Technical Report # 2510

Rethinking "Standardization" for NAEP to Increase Equity and Access

Gerald Tindal PhD

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Reference: Tindal, G. (2025). *Rethinking "Standardization" for NAEP to Increase Equity and Access.* University of Oregon: Behavioral Research and Teaching. Technical Report 2510.

CONTENTS

EXHIBITS

ABBREVIATIONS AND ACRONYMS

ASL	American Sign Language
CRESST	Center for Research on Evaluation Standards and Student Testing
DBA	Digitally based assessment
DIF	Differential item functioning
DTF	Differential test functioning
EL	English learner
ES	Effect size
IDEA	Individuals with Disabilities Education Act
IEP	Individualized Education Program
IRT	Item response theory
LD	learning disability
LEP	Limited English proficiency (Note: LEP is a term used in older publications and has been updated to EL)
NAEP	National Assessment of Educational Progress
NASEM	National Academies of Sciences, Engineering, and Medicine
NCEO	National Center on Educational Outcomes
NCES	National Center for Education Statistics
NCLB	No Child Left Behind
NVS	NAEP Validity Studies
PISA	Programme for International Student Assessment
SBT	Scenario-based task
SD	Student with disabilities
TA	Typically achieving
TEL	Technology and engineering literacy
UD	Universal design
UDE	Universal design elements
UDL	Universal Design for Learning

INTRODUCTION

The National Assessment of Educational Progress (NAEP) is the largest nationally representative and continuing assessment of what students in the United States know and can do in various subjects. Since 1969, NAEP has provided a common measure of student achievement across the country. In service of this purpose, NAEP has been regularly administered, maintaining standardized conditions by holding to the general adage 'If you want to measure change, don't change the measure.'

However, this notion of standardization has been recently questioned given NAEP's transition to digitally based assessments (DBAs) and discussions about equity in assessment. The resulting dialogue has generated momentum for considering whether standardization for NAEP might be conceptualized in terms of 'the experience' rather than in terms of having 'everything the same.' Thus, the purpose of this paper is to explore how NAEP can rethink 'standardization' to generate a more equitable assessment.

Standardized testing in the United States has a long history that is traceable back to the common tests proposed and developed by Horace Mann, an educational reformer in the mid-19th century (Gallagher, 2003). Broadly speaking, a standardized test is "an assessment instrument administered in a predetermined manner, such that the questions, conditions of administration, scoring, and interpretation of responses are consistent from one occasion to another" (American Psychological Association, n.d.). Throughout the 20th century, standardized testing involved increasingly focused attention on controlling the conditions of testing. This traditional view emphasized sameness and comparability to ensure that valid interpretations could be made with an emphasis on consistency (reliability). In recent decades, the notion that measurement error is best controlled through highly standardized testing conditions has been challenged, and a new point of view of standardized tests is emerging. Several trends are contributing to this shift.

An initial trend was the emergence of Universal Design for Learning (UDL) and the use of accommodations in testing. UDL reflects adaptations in response to learner needs, identifying and removing barriers, and attention to learner strengths. According to Rose (2006), three guiding principles of UDL allow different ways for students to succeed that provide multiple means of representation, expression, and engagement. UDL is useful in facilitating the inclusion of students with disabilities (SDs), English learners (ELs), and ELs with disabilities in education systems because it highlights the need to provide students with multiple pathways to achieve learning outcomes. Similarly, UDL clarified the need for altering conditions to provide multiple pathways for students to access standardized tests—often called testing accommodations. Accommodations involve adaptations to test presentation, the environment, content, format, or administration conditions for test takers that do not alter the assessed construct.

A second trend affecting perceptions of standardized testing has been technology advancements. "Every year, technology usage becomes an increasingly more visible and fundamental part of K-12 education, and there is no turning back" (Ross, 2020, p. 2014). The effect has been a shift from teacher- to learner-centered activities, with an array of technological devices used in education. In particular, the advent of digital testing facilitates

the use of multiple representations in test content (e.g., text, video, audio), technology-enhanced item types, adaptive administration, and tools such as highlighters and linked dictionaries that can broaden student access to the testing process without the need for targeted accommodations.

Finally, there is a renewed and more urgent focus on equity in testing and the 'personalization' of the assessment experience (e.g., Hughes, 2023; Sireci, 2020). A theme of this trend is that, although the goal of standardized testing is to promote fairness through consistency, the testing conditions may interact with the personal characteristics of examinees to affect test performance in ways that are not construct relevant. Thus, more flexibility in standardization is necessary to account for the diversity of examinees assessed in today's world. Sireci (2020) coined the term 'understandardization' to represent this changing perspective. "It is important to note from the outset that the key change in moving from standardization to 'understandardization' is not the prefix 'under,' but rather the prefix 'understand' (Sireci, 2020, p. 101).

NAEP first encountered tensions between standardization and inclusion in the mid-1990s after the passage of the Education for All Handicapped Children Act (Public Law 94-142) in 1975 and the Individuals with Disabilities Education Act (IDEA) in 1990. As with state and district testing programs, it was important for NAEP to make adaptations so that reporting samples would include students assessed with accommodations, yet it was important to ensure that NAEP results are comparable to previous assessment cycles. In 1996, NAEP began to study the effect of assessment accommodations on NAEP results, and during the next 5 years initiated a transition in which NAEP official reporting samples would begin to include students assessed with accommodations.[1] The transition was complicated and included reporting results with split samples (i.e., students who were accommodated and not accommodated) in several subjects during the 2000 and 2001 assessment cycles. By 2002, the transition was complete, and NAEP offered accommodations for all assessment subjects as detailed in the current inclusion policy.[2]

A second challenge related to standardization arose in the early 2000s as NAEP began new testing methods and question types that reflected the growing use of technology in education. A series of NAEP research projects compared the performance of students using an online assessment with students who used a paper-and-pencil assessment (Sandene et al., 2005). In 2014 and 2015, NAEP piloted mathematics and reading assessments using Microsoft Surface Pro tablets, which included questions involving audio and/or video as the use of digital tools (such as an onscreen calculator) and scenario-based problems. NAEP officially transitioned from paper-based assessments to DBAs in mathematics and reading in 2017, which was accompanied by a bridge study to evaluate the effect of the mode of administration on performance and to allow for comparisons of the 2017 results to later assessments administered digitally, as well as to the earlier assessments administered on paper (Jewsbury et al., 2020). Although successful, the transition to DBAs revealed a further problem related to standardization because it was not possible to maintain common delivery devices across time. First, the Surface Pros were breaking down, becoming out of date, changing across time, and eventually no longer produced. Second, the model of the program

[1] For details related to this transition, see https://nces.ed.gov/nationsreportcard/about/history_inclusion.aspx.
[2] For details on NAEP's inclusion policy see https://nces.ed.gov/nationsreportcard/about/inclusion.aspx.

buying, maintaining, and replacing thousands of devices for each administration of NAEP was too costly to sustain. In response to this issue, a recent NAEP Validity Study (NVS) paper offered considerations related to device and interface features that might affect student performance in the NAEP testing program (Way & Strain-Seymour, 2021).

NAEP is currently facing a third challenge related to standardization as it considers adaptations to administration. Most prominent is the opportunity to eventually (a) make NAEP a device agnostic assessment in a Next-Gen eNAEP platform and (b) revise administration with reduced contact from NAEP field staff or even become a contactless assessment. Adaptations and other potential innovations in NAEP design, administration, and scoring were addressed in a recent National Academies of Sciences, Engineering, and Medicine (NASEM; 2022) report[3] from an expert panel, which was convened to "recommend innovations to improve the cost-effectiveness of NAEP while maintaining or improving its technical quality and the information it provides" (NASEM, 2022, p. 1).

These adaptations for NAEP are unprecedented and, on the surface, seem contradictory to the 'don't change the measure' adage. On the other hand, change also provides an opportunity for NAEP to create a more equitable assessment. How can NAEP take advantage of these opportunities but still maintain its primary purpose to provide "a fair and accurate measurement of student academic achievement and reporting of trends in such achievement in reading, mathematics, and other subject matter" (20 U.S. Code § 9622 [2021]).

The focus of this paper is to discuss research and possible adaptations for NAEP in the setting, administration, and scoring by extending Sireci's (2020) perspective on standardization: "In educational testing, *students* are the most important part of the measurement process, not the measure itself, or the measurement scale" (p. 100). With this orientation, the primary goal is to better understand testing conditions and how they interact with student characteristics, which may require flexibility. In Sireci's examples, culturally responsive assessments allow students to rely on their funds of knowledge through **translanguaging** (e.g., **bilingual test delivery systems**). Other illustrations address flexibility in the testing environment that allow students to take the test using their own equipment (e.g., computers, devices, and software), selecting their own passages or writing prompts, and adapting the language for taking tests. In this paper, test adaptations are examined that might allow more flexibility in NAEP test administration, citing relevant research and current practices to test adaptations.

This paper has seven major sections.

1. Defining test adaptations, both within the *Standards for Educational and Psychological Testing*; hereafter "*Standards*" (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014) and as practiced within the NAEP program: accommodations, designated supports, and universal designs.
2. Summaries of the research previously conducted on a particular type of test adaptation, covering the extensive research on accommodations, both specific to NAEP and in general, as relevant to large-scale testing programs to establish a more expansive view of accommodations as measured by the significance of difference and consistency of outcomes within and across performance levels.

[3] Future references of this work will be the NASEM report.

3. Focus to an even broader and more practical view of test adaptations by referencing state policies and practices as well as the consistency of adaptations across testing platforms.
4. Understand test adaptations for students beyond labels and categorical characteristics to understand their impact. This issue is critical, given the common lament by researchers that student samples often are only vaguely described.
5. Speculations beyond research and practice, considering both function and format of the process, with two specific explications. An example in writing focuses on constructs and applies this logic to the content and constructs of reading. Three specific questions are posed for defining test adaptations, emphasizing universality for improving equity and access to a more diverse population of students.

Within each section, recommendations are offered that NAEP may consider that provide greater flexibility in administration and measurement in the service of increasing equity and access; these recommendations are preceded by the abbreviation '**Rx**.' Finally, a summary of these recommendations is presented, along with suggestions for future research.

STANDARDIZATION, EQUITY, AND ACCESS

Expanding standardized test protocols may enhance equity and access, but, at some point, comparability (across time or across students) may be compromised. Therefore, any consideration of standardization needs to be in concert with the *Standards* (AERA et al., 2014). Furthermore, expansions in testing programs should be consistent in terminology, whether NAEP or state testing accountability systems. Finally, with this terminology explicated, the specific adaptations with the NAEP testing program require documentation because many adaptations have occurred in the past 30 years.

Standards for Educational and Psychological Testing

Our initial perspectives are guided by the *Standards* (AERA et al., 2014) to allow further discussion in a consistent manner with generally accepted guidelines. The *Standards* note the following:

> Although standardization has been a fundamental principle for assuring that all examinees have the same opportunity to demonstrate their standing on the construct that a test is intended to measure, sometimes flexibility is needed to provide essentially equivalent opportunities for some test takers. In these cases, aspects of a standardized testing process that pose no particular challenge for most test takers may prevent specific groups or individuals from accurately demonstrating their standing with respect to the construct of interest. For example, challenges may arise due to an examinee's disability, cultural background, linguistic background, race, ethnicity, socioeconomic status, limitations that my come from aging, or some combination. Of these and other factors, in some instances, greater comparability of scores may be attained if standardized procedures are changed to address the needs of specific groups or individuals without any adverse effects on the validity or reliability of the results obtained (p. 51).
>
> Adaptations to the original test design, administration, or response can hopefully increase access to the test for a broad range of individuals. Two more specific terms, however, need to be invoked to determine the effects from such adaptations:
>
> - **Accommodation**: An adaptation that maintains score comparability.
> - **Modification**: An adaptation that results in incomplete or partial measurement of a construct.

In distinguishing these terms, the *Standards* specify that accommodations be made only under a clear specification of who should receive them and how they are to be made. The characteristics of individuals and relevant subgroups need articulation with reference to the construct and the test design, development, administration, and scoring to remove critical barriers while not compromising valid interpretations. Instructions also should be clear in test administration, with specification of "instructions to test takers, time limits, form of item presentation, use of devices with balance between flexibility and potential to jeopardize test score interpretation (based on evidence)" (AERA et al., 2014*,* p. 114). Irrespective of the type of accommodation, qualified personnel need to use a formal decision-making process, which would include "policies and procedures for assigning and using accommodations in the administration, scoring, and reporting of educational assessments" (p. 192) with "the

presence of manuals and training materials" (p. 200). Such accommodations can address adaptations in **setting**, **administration** (including qualifications of administrators, time needed, presentation, interface/engagement), and **response** (including the scoring protocols followed), requirements that remove construct-irrelevant barriers and support valid interpretations related to "individual test-takers' needs (e.g., cognitive, linguistic, sensory, physical, and do not change the construct)" (p. 67). Of course, multiple accommodations can be implemented that span these three categories (Thompson et al., 2002).

