). The ECLS-K proficiency levels at grade 8 were related to the NAEP Achievement Levels. "Higher proficiency levels on ECLS-K corresponded to higher achievement levels in NAEP" (p.

199). The authors' findings add to the body of evidence in support of the grade 8 reading NAEP Achievement Levels by establishing their relationship with developmental reading skills.

- Evidence from the relationship between STEM course-taking in high school and grade 12 NAEP Mathematics performance showed that NAEP Achievement Levels can be related to course-taking patterns in grade 12 mathematics (e.g., students taking higher-level STEM courses were more likely to score *NAEP Proficient* or *NAEP Advanced*; Yee et al., 2021). These results suggest that mathematics proficiency, as measured by NAEP, among U.S. students is meaningfully related to course-taking patterns, providing a source of external evidence for the definition of proficiency on the grade 12 NAEP Mathematics assessment across the diversity of students who are performing at *NAEP Proficient* in grade 12 mathematics.
- Evidence from the study of motivation, high school STEM course-taking, NAEP Mathematics achievement, and social networks showed NAEP scores were associated with the probability of students' entrance into a STEM major in college. Also, the direct relationships between other variables, including motivational variables and STEM course-taking variables, and the likelihood of students entering a STEM field in college related to NAEP Achievement Levels (Zhang, Bohrnstedt, Zheng, et al., 2021).
- Evidence from examining motivation and student performance showed that NAEP scores, including NAEP Achievement Levels, can be related to motivation across subgroups. There was evidence that the mathematics motivation model fit the overlap sample data quite well regardless of gender or race/ethnicity. Subgroup comparisons showed a similar pattern, suggesting evidence of measurement invariance across subgroups and supporting the validity argument across subgroups.
- Evidence from college enrollment benchmarks for the NAEP grade 12 mathematics assessment supported the interpretation of a cumulative relationship between *NAEP Basic* and *NAEP Proficient* by external evidence in grade 12 mathematics. The *NAEP Proficient* achievement level in grade 12 mathematics was associated with higher percentages of students matriculating to four-year colleges and higher probability of related factors than *NAEP Basic*.

Table 5-1 summarizes the evidence in relation to the interpretations supported by the study findings.

Table 5-1
Summary of Appropriate Use of Linking Study Findings

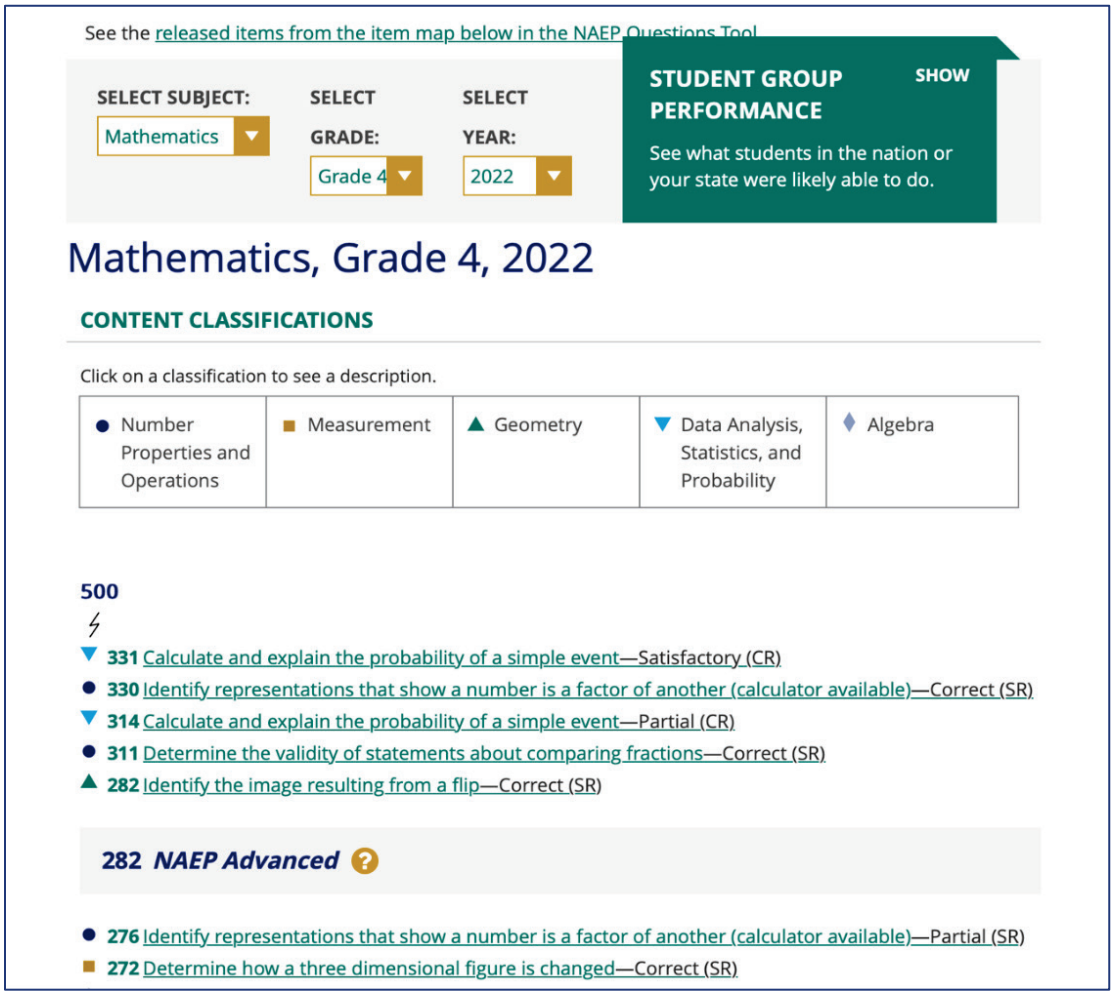
Evidence Source Topic	Citation	Interpretation
Linking NAEP Reading and Mathematics to College Entrance Exams	Moran, Freund, & Oranje, 2012; Moran, Oranje, & Freund, 2012; Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b	Students who performed at higher NAEP Achievement Levels in reading and mathematics were more prepared to meet postsecondary demands and demonstrated greater college readiness.