This flexibility relies on validation anchored to interpretations based on evidence confirming or disconfirming intended interpretations of test scores and use. Furthermore, this process involves evaluating this support by referring to the constructs being measured, in which the evidence may underrepresent the construct (referred to as **construct deficiency**) or may be influenced by processes extraneous (irrelevant) to the construct (referred to as **construct irrelevant variance** or **construct contamination**). In the validation process, evidence must be collected and integrated from the content of the measures, cognitive processes invoked in the measurement process, internal structures of the measures, relationships with other variables (and measures) that can be both concordant or discordant, and consequences from the measurement process. In the end, the emphasis rests on construct representation and the validity of interpretations.

In further and more specific reconsideration of standardization, the *Standards* (AERA et al., 2014) introduce two additional terms that are important to consider: **fairness** and **accessibility**. Standardization should be suitably anchored to fairness in testing, with individual test takers equitably treated first and foremost, as well as reflect characteristics of the measures as (a) being unbiased, (b) providing access to the construct being measured, and (c) serving as the basis for score interpretations that are consistent with the intended use. An important caveat is that access may need to balance several concurrent characteristics of test takers that interact with "contextual features of the testing situation" (p. 53). For example, English language proficiency may possibly relate to cultural experiences and socioeconomic status (see Srikanth, 2022). Fairness in testing is organized within the *Standards* by test content, test context, test response, and the opportunity to learn.

> Professionals may be justified in deviating from standardized procedures to gain a more accurate measurement of the intended construct and to provide more appropriate individual decisions. However, for other contexts and uses, deviation from standardized procedures may be inappropriate because they change the construct being measured, compromise the comparability of scores or use of norms, and/or unfairly advantage some individuals (AERA et al., 2014, pp. 53–54).

In addition to fairness, standardization needs to address accessibility as identified in the *Standards* (AERA et al., 2014): the need to allow all targeted test takers to show their status without either advantage or disadvantage from individual characteristics such as age, disability, cultural background, race/ethnicity, gender, or language. Accessibility…

> demands that the test developers be clear on the construct(s) being measured, including the target of the assessment, the purpose for which scores will be used, and the characteristics of the examinees and subgroups of the intended test population that could influence access (AERA et al., 2014, p. 50).

Definitions of Test Adaptations and Student Populations

Because the *Standards* (AERA et al., 2014) require standardized procedures for implementing accommodations (including who is eligible to receive them and how to administer them) so that comparable scores can be maintained, it is important to be clear on their unique characteristics and how they differ from either designated support features or universal design. However, distinctions among accommodations, designated supports, and universal design are not clear when used by NAEP, researchers, or state education agencies. In general, definitions are categorical with little attention to the specific criteria for placement of the test adaptation into the category. In this section, recommendations clearly distinguish adaptations in the test administration or environments that are accommodated, designated, or universal[4].

Accommodations have been consistently defined as adaptations to the test administration that do not affect score use or interpretation. Both researchers and practitioners have uniformly endorsed this definition. Typically, the choice of implementing an accommodation is dictated by the team of professionals involved in developing the student's individualized education program (IEP) or Section 504 plan. In NAEP and the Smarter Balanced consortium of states, accommodations may be **embedded** (in the digital environment) or **not embedded** (outside the digital environment).

Designated support is a term used by the Smarter Balanced consortium of states.

> Although these tools are available to all students, educators may determine that one or more might be distracting for a particular student and thus might indicate that the tool should be turned off (or not used) for the administration of the assessment to the student (Smarter Balanced, 2021, p. 9).

"The designated supports described in this section are not modifications but yield valid scores that count as participation in assessments that meet the requirements of ESSA [Every Student Succeeds Act] when used in a manner consistent with the Guidelines" (Smarter Balanced, 2021, p. 13). Smarter Balanced also makes the further distinction of designated support features being embedded (within the digital platform) or not embedded (outside the digital platform). For Smarter Balanced, this decision to provide a designated support is guided by an *Individual Student Assessment Accessibility Profile*. As noted earlier, NAEP addresses embedded (within the test delivery system) and not embedded (outside the test delivery system) only as features of accommodations.

Universal design became quite popular with the work of CAST (formerly known as the Center for Applied Special Technology) and the publication of *A Practical Reader in Universal Design for Learning* (Rose, 2006). As noted earlier, universal design provides students with multiple means for representation, expression, and engagement. The principles of universal design consider multiple ways for material to be presented using various media (print and digital), the provision of scaffolds to access material, flexible methods for teaching and multiple examples of concepts to be learned, allowance of student choice and customization of material to fit diverse needs, motivational strategies to ensure student engagement, and

[4] Three types of adaptation are considered, though when the term test "change" is used within a quote, the language from the author is used.

effective deployment of technology (Rose, 2006). Both NAEP and Smarter Balanced include universal design in the description of allowable test adaptations, although they differ on the specifics in both the features and their definitions.

Note: Even though this paper focuses extensively on SDs and ELs, it is likely that rethinking the word "standardization" with these two populations can lead to better understanding for a full range of students: those from socioculturally diverse groups, impoverished backgrounds, geographically distributed areas, and multilingual histories. By broadening the standardization of test setting, administration, and scoring to be inclusive of this broader range of students, the NAEP testing program can make the testing experience relevant for them. In the end, understanding and changing NAEP practices should lead to greater student participation from diverse groups, allow better understanding of the constructs being assessed, and connect to student experiences (including classroom practices used in teaching and learning).

Current NAEP Accommodations Practices

By starting with NAEP policies and practices, evidence-based practices are emphasized, both within the NAEP testing program and in general for large-scale testing programs. This research is separated into studies addressing SDs and ELs. Finally, because of the importance of understanding specific populations, sampling populations are considered as a final topic in reviewing accommodations with NAEP. It is important to realize that only a few NAEP test adaptations have research supporting them.

NAEP has an extensive history of deploying accommodations, going back to 1996 in mathematics and 1998 in reading and science, for ELs and SDs.

> NAEP incorporates inclusive policies and practices into every aspect of the assessment, including selection of students, participation in the assessment administration, and valid and effective accommodations. . . . Just like any other student, SD and EL students are selected to participate in NAEP. Within each selected school and grade to be assessed, students are chosen at random to participate in NAEP. Regardless of race/ethnicity, socioeconomic status, disability, status as an English learner, or any other factors, every student has the same chance of being selected, because NAEP is administered to a sample of students who represent the student population of the nation, and for state level tests, of each individual state (National Center for Education Statistics [NCES], n.d.).

NAEP test adaptations are organized into four categories:

- Accommodations for both SDs and ELs
- Accommodations designed specifically for SDs
- Accommodations appropriate for ELs
- Universal design features built into computer-based assessments (appropriate for all students) in all areas (mathematics, reading, science, writing (DBA), civics, economics, geography, U.S. history, music and visual arts, and writing.

Accommodations for both SDs and ELs refer to the following:

- **Setting**: extended time, small group or one-on-one, one-on-one, and breaks during testing
- **Administration**: directions only read aloud in English, test items read aloud in English occasionally or most/all the time (but not in reading)
- **Response**: None

Accommodations that are specific to SDs have the following characteristics:

- **Setting**: Must have an aide present in the testing room; preferential seating
- **Administration**: calculator, large print version of the test (music only but not visual arts), magnification, use of template/special equipment, cueing to stay on task, presentation in Braille (not in science), presentation in American Sign Language (ASL; not in reading)
- **Response**: responds orally to a scribe (not in writing before DBAs; paper and pencil for performance-based assessments), in Braille, or in ASL (not in music and visual arts; TEL [technology and engineering literacy]; or writing)

For ELs, the focus of NAEP accommodations is on language adaptations:

- **Administration** (a.k.a. presentation): using a bilingual dictionary without definitions in any language (not in reading or writing DBAs or before that, performance-based assessments; paper and pencil), directions read aloud only in Spanish (not in TEL), Spanish/English version of the test (not in Grade 12) only in mathematics, science, and civics-economics-geography-history, test items read aloud in Spanish only in mathematics (but not Grade 12 mathematics), science, and civics-economics-geography-history

Universal design features in mathematics, science, reading, and TEL include the following:

- **Setting**: small group, one-on-one
- **Administration**: zooming, text-to-speech (English) for directions only, text-to-speech (English) occasionally or most or all (but not for reading), volume adjustment, closed captioning
- **Response**: use a computer/tablet to respond, color contrast (mathematics, science, and TEL accommodation for reading), scratchwork/highlighter capability, eliminating capability

Universal design elements for all students in DBA used for writing 2011 and TEL 2013:

- **Setting**: small group, one-on-one
- **Administration**: adjusting font size, directions occasionally read aloud only in English (text to speech), test items occasionally read aloud in English (text to speech), test items mostly or always read aloud in English (text to speech), adjusting contrast or colors, highlighter tool
- **Response:** using a computer or typewriter to respond, eliminating answer choice tool

Exhibit 1 compares universal design elements (UDEs; for setting, administration, and response) with accommodations offered within and outside a digitally based environment. For example, in the first row, 'individual testing experience' is a UDE, but a "separate location" is an accommodation outside the test delivery system. As noted later in this paper, better justification is necessary when classifying an adaptation as an accommodation. For example, it is difficult (if not impossible) to describe a situation in which testing could be conducted individually without being in a separate location. Furthermore, in the DBA environment, many adaptations with read aloud can be carefully controlled (in English or Spanish), which then allows activation of this feature as a universal design feature.[5]

Exhibit 1. Summary Comparison of Universal Design and Accommodations Within and Outside Test Delivery

	Accommodations within test delivery	Accommodations outside test delivery	NAEP UD elements
Setting	• Extended time	• Breaks • Separate location • Familiar person • Preferential seating	• Individual testing experience
Administration	• Magnification • Low mobility version of the test • Calculator version of the test • Hearing impaired version of the test • Directions only translated to Spanish • Directions read aloud (text-to-speech; Spanish) • Spanish/English version of the test • Read aloud (text-to-speech; Spanish): occasionally, most, or all	• Uses template • Special equipment • Cueing to stay on task • Directions only presented in ASL • Presentation in ASL • Braille version of the test	• Zoom • Directions read aloud (text-to-speech; English) • Directions clarified/explained • Read aloud (text-to-speech; English): occasionally, most, or all • Color theming • Volume adjustment • Closed captioning
Response	• None	• Scribe • Response in ASL • Braille version of the test	• Use a computer/tablet to respond • Scratchwork paper • Highlighter capability • Eliminating capability

Note. NAEP = National Assessment of Educational Progress; UD = universal design; ASL = American Sign Language.

[5] See https://nces.ed.gov/nationsreportcard/about/accom_table.aspx for a full comparison of accommodations and UDEs.

RESEARCH ON TEST ACCOMMODATIONS

The research on test accommodations has a long history in both the NAEP testing program and state testing programs. This research became particularly important with the implementation of No Child Left Behind (NCLB) and primarily targeted SDs and ELs, even though both groups reflect considerable diversity in demographics and backgrounds. This research has been conducted on a wide range of accommodations, has investigated many different moderating variables, and has documented varying outcomes.

NAEP Accommodations Research on Students With Disabilities

Beginning in 1996, accommodations for SDs participating in NAEP have been documented in successive years across different subject areas. Lutkus and Mazzeo (2003) conducted one of the first studies with NAEP (using the 1998 Reading Assessment) for SDs tested with and without accommodations. Their results indicated no significant differences in average reading scale scores between the two groups, accommodated and not accommodated (overall or by sex, racial/ethnic group, or grade). However, allowing accommodations increased inclusion rates, with states varying in the percentage of participating students from 1% to 5%.

> Offering accommodations in state NAEP to students who receive them in their regular classroom assessments will increase inclusion in some states and other jurisdictions, but the magnitude of the increase varies across jurisdictions. At Grade 4, the increase in inclusion of special-needs students and the provision of accommodations was associated with lower average scale scores in nine states, but not in the nation. At Grade 8, there was no pattern of statistically significant differences by accommodation status (Lutkus & Mazzeo, 2003, p. 13).

The main limitation of this study was that the study provided little information on the specific accommodations for individual students who may have received them. Rather, the use of accommodations was bundled:

- One-on-one testing
- Small-group testing
- Extended time
- Oral reading directions
- Signing directions
- Magnifying equipment
- Transcription of responses

Note that neither reading (of questions or text) nor bilingual dictionaries were allowed. Furthermore, other than sex, race/ethnicity, and grade, the study provided no information about the disability type or the EL status (other than having received instruction in English for 2 years) for the students.

Tindal and Ketterlin-Geller's (2004) review (of K–12 mathematics tests) considered several specific accommodations:

- Small-group setting
- Calculators
- Reading problems aloud
- Extended time

On the 2000 NAEP Mathematics Assessment, "more students participated in the NAEP when accommodations were permitted. In 4th grade, students taking the NAEP with accommodations scored significantly lower than students not using accommodations" (Lukus & Mazzeo, 2013, p. 8). However, in their further review of other research on accommodations in large-scale mathematics tests, they reported that complex relationships likely exist between the accommodation and the outcome: Calculators may be effective for some problems but not others, read aloud may be more beneficial for younger versus older students, and no **differential boost** may occur with extra time. They argued that the outcomes from specific accommodations are a function of the characteristics of individual test items and the skills of the students, not their disability.

Tavani (2007) analyzed the 2000 NAEP Mathematics Assessment database and found no effects from using accommodations for students with learning disabilities (LDs), although grade level, gender, and race/ethnic background were influential predictors of performance. Ricci (2015) investigated reading item aloud accommodations on the 2011 Reading Assessment restricted data set with Grade 4 students from three states (New York, New Jersey, and Connecticut). "The mean scores for those who received the read-aloud accommodation were lower than those who did not receive the accommodation" (Ricci, 2015, p. iii). Comparisons using effect sizes reflected large negative values across the three states (from −.46 to −1.02) and indicated better performance occurred when the read aloud accommodation was not administered. Although the study was nonexperimental, the findings are important because (a) they represent students identified for this accommodation according to NAEP protocols and (b) students were compared with same grade students who did not receive this accommodation. In the most recent study, Tam's (2020) investigation of read aloud with extended time for the 2013 Reading and Mathematics Assessments (for SDs in Grades 4 and 8), using matched samples of students receiving and not receiving the accommodation, showed that "students with disabilities **benefitted** [emphasis added] from the read-aloud accommodation. The extended time accommodations appeared to have benefitted the 4th grade students and not the 8th grade students" (Tam, 2020, abstract).

NAEP Accommodations Research on English Learners

The research specifically on NAEP accommodations for ELs is dated, occurring in the late 1990s, with only a few studies in the 2000s. Most of the early research was from Abedi and associates through the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

In one of the first studies, Abedi et al. (1997) used the 1990 and 1992 main NAEP assessments with ELs, which serves as a model for its breadth of variables analyzed, including the role of language in mathematics items (linguistic complexity and length of items) as well as language background and student perceptions. Not surprisingly, they reported lower mathematics proficiency scores (and more omitted items) for students who

spoke a language other than English, particularly for longer items. Nevertheless, when comparing original items with linguistic modifications, no statistically significant differences were found overall, but a slight advantage was found for students taking low and average mathematics classes. These combined effects of accommodation strategies and students' background characteristics were more powerful predictors of students' performance than either of them separately (Abedi, 1999; Abedi et al., 2000). For example,

- Extra time was not effective for students enrolled in more basic mathematics classes.
- The use of a glossary was not effective overall, but when combined with extra time, the glossary was effective for all students.
- Linguistic modification of test items was uniquely effective for ELs.

In contrast, Abedi, Hofstetter, et al. (2001), using the 1996 Grade 8 Bilingual Mathematics booklet, reported that most accommodations were effective for all students (ELs and non-ELs) with the exception being modified English (which was most effective with ELs). For their entire sample, extra time resulted in an increase of 1 point, glossary and extra time resulted in a 2-point increase, and linguistically modified items resulted in a differential boost (narrowing the difference between ELs and non-ELs). Again, student characteristics were important considerations in the effectiveness of the accommodations. (e.g., students who were better readers achieved higher mathematics scores). Similarly, Abedi, Lord, et al. (2001) compared no accommodation with a customized dictionary and a glossary, reporting some accommodations benefited all students.

In a summary of research on accommodations with ELs, Abedi, Hofstetter, & Lord (2004) addressed the policy context using test accommodations, the population and definition of ELs, the relationship between language proficiency and test performance, the definition of accommodations and their use, and the empirical research on accommodations—all of which are key issues in deciding among accommodation options, as well as determining the implications for education policy and practice. Their review included the following most frequently investigated accommodations for students with limited English proficiency (LEP):

- Testing in the student's native language
- Linguistic modifications of test items
- Extra time
- Published dictionaries
- Glossary and customized dictionaries
- Oral administration

Abedi & Hejri (2004) had four main conclusions:

- **Translating test items** from English into other languages is not an effective accommodation strategy when students have studied the subject in a classroom using English ("the language of assessment should match students' primary language of instruction" [p. 17]).
- **Some accommodations are more effective with certain student groups** than with others, depending on background factors such as English reading proficiency and time spent in the United States.

- **Reductions of low-frequency vocabulary and complex language structures** (incidental to the content knowledge being assessed) narrows the performance gap between ELs and other students. This conclusion, however, is somewhat in contrast to Abedi and Hejri (2004), who reported that accommodations used in the 1996 Main NAEP failed to reduce performance gaps between LEP and non-LEP students in general, although many accommodations had few students receiving them.
- **Customized dictionaries** help ELs while not affecting the scores of English-proficient students.

In summary, NAEP began incorporating accommodations in the late 1990s; soon after, researchers began addressing their effects using available data. For both SDs and ELs, the accommodations addressed test adaptations in setting, administration, and response, although the bulk of adaptations for ELs focus on linguistic features inherent in administering the test. This early research was less concerned with interactions or differential boosts than simply their effects on the target groups. Since then, accommodations appropriate for both groups have been organized into easily available tables. With large-scale testing programs being emphasized at about this same time (2000 and beyond), researchers also began to investigate these NAEP-identified test adaptations, as well as many more that were specific to states.

Meta-Analytic Research on Accommodations for SDs

Since the early research on NAEP-specific accommodations, another extensive body of research on accommodations has been conducted in the past 30 years. Much of this initial research was with statewide testing programs given the legislative dictates of NCLB focused on full participation (and proficiency) of student populations throughout the first decade of the 2000s. This research provided an extensive analysis on possible adaptations that testing programs can make to accommodate SDs and ELs—the two most prominent groups represented. With hundreds of published studies, researchers eventually turned to summaries and meta-analyses to codify general trends. In this section, the focus is only on these publications (not primary studies), particularly with our interest on standardization*,* not only test accommodations. In this section, research is presented from three studies conducted from 1999 through 2005, one a formal meta-analysis and two as summaries. Then, reviews are provided from an additional seven meta-analyses (after 2010) conducted on accommodations for students with and without disabilities. The reason for this division is that NCLB was enacted early in the decade, and by the time implementation was finalized, the more recent research was more prominent.

In displaying the data from the original studies, the tables are adapted into two ways: (a) only relevant values and variables are displayed (not the entire table), and (b) the cell entry of accommodations by student samples displays a symbol for a binned range of effect size values using a 'consumer report' view. Values from Cohen (1988, 1992) are presented in which effect sizes of .20 or less are considered small and displayed with an empty circle, .21 to .79 are considered medium (the middle being .50) and displayed with a partially shaded circle, and effect sizes of .80 or greater are considered large and displayed with a fully shaded circle. These values are more conservative than Nye (2019), as noted later in this paper.

Chiu and Pearson (1999) provided the first published meta-analysis and established a general framework for reporting on test accommodations for special education (SDs)[6] and ELs in 30 research studies examining extended time or unlimited time, assistive devices, presentation formats, response formats, setting of tests, radical accommodations, and combinations of accommodations. They analyzed outcomes on the effects of accommodations for several student subgroups: 'garden variety' disabilities, LDs, multiple disabilities, physical disabilities, visual impairment, and no disabilities. The designs included repeated measures with and without a comparison group and equivalent group designs. Most of the effects were small (except for presentations, radical accommodations, and setting of tests). Note that with the effects of providing the accommodations, students in general education benefitted more than those receiving special education services. They also noted, however, that "the accommodation effects varied substantially within different types of accommodations, different ways of identifying target populations, and the grade levels of the students" (Chiu & Pearson, 1999, p. 16).

> Almost half (47%) of the accommodations provided extended time or unlimited time. *Setting of tests* (2%) and *response format* (2%) were the least frequently investigated accommodations. Four other frequently examined accommodations included *assistive device* (9%), *combination of accommodation* (11%), *presentation formats* (13%), and *radical accommodation* (17%) (Chiu & Pearson, 1999, p. 14).

Exhibit 2 shows a summary of effect sizes for this meta-analysis.

Exhibit 2. Effect Sizes on Test Scores for Accommodations (Chiu & Pearson, 1999)

Accommodation	Special Education and ESL/LEP	General Education
Assistive devices		
Combination of accommodations		
Presentation formats		
Radical accommodations		
Response format		
Setting of tests		
Timing of tests		

Note. ~~Adapted~~ Values ~~reflected~~ Table 3 (p. 15) from in Chiu and Pearson (1999).

Tindal and Fuchs (2000) summarized research on the following specific accommodations:

- Timing and scheduling of testing
- Test settings
- Computer presentation of tests
- Examiner familiarity
- Multiple adaptations in presentation

[6] Students receiving special education services included students with 'garden variety' disabilities, hyperactive students, students with learning disabilities, and students with no formal status.

- Dictation to a proctor or scribe
- Using an alternative response
- Marking responses in test booklets
- Working collaboratively with other students
- Using word processors
- Using calculators
- Reinforcement
- Instruction on test-taking strategies

They described all the subgroups defined by the primary authors; the grade level of students was reported, with many studies at postsecondary levels. Academic achievement was summarized for mathematics, reading, writing, listening, social studies, and science. A variety of tests and measures were considered, not just state accountability tests. For each type of accommodation, they reported outcomes from specific studies organized by subject and test. In the end, they also provided a qualitative summary noting methodological soundness by referencing three types of designs: **descriptive** (logical analysis of the nature and severity of the disability along with the accommodation), **comparative** (retrospective analysis of data sets to determine effect of accommodations), and **experimental** (prospective research designs to determine differential boosts among groups). They concluded this review by asking six critical questions:

- Are the findings relevant for classroom practice and instructional focus?
- Who has been studied and what tests have been used to study adaptations and for which decisions?
- How well designed is the research on test adaptations and can the results be used?
- Has the research been conducted correctly (with reliability and validity established)?
- Does the research on test adaptations help establish construct validity (construct of the measure, individual need, and differential outcomes)?
- When research is put into practice, what are the consequences at a systems level, for state practices, and in teacher knowledge?

Sireci et al. (2005) extended this research by reviewing 28 empirical studies on the effects of accommodations, including the following:

- Presentation: oral, paraphrasing, technological, Braille/large print, ASL, encouragement, cueing, spelling assistance, and use of manipulatives
- Timing: extended time, multiple days/sessions, and separate sessions
- Responses: use of scribes, use of booklet versus answer sheet, marking task booklet to maintain place allowed, and transcription
- Setting: separate room and no specifics listed

They stated that

> the interaction hypothesis needs qualification. When SDs exhibit greater gains with accommodations than do their general education peers, an interaction is present. When the gains experienced by SDs are significantly greater than the gains experienced by their general education peers, the fact that the general education students achieved higher

> scores with an accommodation condition does not imply that the accommodation is unfair. It could imply that the standardized test conditions are too stringent for all students (Sireci et al., 2005, p. 481).

As these and later authors concluded, any kind of accommodation needs to be interpreted in the context of sample characteristics, grade, level, subject matter, and study design. Later, as the research findings accumulated, these two populations (SDs and ELs serving as the target groups and general education students serving as the reference group) were further separated and refined in different reviews and meta-analyses.

Finally, three early summaries of research on accommodations were published and disseminated by the National Center on Educational Outcomes (NCEO), which has become the primary repository of information on accommodations, including bibliographies on both general education accountability tests and alternate assessments. In four primary reports, several variables from research on accommodations were documented in successive time intervals: Thompson et al. (2002) published summaries of accommodation from 1999 to 2001; Johnstone et al. (2006) reviewed accommodations from 2002 to 2004; Zenisky and Sireci (2007) published reviews of accommodations from 2005 to 2006; and Cormier et al. (2010) reviewed accommodations from 2007 to 2008. Their most important conclusions were as follows:

- State policies have varied in their explicit reference to acceptable accommodations.
- Research on accommodations has deployed various methodologies, from using experimental and quasi-experimental studies to data collection using surveys of perceptions, IEP analysis, and product evaluations. Further, variation occurred in the type of accommodations deployed, the test forms used between accommodated and nonaccommodated groups, and the populations studied. All these variations in methodology and implementation prevent the making of generalized statements.
- Several outcome measures have been deployed in various content areas, including mathematics, reading/language arts, science, writing, and social studies. Relatedly, this variation in content areas often relates to variation in population samples (e.g., elementary, middle, and high school students).
- The effectiveness of accommodations has varied with few consistent outcomes. For example, in the latest summary from NCEO, covering research from 2007 to 2008, Cormier et al. (2010) reported that outcomes from accommodations have (a) increased performance for only the targeted group (representing an interaction effect) and for both groups, but more so for the target group (representing a differential boost), (b) been neutral, or (c) been detrimental (decreasing performance).