Evidence Source Topic	Citation	Interpretation
Linking to International Assessments	NCES, 2013	The NAEP Achievement Levels in reading and mathematics can be used to predict TIMSS scores and evaluate subgroup performance, providing an international point of comparison to participating American states.
NAEP Grade 8 Academic Preparedness Research	Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b	Student performance on the college readiness benchmark measures corresponded to the *NAEP Proficient* achievement level cut scores in grade 8.
NAEP Grade 12 Academic Preparedness Research	Moran, Freund, & Oranje, 2012; Moran, Oranje, & Freund, 2012; Xi et al., 2016a, 2016b, 2016c.	NAEP score scales, which define the NAEP Achievement Levels, are related to postsecondary readiness measures.
Linking NAEP Reading to the Early Childhood Longitudinal Study	Dogan et al., 2015	The NAEP ALDs of reading in grades 4 and 8 can be interpreted to reflect developing reading skills.
Evidence from the Relationship Between STEM Course-Taking in High School and Grade 12 NAEP Mathematics Performance	Yee et al., 2021	NAEP Achievement Levels can be related to course-taking patterns in grade 12 mathematics (e.g., students taking higher-level STEM courses were more likely to score *NAEP Proficient* or *NAEP Advanced*).
Evidence from Motivation, High School STEM Course-Taking, NAEP Mathematics Achievement, and Social Networks	Zhang, Bohrnstedt, Zheng, et al., 2021	NAEP scores were associated with the probability of students' entrance into a STEM major in college, as well as the direct relationships between other HSLS variables, including motivational variables and STEM course-taking variables, and the likelihood of students entering a STEM field in college.
Evidence from Examining Motivation and Student Performance	Zhang, Bohrnstedt, Park, et al., 2021	NAEP scores, including NAEP Achievement Levels, can be related to motivation across subgroups.
Evidence from College Enrollment Benchmarks for the NAEP Grade 12 Mathematics Assessment	Ogut et al., 2021	The cumulative relationship between *NAEP Basic* and *NAEP Proficient* is supported by external evidence in grade 12 mathematics.

NAEP Item Maps

The Nation's Report Card (NCES, 2024a) provides item maps, digital, browser-based tools intended to help readers understand student performance and guide the interpretation of scores (Figure 5-1). These maps can help further describe what it means to perform at different points along the NAEP scale score and within each NAEP Achievement Level.

Figure 5-1
Illustration of an Item Map



Source: https://www.nationsreportcard.gov/itemmaps/?subj=MAT&grade=4&year=2022

The item maps use a logic similar to that of item maps used for standard setting, anchor studies, and alignment studies, as described in earlier chapters. For each assessment, example questions (i.e., test items) are mapped onto the NAEP scale for that subject area. The interactive map provides a description of the knowledge or skill needed to answer each test question in its position on the scale with harder items appearing at the top and easier items at the bottom. The location of the questions on the map indicates that students with that score had a relatively high probability of answering the question correctly. Each item map contains a scale specific to the subject. Scales range from 0–300 or from 0–500, depending on the subject. Scale scores from a given assessment represent the scores for students who were likely to answer a question correctly or to give a complete response. Constructed-response questions for which students could earn partial credit may appear on the map multiple times, once for each level of credit. Constructed-response items are marked with “CR” on the map.

Item maps contain descriptions that indicate what students need to know or do to answer the question correctly. They also contain content classifications that refer to the specific skill area of the subject being assessed; for example, in mathematics, the content classification might be algebra or measurement. Descriptions for the questions are from questions released to the

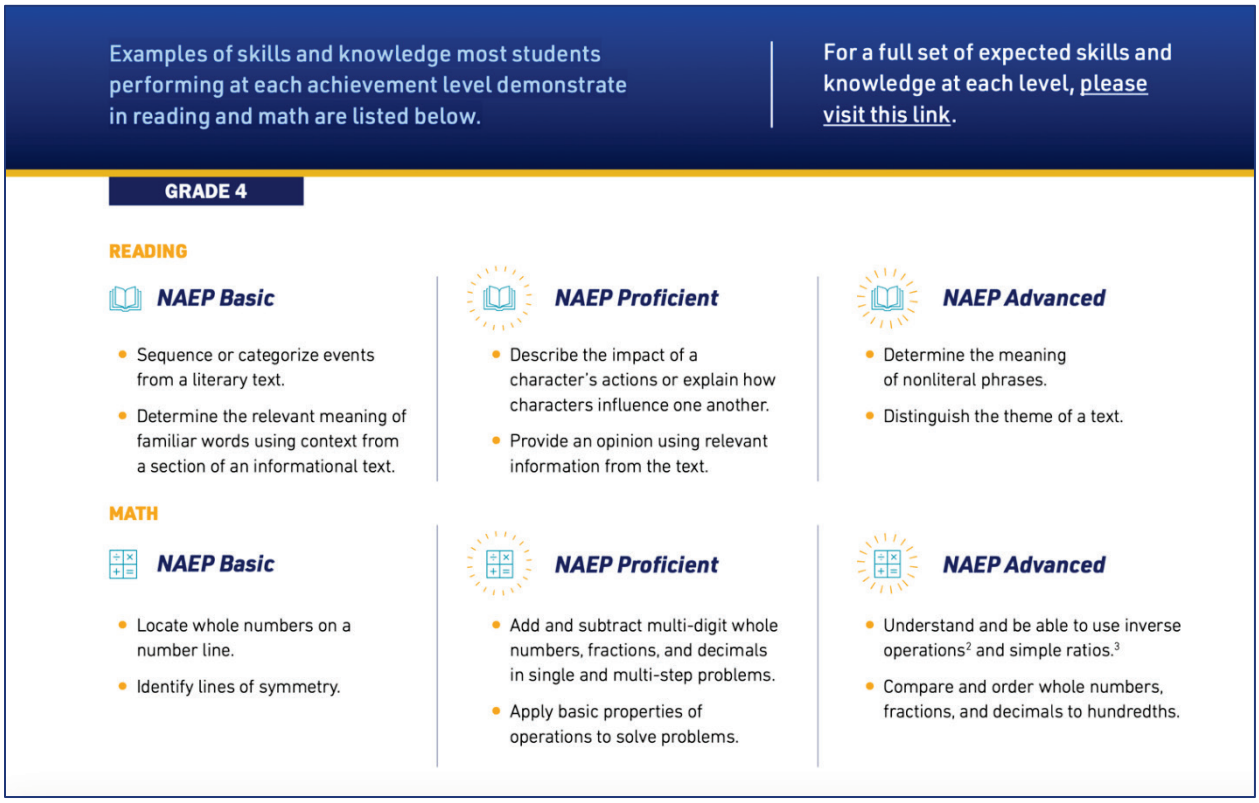
public and thus are no longer used in an assessment and available via hyperlink. For some subjects and years, no items were released and so no item descriptors are linked.

Finally, item maps contain the NAEP Achievement Level cut scores that show whether the student is performing at a *NAEP Basic*, *NAEP Proficient*, or *NAEP Advanced* level. When a user engages with the item map, they choose a subject, grade, and year. For example, for "mathematics," "grade 8," "2013," the map will generate a report that shows eighth-grade students performing at the *NAEP Advanced* level with a score of 348 were likely to correctly answer a question that required them to "solve an algebraic inequality" and would also be likely to demonstrate the skills associated with questions that appear lower than a score of 348.