The most critical information for our interest in explicating standardization focuses on both the accommodation types and population samples.

- Fifteen different accommodations have addressed presentation, timing/scheduling, response, technological aids, and multiple accommodations. For example, accommodations have included oral presentation, extended time, computer administration, and technological aid (computer and dictionary).

- The most frequently studied group of SDs has been those with LDs, although other disabilities have been studied, including cognitive disability; emotional/behavioral disability; communication; reading or mathematical disabilities; and other disabilities to include physical and sensory disabilities, autism, attention deficit–hyperactivity disorder (ADHD), health impairments, and multiple disabilities.
- The main outcome from one of these NCEO reports (and echoed in the other reports) has been that the findings continue to be contradictory. "Research indicated that accommodations were either beneficial or not beneficial for students with disabilities. Likewise, researchers did not reach consensus on whether accommodations change the construct of the item assessed" (Johnstone et al., 2006, p. 15).

Concluding Perspectives on Early Accommodations Research

By the end of the first decade of 2000, researchers generally agreed that accommodations need to have different impacts (a simple interaction effect) for subgroups of students (e.g., work with the targeted group of students needing it and not work with a comparison group of students not needing it). The change in performance for the target group (usually SDs or ELs) should be significantly positive with no such (positive) change occurring in the reference group (usually students in general education). Fuchs and Fuchs (1999) enhanced this simple interaction, however, by requiring that the outcome provide a differential boost, wherein performance may improve in both groups (target and reference) but should be greater for the target group.

With this initial backdrop of research on accommodations, six more recent meta-analyses/reviews (within the past 2 decades) have been conducted on the effects of accommodations for SDs.

- Vanchu-Orosco (2012)
- Gregg and Nelson (2012)
- Harrison et al. (2013)
- Cawthon and Leppo (2013)
- Burzick and Stone (2014)
- Li (2014)

The next subsection describes the various accommodations deployed, their effectiveness, and any moderating variables qualifying the outcomes. Most of the reviews included research using repeated measures with counterbalancing or true experimental studies with random assignment.

Accommodations Studied From 2010 Forward

Three reviews included an array of accommodations. Vanchu-Orosco (2012) used the typology promulgated by NCEO:

- **Setting** (special acoustics)
- **Time/scheduling** (extended time)

- **Presentation** (read aloud, segmented text, and simplified language)
- **Response** (calculators and dictation with a scribe)

Gregg and Nelson (2012) provided the most comprehensive comparison of extended time accommodations for students with LDs.

Harrison et al. (2013) investigated a much broader list of accommodations than previous researchers. For example, their **presentation** of accommodations included choice making, interest, intra-task stimulation and fast paced instruction, and shortened task length. Their **setting** accommodations included adaptive furniture, teacher proximity, extra-task stimulation, and small-group instruction. Their **response** accommodation included "opportunities to respond (OTR), which refers to providing students with frequent opportunities to actively respond to academic requests" (p. 580). Finally, multiple accommodations were "selected through functional assessment or analysis that changes the antecedent to problem behavior to address the function of the maladaptive behavior" (p. 581).

Cawthon and Leppo's (2013) review of 16 studies focused on **linguistic supports** for students who were deaf or hard of hearing, so the accommodations included English item modification, ASL interpreters, extended time, and various computerized supports (all of which provided minimal impact).

The remaining three reviews addressed specific accommodations: extended time or read aloud. The review of 19 studies by Burzick and Stone (2014) focused on read aloud, which has been one of the most frequently implemented accommodations. Finally, Li (2014) conducted a review on read-aloud accommodations for students with and without disabilities by analyzing results from 23 studies.

Effectiveness of Accommodations

Vanchu-Orosco (2012) reported that, overall, students with LDs generally benefitted more than "typically developing peers" whether the effect sizes were calculated for accommodation categories or specific types of accommodations (see Exhibit 3). These average effect sizes, however, also were accompanied by a range that included both negative and positive values and were highly variable. Nevertheless, in both the average effect size and in the number of effect sizes that were small, medium, and large, Vanchu-Orosco concluded that SDs benefited from all four accommodation categories and most specific types of accommodations (except for segmented text and calculators). The effect sizes were smaller for students receiving special education services (which included students with LDs but is more inclusive). Therefore, "it does appear that the more specific we can be regarding type of disability, the better able we are to target appropriate accommodations that have a positive and statistically significant impact" (Vanchu-Orosco, 2012, p. 204). However, test content and student populations accounted for more of the variance than that from the accommodations.

Exhibit 3. Effect Sizes on Test Scores for Accommodations (Vanchu-Orosco, 2012)

General accommodation category	ES: fixed effects model	ES: random effects model
Setting	NA	
Presentation		
Response		
Specific accommodation type	NA	NA
Timing/scheduling		
Read aloud		
Segmented text	NA	
Simplified language	NA	
Calculator		
Dictation (scribe)	NA	
Special acoustics	NA	
Extended time		

Note. ~~Adapted from~~ Values from Table 19 (p. 184) and Table 20 (p. 193) in Vanchu-Orosco (2012).

Gregg and Nelson (2012) focused on effect sizes for extended time using student performance on various standardized tests (including the SAT; state mandated tests; the Nelson-Denny Reading Test; and other non-SAT reading, writing, and mathematics tests). Five group comparisons were made for students with LDs and typically achieving (TA) students. An interaction effect or differential boost was supported: students with LDs exhibited greater gains with accommodations than did typically achieving students: The boost for students with LDs was +.90 when compared with other students with LDs, whereas the boost for typically achieving students was .66 when compared with other typically achieving students. When provided extended time, the effect size for students with LDs was .69 when compared with typically achieving students who also received extended time (see Exhibit 4).

Exhibit 4. Extended Time Effect Sizes for LD Versus TA Students (Gregg & Nelson, 2012)

Comparison of group and accommodation	ES
LD extended time versus TA standard	
LD extended time versus TA extended time	
LD standard versus TA standard	
LD extended time versus LD standard	
TA extended time versus TA standard	

Note. LD = students with disabilities; TA = typically achieving students; ES = effect size. Values from Table 2 (p. 132) in Gregg and Nelson (2012).

Type of test	ES
SAT	
Non-SAT	
Academic skill area	
Reading/writing	
Mathematics	

Note. ES = effect size. Values from Table 4 (p. 134) in Gregg and Nelson (2012).

However, Gregg and Nelson (2012) also concluded that students with LDs performed significantly better when provided extended time, but

> transitioning students with LD still underperform academically as compared to their normally achieving peers whether provided extended time or not on standardized tests. While students with LD perform significantly better when provided extended time, the accommodation does not erase the disability" (p. 136).

Harrison et al. (2013) documented average effect sizes (and ranges) as follows:

- Choice making (−.86 to .49)
- Interest (.85)
- Shortened length tasks (.13 to .53)
- Extra task stimulation (−.91 to .62)
- Small group (.30)
- Extended time (−.1.07 to .30)
- Opportunities to respond (−.67 to .93)
- Multiple accommodations (.94 to 1.0)

It is important to note, however, that these effect sizes were across many different dependent variables, such as the following:

- Task engagement/attention as well as activity level
- Disruptive/desirable/undesirable behavior as well as socially (in)appropriate behavior
- Response rate
- In-seat behavior
- Legible word production
- Aggression
- Noncompliance
- Engagement/disengagement
- Rule violations
- Teacher prompts
- Appropriate/inappropriate behavior
- Work productivity
- Items/problems attempted/correct/completed

Furthermore, the positive or negative sign is important to consider (along with the dependent variables), as the authors note, for example: "Across studies, when participants were provided an opportunity for choice making, task engagement, work productivity, and accuracy increased . . . When choices were provided, undesirable behaviors decreased" (Harrison et al., 2013, p. 570). This is one of the few studies that did not use tests as the dependent variable for gauging the effect of accommodations.

Burzick and Stone (2014) reported outcomes of slightly larger effect sizes with read aloud for reading than for mathematics and for students with and without disabilities, but more so for SDs; moderator variables of content and mode (video, computer, or live) were minor, but younger students showed slightly larger effect sizes (see Exhibit 5). Note that mode included the following: computer; audio CD; human reader; human reader with restricted content; reading pen; video; and video plus highlighting. Extra time was provided on accommodated administration only. Other included content read aloud: not specified; proper nouns and comprehension stems; and entire test.

Exhibit 5. Effect Sizes on Test Scores for Accommodations (Burzick & Stone, 2014)

Content	ES: students with disabilities	ES: students without disabilities
Reading	◔	◔
Mathematics	○	○
Mathematics (human reader)	◔	○

Note. ES = effect size. ~~Adapted from~~ Values from Table 1 (p. 23) in Burzick and Stone (2014).

Burzick and Stone (2014) concluded as follows:

> [R]ead aloud on the reading assessment does appear to be effective at raising test scores for students with disabilities (by an average of .56 standard deviation units)… and… for mathematics assessments, read aloud also increased scores for both student groups, but the average score gains were small for both groups (.13 and .08 standard deviation units, respectively). We found no evidence of differential boost from read aloud on mathematics assessments" (p. 22).

In Li's (2014) study, disability, subject area, delivery, grade, extra time, and research design served as moderating variables in a hierarchical regression analysis. She reported a similar effect as did Burzick and Stone (2014). In reading, irrespective of the delivery method, students with and without disabilities benefitted from the read aloud. In mathematics, however, this benefit for both groups was found only with human proctors reading aloud. Otherwise, when the read aloud was from a computer or a video/audio player, effect sizes were small for SDs and near zero for students without disabilities. (See Exhibit 6 representing the effect sizes displayed in the original publication displayed in Figure 2 of Liu [2014]).

Exhibit 6. Effect Sizes on Test Scores for Accommodations (Li, 2014)

Accommodation	ES: students with disabilities	ES: students without disabilities
Reading (human proctor)		
Reading (computer)		
Reading (video/audio)		
Mathematics (human proctor)		
Mathematics (computer)		
Mathematics (video/audio)		

Note. ES = effect size. ~~Adapted from~~ Values from Figure 2 (p. 10) in Liu (2014).

This finding may be an important caveat for developing a DBA in NAEP, either in formatting the interface or specifying its access: The effect size was greater for human readers than all other modes, for which they noted that "when human proctors read tests, the actual procedure cannot be completely standardized" (Li, 2014, p. 14).

Moderating Variables

Several moderating variables have been concurrently investigated in some of these reviews. In a meta-regression, Vanchu-Orosco (2012) noted that "population description and test characteristic variable sets explained the greatest amounts of variability for change in test score, $R^2 = 0.22$ and $R^2 = 0.35$, respectively" (p. 208).

Gregg and Nelson (2012) concluded as follows:

> [T]he lack of detailed descriptive information about the participants in these studies was even more discouraging . . . only one of the nine studies reported any substantive (i.e., ability and achievement current functioning) or topical marker variables (i.e., cognitive processing current functioning) for the populations investigated. In addition, just three studies reported the type of eligibility criteria used to operationalize LD" (p. 134).

Two studies focused on students with specific disabilities. The review by Harrison et al. (2013) included students with emotional behavior disorder and ADHD, a population rarely studied as part of accommodation research.

Cawthon and Leppo's (2013) review focused only on students who were deaf or hard of hearing. She identified both test-level factors and item functioning as influencing performance; likewise, both educational context and academic proficiency (student-level factors) influenced performance. Finally, she concluded that reading skills and literacy development were important variables with this population, and important contributions on the overall effects were mixed with the type of test and type of accommodation.

Burzick and Stone (2014) focused only on the medium of the accommodation rather than population characteristics.

Finally, the results from Li's (2014) study indicated stronger effect sizes for elementary students than for middle school students (and no significant effects for high school students). Further speculations were offered on the influence of students' decoding skills and test characteristics (e.g., the readability of test items).

Summary

These more recent meta-analyses and reviews of accommodations indicate some positive outcomes and some consistencies. The range of outcome variables is limited primarily to achievement, and differences may exist in the effects on mathematics versus reading tests and student ages (at least for read aloud). All major categories of accommodations may have some positive effects:

- **Setting** (e.g., special acoustics)
- **Administration**, which includes timing/scheduling (e.g., extended time) as well as presentation (e.g., read aloud, simplified language, ASL, and modified items with computer supports)
- **Response** (e.g., scribe)

Differential boosts for extended time may be limited to within student samples overall but may be present with specific accommodations (extended time and read aloud).

The implications of this research for standardization in NAEP are twofold, as noted later in our overall recommendations. First, accommodations with significant effect sizes that reflect an interaction effect, or a differential boost may be added to the list of acceptable accommodations for SDs. Second, when the effect sizes are insignificant, for either targeted groups (e.g., SDs) or all students, the adaptation may be considered a universal design feature.