Typical Interpretations of NAEP Achievement Levels

In many cases, stakeholders seek to understand and characterize the meaning of NAEP Achievement Levels. The Governing Board provides examples to support appropriate interpretations. For example, web content (Governing Board, 2022) is available publicly that succinctly summarizes appropriate inferences that can be drawn for each achievement level in grades 4 and 8.[5] Along with a brief overview consistent with this report, the web page includes a link to a briefing document that offers an explanation of achievement levels along with examples of skills and knowledge most students performing at each achievement level demonstrate in reading and math (Figure 5-2). The Governing Board is dedicated to continuing efforts to improve communications surrounding NAEP Achievement Levels and score interpretations in general. They are working toward developing an interpretive guide to accompany the release of 2024 NAEP data with achievement levels as one of the key foci.

Figure 5-2
Examples of NAEP Skills and Knowledge by Achievement Level



Inappropriate Uses of NAEP Achievement Levels

The inappropriate use of test scores, including achievement levels, is an occupational hazard and known problem for testing programs (AERA et al., 2014), including NAEP. As the Governing Board (2020) acknowledged, while NAEP measures educational achievement and progress, NAEP results alone cannot indicate either why or how achievement or progress has occurred. "Educational policies and practices that concur with NAEP progress may have caused this progress or been coincidental" (Governing Board, 2020, p. 1). Just as a doctor would not judge the effects of a child's overall health and nutrition based on only a height measurement at the annual wellness appointment, we cannot judge educational outcomes based only on NAEP scores, which are produced every other year and are designed technically to support specific interpretations. Multiple indicators are needed to evaluate an educational outcome given the multitude of limitations inherent in using a single measure. Similarly, NAEP Achievement Levels should never be used as an outcome measure to determine cause and effect impacts of state- or district-level interventions. The frequency and design of the assessment does not permit confidence in any one intervention being responsible for academic outcomes. (Noting general validity requirements to use assessments for their designed purpose and pulling from the Governing Board's intended meaning of NAEP could unintentionally help support this inappropriate use.)

As discussed in Chapter II, scores must be interpreted in the context of their validity argument and based on the stated claims (Kane, 2006). This chapter describes some known inappropriate uses and interpretations for NAEP scores and directs users to use them appropriately at all times.

The Governing Board (2020, p. 1) calls out three specific cautions. First, NAEP produces results for the nation and participating states and jurisdictions in public and private schools. It does *not* produce results for individual students or schools. This is due to various design characteristics of the assessment, including its content and student sampling approaches, and legislative constraints. As mentioned early in Chapter V, NAEP neither samples nor assesses at a level appropriate to report on individual students or schools, so users should never make claims about a school's percentage of students at *NAEP Proficient* nor about individual students at all. Because the NAEP assessments are administered only at the national, state, and selected urban district levels every two years to a sample of students, results can be reported on student achievement in grades 4, 8, and 12 only at the levels of national, state, or select urban areas. Also, because the main NAEP assessment is administered to a nationally representative sample of students, it reports on student achievement in the aggregate and does not report on the performance of any individual student or school. Since NAEP is designed to report at the aggregate level, results should never be interpreted as reflecting an individual student or school.

Second, NAEP measures progress based on successive cohorts of students, not a single cohort over time. It is appropriate to note the changes between cohorts between administrations; however, it is not possible to draw inferences about student growth. NAEP therefore does not produce results about the growth of individual students or groups of students over time.

Third, the NAEP assessments are based on independent assessment frameworks developed through a national consensus approach (Governing Board, 2021a, 2021b, 2022a, 2022b). NAEP frameworks do not represent any single state or local curriculum. This means that the states hold the authority to develop and maintain their own content standards and performance standards to guide their curriculum. The NAEP frameworks are independent of state oversight of education systems.

In addition to these cautions against inappropriate use, the Governing Board outlines additional examples. *NAEP Proficient* should never be interpreted as "on grade level." This common misinterpretation is inappropriate because there is no common definition of grade level in the United States. Rather, grade-level expectations are set as state policy and described in content standards and curriculum. These can change over time as lawmakers craft education laws that drive education leaders to set state policies. Even the definition of proficiency can differ from state to state or in comparison with *NAEP Proficient*. In sum, there is no justification for assuming *NAEP Proficient* can be equated with grade level expectations.

Similarly, score interpretations should be based on the specific skills assessed; blanket statements that students can or cannot do math or read are not accurate. For example, it is not accurate and is inappropriate to state that fourth graders in the nation cannot read based on the percentage of students scoring at *NAEP Proficient*. The *NAEP Proficient* level is associated with specific, higher-level skills. Those scoring below *NAEP Proficient* may indeed be able to read, but their reading comprehension skills are likely not at the level NAEP has identified as necessary for meeting the *NAEP Proficient* cut score. To understand what it means to perform at each NAEP Achievement Level, users must focus on the specific reading and math skills assessed in the NAEP grade level and content area by reviewing assessment documentation such as the ALDs and example items available in the assessment frameworks, the reporting ALDs, and/or information from released items and item maps. Careful review of the ALDs is important because the grade-level markers vary across states and NAEP. States define and implement curriculum and instruction; NAEP does not. And since learning is a continuum influenced by a myriad of complex inputs as well as students' opportunity to learn, gross assumptions about student knowledge and skills without context and additional data points are irresponsible and potentially very harmful. Users should always refer to the ALDs to make accurate and valuable statements regarding student knowledge and skills. Since NAEP is a single yardstick, the definitions of each of the achievement levels are not comparable to other assessments.

With any assessment program, NAEP included, it is important to understand appropriate and inappropriate uses of the scores. Unintended uses could result in harmful decisions, harmful policies developed, and other impacts that damage educational systems and the people they intend to serve. Such impacts can have lasting effects on children and families, as well as diverse and far-reaching societal outcomes.

VI. Discussion

In sum, the NAEP Achievement Levels Validity Argument Report compiles and presents accumulated evidence that supports stated claims about the valid use and interpretation of the NAEP Achievement Levels, with particular emphasis on mathematics and reading. While the Governing Board does not draw conclusions about the overall validity of the NAEP Achievement Levels to date, the NAEP Achievement Levels Validity Argument Report endeavors to synthesize evidence that is relevant to the "trial" status of the achievement levels, as indicated by legislation (Improving America's Schools Act of 1994), and in service to external evaluation and determining whether this status should be removed.