Meta-Analytic Research on Accommodations for English Learners

As the accommodations research progressed in the first two decades of the 2000s for SDs, further research in their application to ELs also proceeded with seven prominent meta-analyses or summaries completed. Again, a range of accommodations are considered for ELs, then the outcomes (effectiveness), and the moderating variables found to be influential.

- Francis et al. (2006)
- Kieffer et al. (2009)
- Pennock-Roman and Rivera (2011)
- Li and Suen (2012)
- Cohen et al. (2017)
- Rios et al. (2020)
- Liu et al. (2020)

Most accommodations with ELs focus on linguistic variables (separate from content) and, therefore, involve presentation adaptations. As in the research on SDs, these reviews were based on research designs using repeated measures with counterbalancing or true experimental studies with random assignment.

Accommodations Studied

Francis et al. (2006) reviewed 11 studies using the following accommodations to determine their influence on state tests used for NCLB accountability:

- Bilingual dictionaries/glossary
- Dual language booklets as well as questions/read aloud in Spanish, English dictionaries/glossaries
- Simplified English
- A Spanish version of the test

In addition, extra time was investigated (often because it is necessary to access these accommodations in presentation). Kieffer et al. (2009) reviewed the same accommodations and published the same results in the *Review of Educational Research*.

Pennock-Roman and Rivera (2011) published a meta-analysis based on 14 studies investigating various types of accommodations:

- Dictionaries/glossies (pop-up, English, and picture)
- Plain English
- Read aloud
- Dual language
- Bilingual glossaries

Time limits also were investigated as a subset of these accommodations.

Li and Suen (2012) focused on several accommodations (linguistic simplification, dual-language booklet, Spanish version, dictionary, or glossary, and other) in a meta-analysis of 30 studies.

Cohen et al. (2017) investigated the unique effects of a pop-up glossary without confounding variables (e.g., single item presentation; pop-up glossary; extra time; and a small, novel setting). This study was based on the positive effects from prior research on glossaries (Abedi & Hejri, 2004; Pennock-Roman & Rivera, 2011), which contained an audio file that would be played (through headphones) when the student clicked a speaker icon.

Rios et al. (2020) analyzed 26 studies and computed 95 effect sizes that focused on four accommodations:

- Test translation (combined dual language test book and test translation/adaptation; reference group)
- Simplified English
- Dictionaries/glossaries (combined English dictionary/glossary, dual language dictionary, and picture dictionary)
- Combined accommodations

Finally, Liu et al. (2020) summarized test accommodations (from EL accommodations literature published between 2010 and 2018) using the same descriptive variables as in

previous NCEO documents. Their list of accommodations included Spanish translation or enhancement, modified English, English glossary, read aloud, illustrations, and other. As they noted based on their review of 11 publications, "translation and modified English" were studied slightly more often than other accommodations. All the accommodations examined were presentation accommodations and offered "direct linguistic support" (p. 5). Note that three different terms are referenced by the authors (simplified English, plain English, and modified English) with essentially the same type of adaptation.

Effectiveness of Accommodations

The outcomes from the study by Francis et al. (2006) showed the results reported in Exhibit 7.

Exhibit 7. Effect Sizes on Test Scores for Accommodations (Francis et al., 2006)

Accommodation	ES: fixed effects model	ES: random effects model
Bilingual dictionary glossary		
Dual language booklet		
Dual language questions/read aloud in Spanish		
English dictionary/glossary		
Extra time		
Simplified English		
Spanish version		

Note. ES = effect size. ~~Adapted from~~ Values from Table 2 (pp. 31–32) in Francis et al. (2006)

Dual language questions/read aloud in Spanish, English language dictionaries/glossaries, and extra time produced significant positive effect sizes; Spanish language assessments showed effect sizes that varied significantly; and bilingual dictionaries and glossaries failed to show positive effect sizes.

Kieffer et al. (2009) reported essentially the same effect sizes for most accommodations from their meta-analysis of 11 studies. They concluded that the overall mean effect size (effectiveness) was low (*ES* = .04), and only the provision of English dictionaries or glossaries had a statistically significant impact (as well as dual language questions/read aloud in Spanish); two accommodations (bilingual dictionaries or glossaries and Spanish language assessments) showed significant variance across primary studies (meaning they may have been effective in some studies but not others).

Finally, Pennock-Roman and Rivera (2011) reported most effect sizes across accommodations were in a narrow range from slightly negative to slightly positive (see Exhibit 8).

Exhibit 8. Effect Sizes for Test Scores on Accommodations (Pennock-Roman & Rivera, 2011)

Accommodation	ES: English learner	ES: Non-English learner
Groups with restricted time limits		
Pop-up English dictionary	◔	○
English dictionary/glossary	○	○
Picture dictionary	○	NA
Plain English	○	○
Read aloud	○	NA
Dual language	○	○
Pop-up bilingual glossary	○	NA
Bilingual glossary	○	○
Groups with no time limits		
English dictionary/glossary	◔	○
Plain English	○	○
Dual language	◔	NA
Bilingual glossary	◔	NA
Varied extra time		
English dictionary/glossary	◔	◔
Extra time (when allowed alone)	◔	○

Note. ES = effect size. ~~Adapted from~~ Values from Tables 2 and 4 (pp. 17–18) in Pennock and Rivera (2011).

Most accommodations resulted in significantly improved performance when students were provided sufficient time (generous limits) and materials, which should be generalizable for states deploying accountability tests. "The most promising accommodations with generous time limits appear to be the dual language, the bilingual glossary, and the English glossary/dictionary conditions" (Pennock-Roman & Rivera, 2011, p. 22). Under restricted time limits, a promising accommodation included the pop-up English glossary, which may be particularly effective with power tests, in which time limits are not relevant.

Li and Suen (2012) reported an average of 0.16 *SD* for their accommodations (linguistic simplification, dual-language booklet, Spanish version, dictionary or glossary, and other) compared with non-accommodated in test administration, a value statistically different from zero but small, particularly when also considered in the context of the variance often reported by meta-analytic researchers (see Vanchu-Orosco, 2012).

The key outcomes for Cohen et al. (2017) included the **probability of correct response** in large-scale assessments for pop-up English glossary (for non-construct-related terms) and no glossary; with ELs and non-ELs, at Grades 3 and 7; and with three levels of item difficulty (easy, medium, and difficult). These accommodations appeared only on field test items within the standard statewide accountability assessment guidelines, which thus may be

confounded with extra time. As expected, ELs had lower scores than non-ELs. The results, however, were not uniform in mathematics and English language arts (ELA).

> Surprisingly, in both grades 3 and 7, glossaries on math tests seemed to depress the scores of ELs, although a similar trend was not apparent for ELA. In grade 7, glossaries on ELA tests increased the scores of ELs, while not influencing those of non-ELs" (Cohen et al., 2017, p. 267).

Rios et al. (2020) concluded that accommodations for ELs showed only minor improvement (*ES* = .16). This finding contrasts with earlier reports by Keifer et. al. (2009) who showed only English language dictionaries and glossaries to have a positive effect (*ES* = .18) and by Pennock-Roman (2011) who showed pop-up English glossary to have a significant effect (*ES* = .29) with time limits moderating the outcomes, a finding also reported by Li and Suen (2012). As in the research on SDs, the research on ELs has been conflicting in terms of influence on academic achievement (Liu et al., 2020). Effect sizes for various linguistic supports have been moderate at best with many mixed results. English dictionaries/glossaries (with or without the use of pop-ups) and dual language questions/read aloud in Spanish and are more important than bilingual dictionaries/glossaries or Spanish versions. Similar results were reported by Rios et al. (2020):

> Across all studies, test scores improved by an average of .16 SD (SE = .06; 95% CI: .04, .28) when ELs were provided test accommodations; though a large degree of heterogeneity was noted within the sample (I^2 = 90.72%), indicating the need for a moderator (p. 71).

Moderating Variables

As in the research on SDs, many moderating variables appear with ELs, although most of them focus on **language skill** and **previous experience**; potentially, time limits may moderate the effectiveness of some accommodations.

Francis et al. (2006) noted the importance of proficiency in academic language for ELs for interpreting effect sizes, which includes vocabulary, word complexity, and sentence structure. Some accommodations may be more effective when students have sufficient academic language skills for the adaptation in administration to present a 'just noticeable difference' (as discussed later in this white paper). For example, with extremely limited English skills, a dictionary or glossary may effectively be inert. Quite often, researchers do not document the language skills of participants. Another related issue was the status of ELs, which is removed after 2 years of gaining proficiency to participate in grade-level instruction. The findings also suggest that the effect of accommodations may be very different in different contexts or among different populations of students and may reflect unobserved differences in instruction. "It is also possible that bilingual glossaries are effective for a specific group of ELs—those who are literate in their first language and/or who have received content-area instruction in their first language" (Francis et al., 2006, p. 24). They further noted that simplified English, though frequently the target of research, was not statistically significant, but Spanish language accommodation was positive for students instructed in Spanish and negative for students instructed in English.

Pennock-Roman and Rivera (2011) also examined "effect sizes to determine if there were systematic variations in size according to accommodation type, language background of the students, generosity of time limits, test content (e.g., science, mathematics), and grade level" (p. 16). "Although effect sizes ranged from −1.13 to +1.45, the majority (36 values) were clustered in the range of −0.12 to +0.41" (p. 16). In analyzing whether systematic variation occurred as a function of accommodation type, student background, amount of time, test content, and grade level, they reported systematic differences according to these considerations except for test content and grade level:

- For low EL students, the Spanish language versions of tests were effective.
- For intermediate EL, the plain English accommodation was effective.

They also "identified a clear pattern of interaction effects between having generous time limits and particular accommodations requiring additional printed materials" (p. 19).

Li and Suen (2012) investigated several moderator variables (ethnicity, grade level, test subject, English proficiency, and accommodation type [linguistic simplification, dual-language booklet, Spanish version, dictionary or glossary, and other]).

> However, the meta-analysis presented herein shows an estimated grand-mean effect size of 0.157, which implies that on average the accommodated ELLs scored about 0.157 standard deviation units higher than did the non-accommodated ELLs. This effect size is practically small, but it still shows that providing test accommodations for ELLs may boost their test performance to a certain level… The result indicates that test accommodation could improve ELLs' test performance in a general way, thus supporting the effectiveness of test accommodations for ELLs (p. 21).

In the Cohen et al. (2017) study of glossaries, mathematical performance was depressed for ELs, whereas in ELA, glossaries enhanced performance. They hypothesized that the pop-up glossary presented an extra cognitive load for younger students, likely because only words irrelevant to the construct appeared in the glossary, which may have been a distraction.

Finally, Liu et al. (2020) summarized the focus on accommodations for ELs as showing some consistencies, but English language proficiency and extended time (or time limits) are important moderators. Unlike previous research on ELs, few student characteristics have been considered as moderating variables (Liu et al., 2020).

Summary of Accommodations Research

The most important implication of this research is the degree to which these various accommodations can be applied within the NAEP testing program and with what potential effect. When boiling down the accommodations research to only those showing positive effects in various reviews and meta-analyses, many accommodations have not been found to be differentially effective:

- The accommodation neither reflected an interaction effect nor a differential boost; many also were inconsistent in their effectiveness.

- An interaction effect would mean that the accommodation worked for the targeted group (usually SDs or ELs) but was inert with the comparison group (usually students in general education).
- With a differential boost, all students would have benefited, but the improvement was greater with the target group.

In much of this research on accommodations, neither an interaction nor a differential boost was reported. However, the definition of an effective accommodation should be considered with other metrics, such as an increase in participation by a more diverse group of students (cf. Lutkus & Mazzeo, 2003); an enriched understanding of the construct being assessed; and an impact in the classroom, given the need for accommodations to not be introduced for the first time during testing sessions. These latter three attributes provide a deeper conceptualization of adaptations in testing programs that expand accommodations to more universally designed features in rethinking standardization and allowing them for all students.

For SDs, setting and two administration accommodations (extra time and read aloud) appear to show positive effect sizes (in both reading and mathematics, though clearly controversy is present in the former subject area). Furthermore, read aloud accommodations may be positive whether items are read by a human, by a computer, or from an audio-video source. For ELs, three administration accommodations appeared effective: dual language questions/read aloud in Spanish, English dictionaries/glossaries (either as a traditional source or as a pop up) as linguistic supports, and extra time, all showing positive effect sizes.

To better interpret or qualify these various findings, the effect of adaptations consider (a) the impact in terms of **noticeability** when focusing on the student instead of the test/item, with reference to effect sizes (from earlier content in this paper) and (b) the proficiency categories in which students are placed based on their performance (scores). In both analytic perspectives, the goal is to rethink standardization in an empirical and rational manner that can maintain the integrity of NAEP and provide some sense of order in the manner that adaptations are made and evaluated, particularly within the confines of the *Standards*.

When Is a Difference a Just Noticeable Difference?

In psychology, the term "just noticeable difference" refers to the threshold of "noticeability" and may be applied to variation in test administration relative to performance. In an analysis by Sireci (2020) on 'understandardization', this term can be traced back to its origins in psychophysics with investigations of sensations, scaled to reflect noticeable values. This process relied on carefully controlled procedures that eventually became adopted in the field of testing to document comparisons across individuals uninfluenced by measurement conditions. The question then is whether these adaptations result in a noticeable difference. In the research cited earlier, the effect sizes for many test adaptations are small, even when statistically significant. Cohen (1988, 1992) originally proposed that effect sizes of .20 or less are small, .50 are medium, and .80 or greater are large. However, as Nye et al. (2019) noted, these values are for experimental research, not measurement (non)equivalence research. In their Study 1, they found "values of .20 appear useful for differentiating between negligible effects and small differences" (p. 685) under varying conditions of different sample sizes and the number of items. Therefore "Cohen's guidelines may not generalize to effect sizes for

interpreting measurement nonequivalence" (p. 687). In their Study 2, they further qualified effect sizes and reported that

> cutoffs of .40, .60, and .80 might be considered small, medium, and large effect sizes. These values are somewhat larger than those identified in Study 1 indicating that although an effect size might be considered medium relative to other findings in the literature (e.g., .40), the practical importance of the effect may still be small (Nye et al., 2019, p. 700).

Rx 1: Designate accommodations with effect sizes that are small (according to Nye et al., 2019) from the research on universal design features. A liberal interpretation would therefore consider accommodations to be comparable to a standard administration (e.g., if below an effect size of .20) for both SDs or ELs. Of course, the actual values may be debatable, but certainly many accommodations were well below .10. For SDs, this would include segmented text, simplified language, and calculators (when appropriately allowed for items testing conceptual or procedural skills but not calculational skills; Vanchu-Orosco, 2012), and computer or video/audio read aloud in mathematics (Burzick & Stone, 2014; Li, 2014). For ELs, bilingual dictionaries, dual language booklets, English dictionary/glossary, simplified English, and Spanish versions would be considered standard administrations (Francis et al., 2006; Kieffer et al. 2009). From Pennock-Roman and Rivera (2011), the following would be considered standard administrations with restricted time limits: pop-up English dictionaries, picture dictionary, plain English, read aloud, dual language, pop-up bilingual glossaries, and bilingual glossaries; only plain English would be a standard administration with no time limits.

Impact of Adaptations Within/Across Performance Levels

Many effect sizes have been very modest on scale scores, and most research on accommodations has focused on only score changes, often on state testing programs, with only general reference to categorical performance changes, usually noting that participation is increased with no change in proficiency (Tindal & Ketterlin-Geller, 2004). These latter two metrics may serve as more important (and meaningful) dependent variables than simple gains or losses on student scores. Few studies have investigated the effects of accommodations on changes in performance categories. Yet, performance levels contain considerable variability in score values within them (Tindal et al., 2017).

Rx 2: **Analyze the impact of accommodations within/across proficiency categories.** If no differences continue to exist between proficiency categories, accommodations may be interpreted as having limited effects, possibly adding a clause that universal design features may be considered as adaptations with few substantial (significant or meaningful) outcomes affecting proficiency status that could be made available for all students. With sufficient samples, separate effect size comparisons would be necessary (accommodated versus not accommodated) either within proficiency categories or more importantly at the cutoff values across them. Basically, the focus is not on simple score differences between accommodated and non-accommodated students but on the score differences within/across proficiency categories when students are blocked by race/ethnicity, gender, SD status, or EL status. The rationale for this analysis is to simply affirm the functional effects of accommodations.

CONSISTENCY OF ADAPTATIONS IN TESTING PRACTICES

In expanding test changes, it is possible to compare those offered in NAEP to other states that use the same the typology of changes (accommodations, designated supports, and universal design). In this section, the Smarter Balanced consortium of states is deployed to list state practices as reported by NCEO.

Smarter Balanced Typology of Test Adaptations

Accommodation adaptations **embedded** within Smarter Balanced assessments are relatively confined and include the following:

- ASL*
- Braille*
- Braille transcript
- Closed captioning*
- Text to speech*

All of these accommodations provide access to the content of problems/items, and those in common with NAEP are marked with an asterisk. **Nonembedded** accommodations in Smarter Balanced assessments include the following:

- 100s number table
- Abacus
- Alternate response options
- Braille*
- Calculators*
- Multiplication tables
- Print on demand*
- Read aloud*
- Scribe*
- Speech to text
- Word prediction

Again, test adaptations provide access to the test content or support for solving problems.

Designated supports embedded in Smarter Balanced (2021) assessments include the following adaptations along with their possible functions designed to increase equity and access:

- **Color contrast** is used to create more of a difference in hue between text and the background.
- **Illustrations**, which can range from line drawings to realistic photo-like images, either of which may provide appropriate cues to track in the text. NAEP currently includes illustrations in some items as warranted.

- **Glossaries** would provide students translations of words either in a look-up manner (glossary) or side-by-side or alternate form (dual language). The function is clearly to provide access to the text, which may need to be carefully rendered if focusing on literal comprehension.
- **Masking** would function much like line readers in that they control the stimulus so that the student can expand or contract information in a functional manner (suited to their visual or cognitive load).
- **Mouse pointers** simply allow students to target key elements of the item and could function as a guide in problem solving.
- **Streamline** allows students to reduce text so it would function as a control mechanism for attending to critical information.
- **Text-to-speech*** provides students access to information on the problem and/or items that otherwise may not be read or misread.
- **Translations** (directions, glossary, and dual language) provide alternative text that could ensure that the student has access to the content.
- **Ability to turn any off universal tools** can allow customization of the digital environment to activate only those adaptations warranted as functional.

Several other **nonembedded designated support** features also are included in the Smarter Balanced (2021) nondigital administration:

- Amplification
- Bilingual dictionaries
- Color contrast*
- Color overlays
- Illustration glossaries
- Magnification
- Medical supports
- Noise buffers
- Read aloud (English or Spanish)*
- Scribe
- Separate setting*
- Simplified or translated test directions
- Translations (glossary)

The purpose of most of these designated supports is to ensure that students have access to the content.

The complete list of universal design adaptations listed in NAEP (see Exhibit 1) can be compared with those considered as universal design by Smarter Balanced. The following list presents universal design features in Smarter Balanced (2021) that are allowed with their function listed for each:

- **Breaks** would provide more capacity to study problems and perhaps avoid fatigue.
- **Calculators** would be allowed when the construct involves procedural and conceptual knowledge rather than computational skills.

- **Digital notepads** would serve as a scaffold for test takers to organize key information to be used in solving problems.
- **English dictionaries and glossaries** would be allowed as students select words with which they are unfamiliar, increasing the functional capacity to solve problems in reading and content areas. As noted in the research with ELs, extra time may be needed.
- **Expandable passages and/or items** allow students to function in the manner best suited to them with the size and density of text viewable within the screen. Furthermore, this adaptation would allow more diversity in the type of devices (tablets or computers with varying screen size and flexibility in scrolling).
- **Global notes** could provide students self-written cues for monitoring their performance.
- **Highlighters*** function as cues for students to focus on relevant information.
- **Keyboard navigation** is a functional skill that allows better efficiencies than is present in drop-down menus. This adaptation may need to be accompanied by a menu of shortcuts that students can use as a reference.
- **Line readers** function as control devices for ensuring that text is read contiguously with no lines skipped, which would otherwise result in scrambling content and omitting information.
- **Mark for review** would allow test takers to track responses that may be answered with uncertainty.
- **Mathematics tools** can include rulers, measurement devices, triangles, abacuses, and others as deemed appropriate per the construct being assessed. Their function would allow students to focus on the critical concepts being assessed. Again, with scenario-based tasks, this functional adaptation is currently allowed in NAEP.
- **Spell checkers** would be useful in writing with the primary function being to create compositions that are readable so that readers can focus on the content (e.g., events, sequence, ideas, organization).
- **Strike through** would be the opposite of highlighting critical information but would allow test takers to discard unnecessary information, allowing them to focus on only relevant information.
- **Thesaurus** would allow students to edit words (antonyms and synonyms) so that they can track those used from those not used.
- **Writing tools** would function in a comparable way to mathematics tools, allowing compositions to be created with varying text characteristics (e.g., highlighting certain words, indenting certain paragraphs).
- **Zoom*** has a clear function of making test content more visible.

Finally, other nonembedded adaptations considered universal design are invoked outside the digital environment:

- Breaks (allowing respite from concentration to avoid fatigue)
- English dictionary (to understand the meaning of words)
- Scratch paper* (for practicing operations or taking notes)
- Thesaurus (for knowing synonyms and antonyms)

These adaptations potentially increase access to content or support for responses. The critical finding is that only two test adaptations are considered universal design by both testing programs: zooming and scratchwork/highlighting. However, NAEP also includes many more features as universal design that are not adopted as universal design by this state consortium using Smarter Balanced.

In addition, it is important, however, to consider Smarter Balanced accommodations adaptations and designated supports that may not be present in NAEP. As noted earlier, all test adaptations are either embedded within the DBA or not embedded, appearing outside it. Furthermore, designated supports may be used by all students, but on NAEP, appropriate educators make the decision to apply them as they are deemed necessary. When NAEP is administered in schools, only students who participate in NAEP-supported test adaptations (either as an accommodation or as part of a universal design feature) are included. If the student has an adaptation allowed for the state testing program that is not listed with NAEP, they may be excluded from the original ~~roster~~ assessment. In contrast, the reverse is not true: Students receive adaptations in state testing programs (whether accommodations, designated supports, or universal designed features) irrespective of whether they receive them in the NAEP testing program. In all three types of test adaptations in Smarter Balanced assessments, when they also are included in NAEP, they have been marked with an asterisk.

State Practices in Typologies of Test Adaptations

NCEO has compiled a list of test adaptations adopted by states, categorizing them as accommodations, designated supports, or universal design. These test adaptations (n = 24) appear in the most recent Tool Kit (NCEO, 2021). This list can provide guidance for state assessment directors and other state education agency personnel, as well as members of technical advisory committees. Again, these test adaptations have been adopted to function as supports for students that provide greater equity and access. The following summaries include only those adaptations listed as universal design features with some states. Again, adaptations in common with NAEP are indicated with an asterisk. See the appendix for a list of test adaptations adopted by states as accommodations, designated supports, or universal design.

- **Calculator** allows students to solve problems without making careless errors in operations (though it would not be appropriate for mathematical operations problems).
- **Clarify/simplify/repeat directions** ensures that students hear the directions and problems accurately.
- **Color contrast** provides students better access to printed text or text on screens and therefore allows them to better understand and interpret the problem or item.
- **Extended time** would function like breaks, allowing students to take more time in reading the problems, reviewing the options (for selected response types), or composing responses in production tasks.
- **Familiar proctor/test administrator** provides the student a setting in which they have experience (perhaps in the directions being read in a more appropriate manner or in the prompting to move along) to increase access to a wider range of problems.

- **Highlighting*** allows students to make critical text stand out more from other text, thereby reducing the need to mentally sort/focus on critical content.
- **Human read aloud** provides students access to problems and options that otherwise may be not read or misread.
- **Magnification** is a simple strategy to ensure that the student can see/read the item.
- **Manipulatives** would allow students to organize or sort objects to represent the problem concretely.
- **Mathematics charts/tables** function as a scaffold to ensure accurate information can be retrieved (e.g., conversion of measures across different metrics).
- **Multiple days** would function like breaks in which students can avoid fatigue and maintain attention.
- **Noise reduction** can occur in any manner that allows students to maintain attention, by either reducing excess noise, providing white noise, or playing music from the student's playlist.
- **Paper format** would function to ensure that students can see/read items and problems without having to access content displayed on a computer screen and potentially reduce glare or avoid scrolling.
- **Preferential seating** may function to reduce student anxiety or ensure that directions are heard (e.g., sitting in the front of the room).
- **Signed administration** is designed so that students with hearing impairments or who are deaf can participate in the test.
- **Small-group and individual administration** potentially provide students a less distracting environment.
- **Spell check** ensures that students' (written) responses are legible for more accurate scoring.
- **Student reads aloud to self** is a simple way for students to "subvocalize" when reading text, functioning like masking or highlighting.
- **Test breaks** provide students time to avoid fatigue and maintain concentration.
- **Text-to-speech (computer-generated voice)*** provides students the correct problem to be addressed, which might otherwise be misread (~~and~~ or misinterpreted) by the student.
- **Word prediction** allows students to automatically spell-check words as they are written.

Rx 3: Revamp distinctions between accommodations, designated supports, and universal designs. Every jurisdiction decides eligible accommodations for students.[7] The National Assessment Governing Board (2014) confirms that

> allowable accommodations are any changes from standard test administration procedures, needed to provide fair access by students with disabilities that do not alter the constructs being measured and produce valid results. In cases where non-standard procedures are permitted on state tests but not allowed on NAEP, students will be urged to take NAEP without them, but these students may use other allowable accommodations that they need (p. 3).