The NAEP Achievement Levels Validity Argument Report begins with a review of the purpose of NAEP and the achievement levels, including a discussion of the historical and technical contexts and major claims. The major claims are (1) NAEP Achievement Levels are established based on defensible standard-setting methods that are implemented with fidelity, (2) NAEP ALDs are defensible definitions of what students know and can do at each level, and (3) NAEP Achievement Levels meaningfully relate to other measures of student achievement and other indicators of educational outcomes for all students.

In the third section, the report illustrates how the NAEP Achievement Levels relate to the principles set forth by the Governing Board to guide and evaluate adherence to best practices for testing and measurement. In subsequent sections, the NAEP Achievement Levels Validity Argument Report presents research evidence that supports the valid use and interpretation of NAEP Achievement Levels, grouping the studies by their approaches and purposes. Finally, the report presents guidance on the appropriate uses and interpretations of NAEP Achievement Levels, noting the relationships to the presented evidence.

Overall, the studies used for the NAEP Achievement Levels Validity Argument Report (see Appendix A) employed procedures that reflected industry standards for the years in which they were conducted, with earlier studies informing subsequent studies. The recent reporting ALD studies (e.g., Moyer & Galindo, 2022) showed strength in alignment for all grades and both math and reading, except for grade 12 *NAEP Advanced*, a finding which will be explored further once the new math frameworks with updated ALDs are in place.

As a body of work, external evidence reviewed tied the NAEP Achievement Levels to other academic measures of achievement and outcomes with some utility. In some cases, these findings could inform methodologies, and, in other cases, they could inform the validity arguments for the existing NAEP assessments. Examples of such research highlighted in this report follow:

- A series of working papers commissioned by NCES and conducted by researchers at the American Institutes for Research used "overlap samples" of students from the national HSLS of 2009 who were also part of the 2013 grade 12 NAEP Mathematics assessment sample. The four studies provide a body of evidence that examines relationships between NAEP scores, including NAEP Achievement Levels, and various relevant student characteristics and performance factors.
- The ECLS-K was a longitudinal study conducted by NCES that followed a cohort of students who entered kindergarten during the 1998–1999 school year through their eighth-grade year in 2006–2007, collecting data from students, parents, teachers, and schools. Unlike NAEP Reading, the reading portion of the ECLS-K was reported at the student level

and at 10 developmentally descriptive ECLS-K reading proficiency levels. Dogan et al. (2015) looked at the correspondence between NAEP Achievement Levels and ECLS-K proficiency levels at grade 8.

- Looking on NAEP at preparedness of U.S. grade 12 students for postsecondary education or entry into job-training programs by researchers from the Educational Testing Service (Moran, Freund, & Oranje, 2012; Moran, Oranje, & Freund, 2012), studies were conducted to statistically relate performance on NAEP with results from other assessments that serve as indicators of college readiness, course placement, and workforce entry.
- Research looking at whether students in grade 8 were on track to be ready for college and career evaluated readiness for postsecondary performance by relating NAEP to external measures previously benchmarked for readiness for postsecondary demands (Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a; Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b).

Evidence from linking and mapping studies described in Section IV allows for some degree of comparison with these external sources of information that align with the expectations captured in the performance standards. All of these types of evidence are relevant in the validity argumentation for NAEP Achievement Levels.

Limitations

In this discussion, the Governing Board acknowledges the limitations of the report and where more specific evaluations could be made by grade level and subject area to further validation of the NAEP Achievement Levels. Some limitations were known at the outset of the project. First, as one of many activities outlined in an Achievement Levels Work Plan in response to recommendations made by the National Academies of Sciences, Engineering, and Medicine (NASEM) (2017), this report does not act as an independent evaluation of the NAEP Achievement Levels overall or for any specific NAEP assessment. Furthermore, the report does not build a singular validity argument for any one or all of the NAEP assessment program's achievement levels; this is beyond the scope of the effort. Also, the report represents the time frame in which it was written and does not speculate about the impact of developments (e.g., AI) on the validity of the NAEP Achievement Levels. Finally, the report describes studies of various methodologies and levels of academic rigor. The Governing Board has endeavored to be forthcoming about flaws that could impact its interpretation of study results. However, the Governing Board acknowledges the potential for misinterpretation or omission of relevant cautions or caveats.

In order to safeguard against these known limitations of the NAEP Achievement Levels Validity Argument Report, initial drafts of this validity argument were reviewed internally and externally prior to finalization. Specifically, the following reviews were conducted:

- Internal reviews by members of the Governing Board's Committee on Standards, Design and Methodology (COSDAM)
- The psychometricians on the staff of the Governing Board
- NCES staff
- Two external reviewers identified as experts in assessment and specifically in achievement levels and validity: Marianne Perie, the senior research director of assessment and accountability with WestEd, and Henry Braun, Boisi Professor of

Education and Public Policy and director of the Center for the Study of Testing, Evaluation, and Education Policy in the Lynch School of Education at Boston College.

In addition to providing input to improve the clarity of this report and the quality of its contents, reviewers offered input on limitations and future work to strengthen validity.

In regard to external evidence, and linking studies in particular, external and NCES reviewers noted that additional studies that focus on linking to the NAEP Achievement Levels specifically would be beneficial. Many linking studies that have been completed (Tables 4-2 and 4-3) have focused primarily on NAEP scale scores, and those that have offered input into achievement levels have primarily focused on grades 8 or 12, not 4.

The NCES reviewers noted the lack of attention to validity as it pertains to the range of performance below *NAEP Basic*. In 2022, a significant percentage of students fell below *NAEP Basic*; for example, 25 percent of grade 4 students performed below this level in mathematics. The Governing Board has had discussions regarding this level of performance and agrees with taking measures to ensure more information is available at the low end of the achievement scale. The Governing Board has addressed this in recently updated NAEP frameworks. The current focus is on efforts to ensure more items are available at the low end rather than on assigning a new achievement level.

Additional feedback expressed the need to better communicate that while state mapping studies and linking studies offer insight into how NAEP Achievement Levels can be communicated, they are not necessarily validity evidence in and of themselves. And, on a similar note, the external evidence may offer insight into meaningful differences between *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced*, but this should not imply that this means the NAEP Achievement Levels are considered correct. In fact, external evidence should be interpreted only as information to help aid in interpretations and add meaning through commonly understood external measures and outcomes; internal studies to examine how well the NAEP assessment findings align with the defined ALDs provide information needed to assess whether the NAEP Achievement Level results accurately reflect what the Governing Board's assessments claim they do. It is an entirely different debate whether the cut scores associated with these levels are set at the right point. As noted, the Governing Board's internal processes for developing ALDs include various content experts and stakeholders with experience working with the grade levels and content areas assessed; however, there will always be some level of subjectivity regarding what constitutes performance at *NAEP Proficient*.