[7] https://nces.ed.gov/nationsreportcard/about/inclusion.aspx.

However, in applying these criteria, little distinction is made between accommodations and universal design features. Yet many of the adaptations listed as accommodations could be considered universal design and incorporated as part of a NAEP research agenda to determine the effects on both performance and participation. Furthermore, it would be possible to classify them as universal design in line with the *Standards* and include more information about them while concurrently addressing setting, administration (including qualifications of administrators, time needed, presentation, interface/engagement), and responses (including scoring protocols).

To provide a more accessible and equitable test, NAEP may consider developing/adopting criteria to organize test adaptations in these three groups. For example, NAEP states that "accommodations in the testing environment or administration procedures are available for SD and EL students to support their participation in the assessment. Some accommodations are built-in features – or Universal Design Elements – of the digitally based assessments that are available to all students. Other accommodations, such as additional test time, are available upon request (see footnote 7)." As noted later in this paper, the new NAEP reading frameworks[8] reflect advancements systemic to the development and delivery of NAEP.

Rx 4: Allow students to take NAEP tests using state-adopted accommodations, designated supports, or universal designs. Currently, NAEP allows adaptations classified as only accommodations or universal design. To provide more consistency with state testing programs, adaptations considered as designated supports by the state may be considered as universal design by NAEP or expanded definitions of adaptations in NAEP could include this category. This adaptation would allow more students who are rostered to then be included in the NAEP sampling plan. As stated in the guidelines, nonstandard NAEP accommodations, though allowed by the state, are generally not allowed when students take NAEP tests. Yet, many of these test adaptations are both familiar to the students and teachers and represent viable options for large-scale testing programs. For example, extended time is limited, with no allowance for breaking the time into different sessions (extra time and possibly different setting or scheduling). This adaptation would be allowed as a function of the students' classification (SD/EL, SD, and EL as listed earlier in this paper).

[8] https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/reading/2022-nagb-reading-framework-508.pdf or https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/reading/2026-reading-framework/naep-2026-reading-framework.pdf?

IMPACT OF TEST ADAPTATIONS FOR STUDENTS

Although our presentation focuses primarily on standardization of testing programs (both NAEP and other state accountability systems), it is important to also consider the populations of students sampled in them. Although most state testing programs attempt to sample the entire population (with various rules allowed for opting out), NAEP deploys a well-structured **sampling plan** that represents the demographics of states and districts.

In either case, accommodations research is very clear that (a) insufficient attention has been given to student characteristics (other than broad categorical variables such as disability or language) and (b) the presence of an interaction or differential boost is an indicator of effectiveness (it has positive effects for some students but either no effect or even more effect for other students). Therefore, in this section, two issues address sampling students, both in the measurement of variables beyond broad characteristics, as well as the sampling itself. For NAEP to move forward in "rethinking standardization," any adaptations in the testing program (setting, administration, or response) need to be fair and not provide unwarranted advantages. If such adaptations are desirable to certain subgroups, they can/should be offered. In contrast, should any interaction or differential boost occur, explicit deliberation needs to occur on potential reasons and consideration of the validity of inferences that are warranted.

Current NAEP Data Collection on Students

NAEP collects considerable information about students. Even though most of this information is not collected within the test session, it is carefully collected and used to monitor various impacts of the NAEP testing program. In addition to information on students collected from administrative records (grade level, race/ethnicity, gender), NAEP collects the following information on students:

- Student perceptions (e.g., familiarity with technology, level of effort, interests, attention, skills, classroom experiences and activities)
- Home environment issues (e.g., books, receiving help, lives at home with)
- Instructional content and practice (e.g., engaging in various educational activities, receiving assistance)
- Factors beyond school (e.g., participating in activities, talking about studies, getting help beyond the school environment)

Other student information collected includes disability (type and severity), EL status, participation with accommodations, and days absent (with breakdown in sets of 1–2, 3–4, 5–10, and more than 10 and reported with significance tests). For SDs, see https://nces.ed.gov/nationsreportcard/about/inclusion.aspx; for ELs, see https://nces.ed.gov/nationsreportcard/pDS/bgq/sch-sdlep/BQ11_NAEP_ELL.pDS.

Most of this information is of great value in addressing consequential validity and making inferences about the quality of education in the United States. This purpose is different from understanding test adaptations as they relate to specific student characteristics, such as the original research conducted on the inclusion of SDs and ELs (NCES, 2000). This research

addressed several variables in analyzing accommodations in both reading and mathematics for students in Grades 4, 8, and 12. For SDs, these variables included the following:

- Participation (with and without accommodations)
- Severity of disability (mild, moderate, severe)
- Type of disability
- Amount of time spent in the general education classroom (mainstream)
- Grade level of instruction
- Responses on the NAEP questionnaire

For ELs, these variables included the following:

- Time in the United States
- Native language
- School attendance
- Enrollment in English language schools
- Instructional experiences and practices (e.g., academic instruction in English)
- Grade level of instruction
- Responses on the NAEP questionnaire

This research is important to better understand participation (inclusion) rates of both populations. Another important study on inclusion rates of SDs was conducted later (NCES, 2011) to document state variation and the influence of student characteristics (disability and severity), IEPs, and use of non-NAEP accommodations. Based on the results from 2007 to 2009, adaptations in inclusiveness indicated that in Grades 4 and 8 in reading and mathematics, either no adaptation or increases in performance occurred across the states.

These studies are important in not only guiding the trend toward more inclusive practices but also documenting the influence of specific variables and could help guide the NAEP outcome reports to be more granular. For example, the student groups selected for sub-reports included race/ethnicity, gender, National School Lunch Program eligibility, highest level of parental education, type of school, charter school, school location, region, with (or without) a disability, and status as an EL. Based on these two reports in 2000 and 2011, results on adaptations to testing practices could be analyzed with more refined student groups and used to not only rethink standardization but also use any pilot research to guide adoption of testing practices.

Rx 5: Expand reporting of participation in NAEP results to reflect the important research previously conducted that reflects the diversity of the population participating (knowing that samples may be too small for releasing score reports to the public but may guide further internal research).

- **5.1**: For SDs, these variables would include the severity of disability (mild, moderate, severe), the type of disability, the amount of time spent in the general education classroom (mainstream), grade level of instruction, and embedded responses on the NAEP questionnaire.

- **5.2**: For ELs, these variables would include time in the United States, native language, school attendance, enrollment in English language schools, instructional experiences and practices (e.g., academic instruction in English), grade level of instruction, and embedded items on the NAEP questionnaire.
- **5.3**: Consider analyzing results from students who both have a disability and are classified as an EL in proportion to their presence at the school level. As Wu et al. (2021) noted, "[T]he percentage of English learners with disabilities increased from 1.2% (approximately 0.5 million students) in 2012-13 to 1.6% (approximately 0.7 million students) in 2018-19 (1.5% in 2017-18) for students age 6-21 who were enrolled students in grades 1-12" (Abedi, 2009, p. 2). This group of students has rarely been studied for the effects of accommodations.

Granularity and Specificity of Student Population Descriptions

Although inclusive practice was notable in the 1990s, when NAEP began to include SDs and ELs, more refined attention is necessary to further articulate student characteristics, particularly because the identification rates of both subgroups is currently much higher. In doing so, it is important to address three problems that exist with descriptions of populations taking NAEP.

First, Dabbs (2003) proposed a more detailed analysis of procedures used by NAEP in sampling schools (both in developing lists from which schools are sampled and in sampling students from those lists). In sampling schools, lists may be incomplete, and information may be missing on the nonparticipation of schools. Furthermore, student populations in elementary and secondary settings have changed considerably since most of the research on accommodations was conducted. Presently, the samples of students participating in NAEP are in proportion to the percentages of students nationwide. This sampled population could be more extensively described, with concurrent information collected, and new samples of students could be added.

A second problem is that change in demographics is not documented concurrently with change in trend. For example, in California, the percentage of ELs was 18.11% in 2021–22, but it also has witnessed a decrease of nearly 7% since 2000 (https://nces.ed.gov). Nearly 75% of Mississippi's public-school students are eligible for free or reduced-price lunch, which represents a 10% increase since 2001. By contrast, New Hampshire (with the lowest rates) has only 24% eligible for free or reduced-price lunch (but with a comparable increase of nearly 10% in the past decade). (See https://nces.ed.gov for the years 2000 through 2019.)

Finally, the research conducted on NAEP may have sampled different populations than the populations reported for NAEP administrations (NCES, 2005). Although the research on SDs and ELs often is conducted on samples of students within schools or districts, NAEP samples are based on representative samples for states and TUDAs (Trial Urban District Assessment), which may not capture between school differences of student demographics (e.g., at-risk populations, SDs, ELs, poverty status) that typically occur. For example, individual schools within districts are likely affected by district policies and geographical factors that may interact with student characteristics (e.g., students who are homeless being served in schools based on services provided. Likewise, the identification (and proportion) of SDs may be drastically different between schools within a district, depending on the school's

status as a residential or attendance school within the district. Although these differences in populations should be reflected in the NAEP sampling plan, they are likely masked when the sampling plan is within the state or district. Furthermore, instructional practices are implemented at the school level (e.g., response to intervention), EL immersion programs, mainstream versus pullout programs), all of which affect the opportunity to learn. In addition, immigrant students, a population with both limited cultural experiences in the United States and limited English knowledge, are likely to vary among schools, particularly in large urban centers.

Rx 6: Conduct research and report on specific student groups to represent the diversity of student groups for each NAEP administration, document adaptations across time, and note differences among student samples when they make a difference.

- **6.1**: Respond to issues noted by Dabbs (2003) and sample schools and student subgroups that are currently missing or incompletely sampled. The school lists should include schools for the blind or deaf and schools in separate settings (such as hospitals and prisons). Furthermore, in sampling students, the lists may be incomplete, and nonparticipation of students may be present, "which includes students who fail to appear for the assessment and students who are excluded" (Dabbs, 2003, p. 3). Other examples of student subgroups can be accessed in NCES statistics reports on school-age students: As of 2017, 0.2% are in prisons, 0.4% are homebound or in hospitals, 1.4% are in regular private schools, 0.2% are in separate residential facilities, and 2.8% are in separate schools for SDs—a total of 5%. Although NAEP has a well-developed sampling plan, further articulation may be needed for subpopulations as the demographics shift (e.g., the turn to private education with the onset of COVID-19). Even within schools, the timing of NAEP tests may inadvertently exclude some populations: For example, SDs may not be present during NAEP testing because they were not counted in the December 1 child count but identified with a disability later in the school year.
- **6.2**: Dynamically, document changes in student demographics across time and sample them in proportion to school rates rather than limiting sampling in proportion to the district and state numbers using state websites and current population estimates. This sampling plan would stratify on schools within large districts.

SPECULATIVE PERSPECTIVES AND ISSUES

In this section, testing is analyzed in general, which then leads to possible adaptations in which standardization can be reframed. Such adaptations need to be considered initially in terms of purpose and then the strategy for accomplishing this purpose to emphasize function over format of the process. In considering function, accommodations represent individualized adaptations for SDs or ELs to provide access that would otherwise be prevented by these students' characteristics (labels). In contrast, designated supports and universal design represent adaptations for all students and primarily focus on flexibility in test administration without changing the content of the item or the construct of the measure.

Focus Primarily on Function not the Format of the Process

Although traditionally the emphasis on standardized test administration is exact sameness of the process (setting, administration, and response) in which behavior is solicited, occasional adaptations may indeed become part of the construct, supporting a common interpretation of performance. The nexus of the issue is to ensure that the adaptations do not influence the function of the behavior, even though they change the format of the process.

For example, writing can be completed with a paper and pencil, a typewriter, or a computer (which can vary from model to model but presumed to have similar features in composing and editing). In 2017, the NAEP Writing Assessment began using DBAs to measure three purposes for writing: (a) persuasion in changing a reader's point of view or influence the reader's action, (b) explanation for informing the reader's understanding, and (c) conveyance of an experience (real or imagined) for the reader's appreciation. Each NAEP Writing booklet contains two writing prompts, each addressing one of the three purposes based on "real-world, age- and grade-appropriate issues that are familiar and accessible" (Mazany et al., 2017, p. 19) and oriented to a particular audience using any of several formats (e.g., letter, essay, opinion piece). Each type of discourse is scored based on the development of ideas, the organization of ideas, and facility in using language and conventions, each with more specific criteria for evaluating responses. The proficiency levels also focus on communicative purposes with appropriate text structure; details; voice; phrasing; and syntax, grammar, and spelling (see Mazany et al., 2017, p. 45). In summary, the function of the behavior has been carefully considered to ensure that assessment of writing remains relevant to the current socioeducational context, in which most forms of written communication are completed in a digital environment (whether email, text messages, or documents). Perhaps the only context in which handwriting occurs is within cards or notes used for birthdays, holidays, and so forth. In this example, the function of the response is primary, not the format of the process.