References

Alexander, L. (1986). *Time for results: An Overview.* The Phi Delta Kappan (86)3, 202-204.

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999, July). *The NAEP 1996 technical report* (NCES 1999-452). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES 2001-509). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

American College Testing (ACT). (1995). *NAEP reading revisit: An evaluation of the 1992 achievement level descriptions*.

American College Testing American College Testing (ACT). (2010). *Developing achievement levels on the 2009 National Assessment of Educational Progress in Science for grades four, eight, and twelve: Process report.*

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*.

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). American Council on Education.

Bay, L., Dunn, J., Wonsuk, K., McGuire, L., & Sukin, T. (2012). *Developing achievement levels on the 2011 National Assessment of Educational Progress in grades 8 and 12 writing* [Technical report]. National Assessment Governing Board.

Beaton, A. E., & Gonzales, E. J. (1993). *Comparing the NAEP trial state assessment results with the IAEP international results*. National Academy of Education, Panel on the NAEP Trial State Assessment.

Beatty, A. S., Reese, C. M., Persky, H. R., Carr, P. (1996, April). *NAEP 1994 U.S. history report card: Findings from the National Assessment of Educational Progress*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main1994/96085.pdf

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 1–30). JAM Press.

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

Bollen, K. (1989). *Structural equations with latent variables*. John Wiley and Sons.

Bollen, K. (1993, November). Liberal democracy: Validity and method factors in cross-national measures. *American Journal of Political Science, 37*(4), 1207–1230.

Bourque, M. L. (1999, July). Report on developing achievement level descriptions for the 1996 NAEP science assessment. In The NAEP 1996 technical report (Appendix G, pp. 759-768). U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main1996/1999452.pdf

Bourque, M. L. (2009). *A history of NAEP achievement levels: Issues, implementation, and impact 1989–2009*. National Assessment Governing Board.

Bourque, M. L., & Byrd, S. (Eds.). (2000). *Student performance standards on the National Assessment of Educational Progress: Affirmations and improvements.* National Assessment Governing Board.

Braswell, J., & Haberstroh, J. (2004). *Report on the 2003 mathematics scale-anchoring study* [Technical report]. National Assessment Governing Board. https://www.nagb.gov/content/dam/nagb/en/documents/naep/np-Scale-Anchoring-Study-Report-05-12-2004-RED.pdf

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Praeger.

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R.K., Zenisky, A. L., & Wells, C. S. (2009). *Evaluation of the National Assessment of Educational Progress: Study reports*. U.S. Department of Education.

Byun, S. Y., Irvin, M. J., & Bell, B. A. (2015). Advanced math course taking: Effects on math achievement and college enrollment. *Journal of Experimental Education*, *83*(4), 439–468.

Campbell, J. R., Donahue, P. L., Reese, C. M., Phillips, G. W. (1996, January). *NAEP 1994 reading report card for the nation and the states: Findings from the National Assessment of Educational Progress and Trial State Assessment*. U.S. Department of Education, National Center for Education Statistics. https://eric.ed.gov/?id=ED388962

Cohen, J. (2005). AM statistical software (Version 0.06.03 Beta) [Computer software]. American Institutes for Research. http://am.air.org/default.asp

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage.

Dogan, E., Ogut, B., & Kim, Y. (2015). Early childhood reading skills and proficiency in NAEP eighth-grade reading assessment. *Applied Measurement in Education, 28*(3), 187-201.

Donahue, P., Beaulieu, N., Freund, D., & Pitoniak, M. (2009). *Report on the 2009 Reading achievement level and scale-anchoring study - Draft* [Technical Report]. Educational Testing Service.

Donahue, P., Pitoniak, M., & Beaulieu, N. (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*. Educational Testing Service.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 179-198). Springer.

Ebel, R. L. (1972). *Essentials of educational measurement*. Prentice Hall.

Erwin, B., Brown, D., & Mann, S. (2023, May 23). *50-state comparison: High school graduation requirements. Education Commission of the States.* https://www.ecs.org/50-state-comparison-high-school-graduation-requirements-2023/

Egan, K. L., Schneider, M. C., & Ferrara, S. (2011). The 6D framework: A validity framework for defining proficient performance and setting cut scores for accessible tests. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 275–292). Springer.

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice, 30*(4), 3-15.

Hambleton, R. K. (2001). Setting performance standards in educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 90–116). Lawrence Erlbaum Associates.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). American Council on Education; Praeger Publishers.

Hansche, L. N. (Ed.). (1998). *Meeting the requirements of Title I: Handbook for the development of performance standards*. U.S. Department of Education.

Improving America's School Act of 1994, 20 U.S.C. ch. 70, subch. I § 6301 et seq. (2018).

Ji, C. S., Rahman, T., & Yee, D. S. (2021). *Mapping state proficiency standards onto the NAEP scales: Results from the 2019 NAEP reading and mathematics assessments* (NCES 2021-036). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/pubsearch

Johnson, E. G. (1992). *The NAEP 1992 technical report*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Johnson, E. G., & Siengondorf, A. (1998). *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: Eighth grade results*. (NCES 98-500). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*.

Kahl, S. R., Crockett, T. J., DePascale, C. A., & Rindfleisch, S. L. (1995, April 18-22). *Setting standards for performance levels using the student-based constructed-response method* [Paper presentation]. American Educational Research Association 1995 Annual Meeting, San Francisco, CA, United States.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53-88). Erlbaum.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education; Praeger.

Kane, M. T. (2011). The errors of our ways. *Journal of Educational Measurement, 48*(1), 12–30.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Erlbaum.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-0317-7

Lane, S. & Marion, S.F. (in press). Validity in educational measurement. In Cook, L. & Pitoniak, M. (eds.) *Fifth Edition of Educational Measurement.* Oxford University Press.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996, June). Standard setting: A bookmark approach. In D.R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium presented at the Council of Chief State School Officers National Conference on Large-scale Assessment, Phoenix, AZ.

Linn, R. L. (1994). *Assessment-based reform: Challenges to educational measurement* [Monograph]. Educational Testing Service. https://www.ets.org/Media/Research/pdf/PICANG1.pdf

Loomis, S. C. (2018). *Anchor studies for analysis of NAEP achievement levels*. National Assessment Governing Board. www.nagb.gov/content/dam/nagb/en/documents/publications/achievement/Anchor-Studies-for-Analysis-of-NAEP-Achievement-Levels.pdf

Loomis, S. C., & Bourque, M. L. (Eds.). (2001). *National Assessment of Educational Progress achievement levels: 1992–1998*. National Assessment Governing Board.

Loomis, S.C., & Hanick, P.L. (2000). *Setting standards for the 1998 NAEP in civics and writing: Finalizing the Achievement Levels Descriptions*. National Assessment Governing Board. https://www.nagb.gov/content/dam/nagb/en/documents/publications/achievement/setting-standards-1998-naep-civics-writing-finalizing-descriptions.pdf

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13- 103). American Council on Education.

Lutkus, A.D., Weiss, A.R., Campbell, J.R., Mazzeo, J., & Lazer, S. *NAEP 1998 Civics Report Card for the Nation* (NCES 2000-457). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Mills, C. N. and Jaeger, R. J. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.) *Handbook for development of performance standards* (pp. 75–85). U.S. Department of Education; Council of Chief State School Officers.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Educational Testing Service, Policy Information Center.

Moran, R., Freund, D., & Oranje, A., (2012). *NAEP 12th grade preparedness research: Analyses relating Florida students' performance on NAEP to preparedness indicators and postsecondary performance* [Technical report]. Educational Testing Service. https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/statistical-relationships/Florida_Statistical_Study.pdf

Moran, R., Oranje, A., & Freund, D. (2012). *Establishing a statistical relationship between NAEP and SAT®* [Technical report]. Educational Testing Service. https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/statistical-relationships/SAT-NAEP_Linking_Study.pdf

Moyer, E. L., & Galindo, J. (2022). *Achievement level description review for the National Assessment of Educational Progress mathematics and reading assessments*. National Assessment Governing Board.

Moyer, E. L., & Galindo, J. (2023). *Achievement level description review for the National Assessment of Educational Progress grade 8 science, U.S. history, and civics assessments*. National Assessment Governing Board.

Mullis, I.V. (1993). *NAEP 1992-Reading Report Card for the Nation and the States: Data from the National and Trial State Assessments.* National Center for Educational Statistics.

National Academies of Sciences, Engineering, & Medicine (NASEM). (2017). *Evaluation of the achievement levels for mathematics and reading on the National Assessment of Educational Progress*. The National Academies Press.

National Assessment Governing Board. (2018, November 17). *Developing student achievement levels for the National Assessment of Educational Progress* [Policy document]. http://www.nagb.gov/content/dam/nagb/en/documents/policies/ALS-revised-policy-statement-11-17-18.pdf

National Assessment Governing Board. (2020, March 7). *The intended meaning of NAEP scores*. https://www.nagb.gov/content/dam/nagb/en/documents/policies/Intended-Meaning-of-NAEP.pdf

National Assessment Governing Board. (2021a). *Mathematics assessment framework for the 2026 National Assessment of Educational Progress*.

National Assessment Governing Board. (2021b). *Reading assessment framework for the 2026 National Assessment of Educational Progress*.

National Assessment Governing Board. (2022a). *Mathematics assessment framework for the 2022 and 2024 National Assessment of Educational Progress*.

National Assessment Governing Board. (2022b). *Reading assessment framework for the 2022 and 2024 National Assessment of Educational Progress*.

National Assessment Governing Board. (2022, March 3). *Assessment framework development* [Policy document].

https://www.nagb.gov/content/dam/nagb/en/documents/policies/assessment-framework-development.pdf

National Assessment Governing Board. (2022). *The Nation's Report Card: Reading and Mathematics Achievement Levels*. https://www.nagb.gov/content/dam/nagb/en/documents/naep/naep-day/2022/the-nations-report-card-reading-and-mathematics-achievement-levels.pdf

National Assessment Governing Board. (2023*). Resolution to encourage prioritization of NAEP linking studies.* https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/quarterly-board-meeting-materials/2023-08/11-resolution-on-naep-linking-studies.pdf

National Assessment of Educational Progress Improvement Act of 1988, 20 U.S.C. §§ 3401–3403 (2018).

National Center for Education Statistics. (2011). *NCES handbook of survey methods.* U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/pubs2011/2011609.pdf

National Center for Education Statistics. (2012a). *Statistical standards*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/statprog/2012/

National Center for Education Statistics. (2012b, July 5). *The setting of achievement levels.* U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/set-achievement-lvls.aspx

National Center for Education Statistics. (2013). *The Nation's Report Card: U.S. states in a global context: Results from the 2011 NAEP-TIMSS Linking Study* (NCES 2013–460). U.S. Department of Education, Institute of Education Sciences.

National Center for Education Statistics. (2024a). *The Nation's Report Card: Mapping state proficiency standards*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/studies/statemapping/

National Center for Education Statistics. (2024b). *The Nation's Report Card: NAEP achievement levels.* U.S. Department of Education, Institute of Education Sciences. http://www.nagb.gov/naep/NAEP-achievement-levels.html

National Center for Education Statistics. (2024c). *The Nation's Report Card: NAEP frameworks.* U.S. Department of Education, Institute of Education Sciences. https://www.nagb.gov/naep/frameworks-overview.html

National Center for Education Statistics. (2024d). *The Nation's Report Card: Reading and mathematics achievement levels* [Technical document]. U.S. Department of Education, Institute of Education Sciences. https://www.nagb.gov/content/dam/nagb/en/documents/naep/naep-day/2022/the-nations-report-card-reading-and-mathematics-achievement-levels.pdf

National Research Council. (1999). *Grading the Nation's Report Card: Evaluating NAEP and transforming the assessment of educational progress*. The National Academies Press.

Nebelsick-Gullet, L., & Fitzpatrick, S. (2016). *Developing achievement levels on the 2014 National Assessment of Educational Progress in grade 8 technology and engineering*

literacy: Process report [Technical report]. Pearson. https://www.nagb.gov/content/dam/nagb/en/documents/publications/achievement/developing-achievement-2014-naep-grade-8-tel-process-report.pdf

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement 14,* 3–10.

No Child Left Behind Act of 2001, 20 U.S.C. § 6311 et seq. (2002).

Ogut, B., Bohrnstedt, G., & Broer, M. (2015). *Imputing 12th-grade NAEP mathematics scores for the full HSLS sample*. American Institutes for Research.

Ogut, B., Bohrnstedt, G., & Broer, M. (2021). *College enrollment benchmarks for the NAEP grade 12 mathematics assessment* (AIR-NAEP Working Paper #2021-04). American Institutes for Research.

Ogut, B., Bohrnstedt, G., & Broer, M. (2023). *Updated college enrollment Benchmarks for the grade 12 NAEP mathematics assessment* (AIR-NAEP Working Paper #2023-03). American Institutes for Research.

Ohio Department of Education & Workforce. (2024). *Third grade reading guarantee* [Policy document]. https://education.ohio.gov/getattachment/Topics/Learning-in-Ohio/Literacy/Third-Grade-Reading-Guarantee/Third-Grade-Reading-Guarantee-Guidance-Updates-2024-1.pdf.aspx?lang=en-US

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. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main1994/96087.pdf

Phillips, G. (2014, February). *Linking the 2011 National Assessment of Educational Progress (NAEP) in reading to the 2011 Progress in International Reading Literacy Study (PIRLS)* [Technical report]. https://files.eric.ed.gov/fulltext/ED545246.pdf

Phillips, G., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993, April). Interpreting NAEP scales. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. https://files.eric.ed.gov/fulltext/ED361396.pdf

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*. Educational Testing Service.

Reckase, M. D. (2000). *The evolution of the NAEP achievement level setting process: A summary of the research and development efforts conducted by ACT*. ACT.

Schulz, E. M., & Mitzel, H. (2011). A Mapmark method of standard setting as implemented for the National Assessment Governing Board. *Journal of Applied Measurement, 12*(2), 165-93.

Sgammato, A., Lin, M., Jerry, L., Freund, D., Michel, R., & Oranje, A. (2016). *NAEP grade 8 academic preparedness research: Establishing a statistical relationship between the NAEP and EXPLORE® grade 8 assessments in reading and mathematics for Kentucky*

students [Technical report]. Educational Testing Service. https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/8th-grade/statistical-relationship/ky-preparedness-grade-8.pdf

Sgammato, A., Lin, M., Jerry, L., Freund, D., Michel, R., Xi, N., & Oranje, A. (2016a). *NAEP grade 8 academic preparedness research: Establishing a statistical relationship between the NAEP and EXPLORE® grade 8 assessments in reading and mathematics for North Carolina students* [Technical report]. Educational Testing Service. https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/8th-grade/statistical-relationship/nc-preparedness-grade-8.pdf

Sgammato, A., Lin, M., Jerry, L., Freund, D., Michel, R., Xi, N., & Oranje, A. (2016b). *NAEP grade 8 academic preparedness research: Establishing a statistical relationship between the NAEP and EXPLORE® grade 8 assessments in reading and mathematics for Tennessee students* [Technical report]. Educational Testing Service. https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/8th-grade/statistical-relationship/tn-preparedness-grade-8.pdf

Sireci, S. G., Hauger, J. B., Wells, C. S., Shea, C., & Zenisky, A. L. (2009). Evaluation of the Standard Setting on the 2005 Grade 12 National Assessment of Educational Progress Mathematics Test. *Applied Measurement in Education*, *22*(4), 339–358.

Wang, M. T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review, 33*(4), 304–340.

Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal, 50*(5), 1081–1121.

Weiss, A. (2003). *Report on 2002 Geography Scale-Anchoring Study - Draft* [Technical Report]. Educational Testing Service.

Williams, P.L., Reese, C.M., Lazer, S., & Shakrani, S. (1995). *NAEP 1994 Geography: A first look findings from the National Assessment of Educational Progress* [Technical Report]. Educational Testing Service under contract with the National Center for Education Statistics. https://www.govinfo.gov/content/pkg/GOVPUB-ED-PURL-gpo71163/pdf/GOVPUB-ED-PURL-gpo71163.pdf

Xi, N., Lin, M., Jerry, L., Freund, D., & Oranje, A. (2016a). *NAEP grade 12 academic preparedness research: Establishing a statistical relationship between the NAEP and ACT assessments in reading and mathematics for grade 12 Michigan students* [Technical report]. Educational Testing Service. https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/statistical-relationships/Preparedness-Grade-12-Michigan_508.pdf

Xi, N., Lin, M., Jerry, L., Freund, D., & Oranje, A. (2016b). *NAEP grade 12 academic preparedness research: Establishing a statistical relationship between the NAEP and ACT assessments in reading and mathematics for grade 12 Tennessee students* [Technical report]. Educational Testing Service. https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/statistical-relationships/Preparedness-Grade-12-Tennessee_508.pdf

Xi, N., Lin, M., Jerry, L., Freund, D., & Oranje, A. (2016c). *NAEP grade 12 academic preparedness research: Establishing a statistical relationship between the NAEP and SAT assessments in reading and mathematics for grade 12 Massachusetts students* [Technical report]. Educational Testing Service. https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/statistical-relationships/Preparedness-Grade-12-Massachusetts_508.pdf

Yee, D., Ogut, B., Bohrnstedt, G., Broer, M., & Circi, R. (2021). *Examining the relationship between STEM coursetaking in high school and grade 12 NAEP mathematics performance.* (AIR-NAEP Working Paper #2021-25). American Institutes for Research.

Zhang, J., Bohrnstedt, G., Park, B. J., Ikoma, S., Ogut, B., & Broer, M. (2021). *Mathematics motivation and the relationship with student performance: Evidence from the HSLS overlap sample*. (AIR-NAEP Working Paper #2021-03). American Institutes for Research.

Zhang, J., Bohrnstedt, G., Zheng, X., Bai, Y., Yee, D., & Broer, M. (2021). *Choosing a college STEM major: The roles of motivation, high school STEM coursetaking, NAEP mathematics achievement, and social networks*. (AIR-NAEP Working Paper #2021-02). American Institutes for Research.

Appendix A. Summary of Validity Evidence by Content Area and Evidence Source

Table A
Summary of Validity Evidence by Content Area and Evidence Source

READING

Evidence Source: Year of Publication	*Evidence Source:* Study	*Evidence Type:* Procedural	*Evidence Type:* Internal	*Evidence Type:* External
2024	*Reading and Mathematics Achievement Levels* [Technical document] (NCES, 2024d)	X		
2022	*Achievement Level Description Review for the National Assessment of Educational Progress Mathematics and Reading Assessments* (Moyer & Galindo, 2022)		X	
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* (Donahue et al., 2010)		X	
2009	*Report on the 2009 Reading achievement level and scale-anchoring study - Draft* (Donahue et al., 2009)		X	
1996	*NAEP 1994 Reading Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress and Trial State Assessment* (Campbell et al., 1996)		X	
1995	*NAEP Reading Revisit: An Evaluation of the 1992 Achievement Levels Descriptions* (ACT, 1995)		X	
1993	*Interpreting NAEP Scales* (Phillips et al., 1993)		X	
2021	*Mapping State Proficiency Standards Onto the NAEP Scales: Results From the 2019 NAEP Reading and Mathematics Assessments* (NCES 2021-036) (Ji et al., 2021)			X
2016	*NAEP Grade 8 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for Kentucky Students* [Technical report] (Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016)			X

Evidence Source: Year of Publication	*Evidence Source:* Study	*Evidence Type:* Procedural	*Evidence Type:* Internal	*Evidence Type:* External
2016	*NAEP Grade 8 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for North Carolina Students* [Technical report] (Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a)			X
2016	*NAEP Grade 8 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for Tennessee Students* [Technical report] (Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b)			X
2016	*NAEP Grade 12 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and ACT Assessments in Reading and Mathematics for Grade 12 Michigan Students* (Xi et al., 2016a)			X
2016	*NAEP Grade 12 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and ACT Assessments in Reading and Mathematics for Grade 12 Tennessee Students* (Xi et al., 2016b)			X
2016	*NAEP Grade 12 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and SAT Assessments in Reading and Mathematics for Grade 12 Massachusetts Students* (Xi et al., 2016c)			X
2015	*Early Childhood Reading Skills and Proficiency in NAEP Eighth-Grade Reading Assessment* (Dogan et al., 2015)			X
2014	*Linking the 2011 National Assessment of Educational Progress (NAEP) in Reading to the 2011 Progress in International Reading Literacy Study (PIRLS)* [Technical report] (Phillips, 2014)			X
2012	*Establishing a Statistical Relationship between NAEP and SAT®* (Moran, Oranje, & Freund, 2012)			X
2012	*NAEP 12th Grade Preparedness Research: Analyses Relating Florida Students' Performance on NAEP to Preparedness Indicators and Postsecondary Performance* (Moran, Freund, & Oranje, 2012)			

MATHEMATICS

Evidence Source: Year of Publication	*Evidence Source:* Study	*Evidence Type:* Procedural	*Evidence Type:* Internal	*Evidence Type:* External
2024	*Reading and Mathematics Achievement levels* [Technical document] (NCES, 2024d)	X		
2012	*Statistical Standards* (NCES, 2012a)	X		
2001	*National Assessment of Educational Progress: Achievement Levels (1992-1998) for Mathematics* (Loomis & Bourque, 2001)	X		
2022	*Achievement Level Description Review for the National Assessment of Educational Progress Mathematics and Reading Assessments* (Moyer & Galindo, 2022)		X	
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* (Pitoniak et al., 2010)		X	
2004	*Report on the 2003 Mathematics Scale-Anchoring Study* (Braswell & Haberstroh, 2004)		X	
1993	*Interpreting NAEP Scales* (Phillips et al., 1993)		X	
2021	*College Enrollment Benchmarks for the NAEP Grade 12 Mathematics Assessment.* (AIR-NAEP Working Paper #2021-04) (Ogut et al., 2021)			X
2021	*Examining the Relationship Between STEM Coursetaking in High School and Grade 12 NAEP Mathematics Performance* (AIR-NAEP Working Paper #2021-25) (Yee et al., 2021)			X
2021	*Mapping State Proficiency Standards Onto the NAEP Scales: Results From the 2019 NAEP Reading and Mathematics Assessments* (NCES 2021-036) (Ji et al., 2021)			X
2021	*Choosing a college STEM major: The roles of motivation, high school STEM coursetaking, NAEP mathematics achievement, and social networks.* (AIR-NAEP Working Paper #2021-02). (Zhang, Bohrnstedt, Zheng, et al., 2021)			
2021	*Mathematics Motivation and the Relationship with Student Performance: Evidence from the HSLS Overlap Sample.* (AIR-NAEP Working Paper #2021-03) (Zhang, Bohrnstedt, Park, et al., 2021)			X

Evidence Source: Year of Publication	*Evidence Source:* Study	*Evidence Type:* Procedural	*Evidence Type:* Internal	*Evidence Type:* External
2016	*NAEP Grade 8 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for Kentucky Students* (Sgammato, Lin, Jerry, Freund, Michel, & Oranje, 2016)			X
2016	*NAEP Grade 8 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for North Carolina Students* (Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016a)			X
2016	*NAEP Grade 8 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for Tennessee Students* (Sgammato, Lin, Jerry, Freund, Michel, Xi, & Oranje, 2016b)			X
2016	*NAEP Grade 12 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and ACT Assessments in Reading and Mathematics for Grade 12 Michigan Students* (Xi et al., 2016a)			X
2016	*NAEP Grade 12 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and ACT Assessments in Reading and Mathematics for Grade 12 Tennessee Students* (Xi et al., 2016b)			X
2016	*NAEP Grade 12 Academic Preparedness Research: Establishing a Statistical Relationship between the NAEP and SAT Assessments in Reading and Mathematics for Grade 12 Massachusetts Students* (Xi et al., 2016c)			X
2013	*The Nation's Report Card: U.S. States in a Global Context: Results From the 2011 NAEP-TIMSS Linking Study* (NCES 2013-460) (NCES, 2013)			X
2012	*Establishing a Statistical Relationship between NAEP and SAT®* (Moran, Oranje, & Freund, 2012)			X
2012	*NAEP 12th Grade Preparedness Research: Analyses Relating Florida Students' Performance on NAEP to Preparedness Indicators and Postsecondary Performance* (Moran, Freund, & Oranje, 2012)			X

SCIENCE

Evidence Source: Year of Publication	*Evidence Source:* Study	*Evidence Type:* Procedural	*Evidence Type:* Internal	*Evidence Type:* External
2012	*Statistical Standards* (NCES, 2012a)	X		
2023	*Achievement Level Description Review for the National Assessment of Educational Progress Grade 8 Science, U.S. History, and Civics Assessments* (Moyer & Galindo, 2023)		X	
2010	*Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Science for Grades Four, Eight, and Twelve: Process Report* (ACT, 2010)		X	
1999	*Report on Developing Achievement Level Descriptions for the 1996 NAEP Science Assessment* (Bourque, 1999)		X	
2013	*The Nation's Report Card: U.S. States in a Global Context: Results From the 2011 NAEP-TIMSS Linking Study* (NCES 2013-460) (NCES, 2013))			X

OTHER CONTENT AREAS

Evidence Source: Year of Publication	*Evidence Source:* Study	*Evidence Type:* Procedural	*Evidence Type:* Internal	*Evidence Type:* External
2012	*Statistical Standards* (NCES, 2012a)	X		
2011	*Developing Achievement Levels on the 2011 National Assessment of Educational Progress in Grades 8 and 12 Writing* [Technical report] (Bay et al., 2012)	X		
2001	The NAEP 1998 Technical Report (Allen et al., 2001)	X		
2023	*Achievement Level Description Review for the National Assessment of Educational Progress Grade 8 Science, U.S. History, and Civics Assessments* (Moyer & Galindo, 2023)		X	
2003	*Report on 2002 Geography Scale-Anchoring Study - Draft* (Weiss, 2003)		X	
1996	*NAEP 1994 Geography Report Card: Findings from the National Assessment of Educational Progress* (Persky et al., 1996)		X	
1996	*NAEP 1994 U.S. History Report Card: Findings from the National Assessment of Educational Progress* (Beatty et al., 1996)		X	