This emphasis on function over the format of the process represents a bold move for NAEP, even though an extensive literature exists on writing assessments with paper-pencil administration compared with computer-based administration. Yet, these two forms of behavior (handwritten versus digitally written) are considerably different and lead to significant differences in the format of the process. "For students accustomed to writing on computer, responses written on computer are substantially higher than those written by hand (effect size of 0.9 and relative success rates of 67% versus 30%)" (Russell & Haney, 1997, p. 1). Yet the constructs targeted by NAEP Writing are sufficiently broad in their (three)

purposes (types of discourse), scoring rubrics, and proficiency levels. Furthermore, the constructs are more easily accessed with features of a DBA: copy/cut paste text, highlight, word search, formatting within and across paragraphs, text changes (e.g., bold, italics, font). These and many other features make the text much more readable with the author better able to control both the writing and therefore the reading experience. In summary, it is the function of behavior that is important in reflecting the construct, not the format of the process. In writing, the construct is moot on format.

As in writing, the reading framework is broad by incorporating both literary and informational texts, focusing on locating or recalling information, integrating and interpreting what has been read, and critiquing and evaluating perspectives. Also, like writing, the assessment is digitally based (in Grades 4 and 8) with various tools (on-screen pencil/highlighter, color theming, zooming, and text-to-speech). Performance is a function of constructs such as word meaning, the importance of details, the sequence of events, inferences from evidence in the story, opinions, themes, text structure, and conclusions.

With this construct in mind and a focus on function, the assessment could be completed in any language (for either presentation of the story or in the items generating the response). The layout of the test itself could be landscape or portrait, if similar features are allowed (e.g., screen breaks, masking, highlighting). In several accommodations currently allowed by NAEP, a functional focus appears, although it is limited (e.g., translation into Spanish but not in other languages). This focus within DBAs is typically based on the function of behavior, not the format of the process, and is expandable to several other issues in test administration. For example, noting differences among devices would misdirect attention to the format of the process (e.g., tablets and computers of varying sorts) rather than the function of the behavior (responding to different reading tasks). In reading, the focus should ignore the layout of passages (vertical or horizontal), the presence of scrolling features, or keyboard controls, all of which misdirect attention to format of the process, not the function of the response.

Rx 7: Increase diversity by continuing to adopt a variety of alternate passages with different content (but the same genre and text structure). For example, student choice in reading passages echo recommendations provided by Hughes (2023) for increasing the diversity of NAEP reading content. To illustrate, the test items following the "Wanted: News Carrier" text could provide alternate text with different positions (e.g., advertising for a community theatre, providing directions at a local community event, working at a food cart). The same text structure would be followed in each of these "positions" so that students could view more varied "want ads," to apply for the positions in which they are interested. This adaptation would be consistent with the current emphasis on "authentic" text. It might be possible to use the same questions if the content was consistent, such as the approach taken in the Campbell and Donahue (1997) study. Other stories could be adapted (diversified) to reflect different characters or events with which students identify culturally and socially. These suggestions are designed to rotate through the matrix sampling of students and provide a broader bank of stories and questions, so they are potentially more relevant to the increasingly diverse sample of students in U.S. schools.

Rx 8: **Rearrange the format (spacing as well as font type and size) of the test passage and items, to allow flexibility in orientation (vertical or portrait view printed or on screen with scrolling if a DBA)**. Currently, the practice items appear above the questions, and a horizontal split screen allows scrolling from item to item. Another option would be to provide a vertical split screen with highlighting in the passage allowed (already a universal design feature in NAEP and Smarter Balanced, as well as many states but not available in the practice tests). A search box could be provided so that the student could reference a word within the test item to search the text of the passage where that word appears. This feature would allow the student more time to focus on the specific text for responding to the item[9]. Text could be chunked in successive groups with questions addressing the content, making it more accessible. With either option, it would be important to ensure that the practice items are consistently formatted in the same manner as the operational test.

Distinguish Accommodations From Universal Design Features

To distinguish among accommodations, designated supports, and universal design, three questions are proffered for consideration. These questions focus on the impact of function in making test adaptations available to a wider range of students, with the goal of including all students in the key NAEP subject areas and grade levels. They are designed to clarify when a construct changes and guide the rationale for making adaptations and rethinking standardization.

- **Maintain Construct:** *Is it likely that the adaptation will not affect the construct being measured?* In the case of both Braille and ASL, indeed it is possible for such adaptations to possibly affect the score and interpretation, therefore requiring more extensive consideration of the issues raised in the *Standards*. In contrast, writing directly in booklet or having an aide in the testing room is not likely to alter the administration or response (process) and affect the construct. Therefore, they would be considered as universal design, not an accommodation.
- **Increase Equitable Access:** *Does the adaptation allow equitable access for students who would otherwise be excluded?* For example, students with orthopedic impairments may be excluded from taking either a paper-and-pencil NAEP or participating in a DBA without the use of a scribe. In this case, the adaptation may be cast as universal.
- **Match Adaptations With Student Experiences**: Are the adaptations present in the classroom environment, familiar to the student, and implemented with specialized training? These criteria would emphasize the requirement in the Standards for qualifications of administrators, time needed, and scoring protocols followed. Again, Braille and ASL would be considered as accommodations, whereas many others currently listed in NAEP would be shifted to universal design.

All three questions would be addressed when an adaptation is being considered. If the answer is yes to all three questions, it would be a universal design feature; otherwise, it would be an accommodation. For example, extended time is not likely to result in much variation of administration or scoring within a proctored environment (assuming that the rate of response is not part of the construct), allows equitable access, and is likely to have been used in the classroom (in fact some students may expect it). On the other hand, both Braille and

[9] Note: This feature may need to be limited to certain passage types.

ASL could be deployed with variation, though it may provide access to targeted students who need it and is likely to have been part of their educational program. About the variation: Braille may be contracted or uncontracted, and many types of ASL exist. Furthermore, students may vary in their familiarity with the type of Braille or ASL used. In either case, because of variation in the administration and scoring or in the ability of the student to respond appropriately (e.g., variation among students' experience and familiarity), different testing environments may be created, thus affecting the construct.

These three questions can be applied for each adaptation that NAEP delineates and the targeted populations (SDs and EL. In Exhibit 9, the questions are addressed for adaptations available for both SDs and ELs; in Exhibit 10, they are addressed for adaptations available only for SDs; and in Exhibit 11, the questions are addressed for adaptations available only for ELs. If all three questions are answered "Yes," then we argue that the NAEP accommodations could be considered universal design. If fewer than ALL three answers are "Yes," then the adaptation would be an accommodation. After answering each question, the last column reflects whether the adaptation would be an accommodation or universal design.

Exhibit 9. NAEP Standard Accommodations for SDs and ELs

Adaptations	Not vary?	All access?	Experience?	Category
Extended time	Yes	Yes	Yes	Universal
Small group or one-on-one	Yes	Yes	Yes	Universal
Breaks during testing	Yes	Yes	Yes	Universal
Directions read aloud only in English	Yes	Yes	Yes	Universal
Test items read aloud in English—occasional or most/all (but not in reading)	Yes	Yes	Yes	Universal

Note. NAEP = National Assessment of Educational Progress; SDs = students with disabilities; ELs = English learners.

Exhibit 10. NAEP Standard Accommodations for SDs

Adaptations	Not vary?	All access?	Experience?	Category
Must have an aide present or preferential seating	Yes	Yes	Yes	Universal
Calculator for mathematics for FN3	Yes	Yes	Yes	Universal
Large print version of the test (music only, not visual arts)	Yes	Yes	Yes	Universal
Magnification	Yes	Yes	Yes	Universal
Use of template/special equipment	Yes	Yes	Yes	Universal
Cueing to stay on task	Yes	Yes	Yes	Universal
Presentation in Braille (not in science), or ASL (not in reading)	No	Yes	Yes	Accommodation
Responds orally to a scribe (not in writing)	No	Yes	Yes	Accommodation
Response in Braille or ASL (not in music, visual arts, TEL, or writing)	No	Yes	Yes	Accommodation

Note. NAEP = National Assessment of Educational Progress; SDs = students with disabilities; FN3 = NAEP Mathematics Test Form; ASL = American Sign Language; TEL = technology and engineering literacy.

Exhibit 11. NAEP Standard Accommodations for ELs

Adaptations	Not vary?	All access?	Experience?	Category
Bilingual dictionary without definitions in any language (not in reading or writing)	Yes	Yes	Yes	Universal
Directions only read aloud in Spanish (not in TEL)	Yes	Yes	Yes	Universal
Spanish/English version of the test—not Grade 12 and only in mathematics, science, and civics-economics-geography-history	Yes	Yes	Yes	Universal
Test read aloud in Spanish (not Grade 12 mathematics) only in mathematics, science, and civics-economics-geography-history	Yes	Yes	Yes	Universal

Note. NAEP = National Assessment of Educational Progress; ELs = English learners; TEL = technology and engineering literacy.

In this analysis of function of the behavior over format of the process, little influence or differential boost should be expected on performance (with or without the adaptation). More importantly, conducting a functional analysis (as in a behavioral approach with an emphasis on discriminative stimuli and reinforcing consequences) may result in a positive impact for students, such as staying more engaged, being more attentive, and perhaps performing with more accuracy. This analysis also may suggest changes in the schedules of reinforcement from negative to positive. A negative reinforcement schedule occurs when students behave to terminate an aversive stimulus. In many large-scale testing programs, when low-performing students, those with disabilities or learning English, are confronted with texts and tasks that are unfamiliar and difficult, testing becomes an aversive stimulus that students want terminated. To mitigate this effect, the test purpose needs to be clear to the student, but, more importantly, barriers need to be eliminated (Bolt & Ysseldyke, 2008).

Rx 9: **Emphasize function over format of the process to reconsider classification of accommodations as universal design.** This reclassification, however, needs to involve an orderly process in addressing critical issues that negatively influence the standardization process. In this examination, questions can be asked about the emphasis on function, rather than the format of the process, as well as sources of influence for interpreting the construct. When adaptations can be made that do not vary in administration, provide access to all, and reflect the experiences of students, they could be classified as universal design.

SUMMARY OF RECOMMENDATIONS AND RESEARCH

Rethinking standardization in the NAEP testing program is complex with the need to be inclusive while continuing to use best practices in measurement that are consistent with the *Standards* (AERA et al., 2014). In this paper, several areas are addressed in which traditional NAEP notions of standardization might be challenged by considerations related to equity and access, especially in the context of evolving DBA: applying testing accommodations versus more provisions of universally available supports as well as the role of digital devices, interfaces, and administration. Several measurement features NAEP is pursuing or might consider relate to the notion of 'understandardization' as described by Sireci (2020). Within each section of the paper, specific recommendations are provided for future research and potential adaptations that NAEP may wish to consider in rethinking standardization in its context.

The long history of NAEP testing, particularly the past 30 years, includes attention to expanding the range of accommodations and universal design features for SDs and ELs. This practice has been extensively studied during this time span, with mixed findings. The general conclusion is that some accommodations are indeed effective for some students on some occasions. Generally, this research has been based on either an **interaction effect** (e.g., the accommodation improves performance for the target group, not the comparator group) or a **differential boost** (e.g., the accommodation improves performance for all students but more so for the target group). At the same time, many accommodations have resulted in small effect sizes, as reflected in many independent meta-analyses for both SDs and ELs, so our first two recommendations addressed this issue, using the research results to reconsider the definition of an accommodation. Using the concept of a just noticeable difference and the impact on proficiency categories, it might be possible to reconsider certain accommodations as universal design features.

Following the research on accommodations, comparisons are made on practices among NAEP and the states. With many specific universal design features, the NAEP testing program could be more consistent with large-scale testing practices. With both Smarter Balanced states and in the sample of all states from NCEO, many embedded and not embedded adaptations in testing programs are classified as designated supports or universal design. Another difference between NAEP and state testing programs is targeted sampling rather than census testing of student populations. To address this difference, a more refined and expansive sampling of students to reflect more diverse groups and subgroups of students in environments previously not considered might be explored. Part of this refinement is the sampling plan of institutions and students, which needs stratification by schools rather than states and districts to reflect the important population variations and programs more sensitively. Again, further research would be warranted with NAEP conducting small-scale studies.

These two changes in what is determined to be an acceptable test adaptation and for whom it could be used would allow for more consistency in NAEP definitions of test adaptations and those adopted by the states. Given the specificity in design and implementation of accommodations, including the population for whom they are targeted, a much simpler strategy would be to allow the adaptation to be provided to all students, who differ

extensively in ways other than the presence of a disability or learning English. Other student characteristics could be considered in allowing test adaptations to be deployed (e.g., using ZIP codes as a stand-in for geographic locations [urban, rural, remote]), including students from impoverished backgrounds, testing students with languages other than Spanish, and testing students who attend specialized environments (e.g., special school districts, prisons) and with various histories in learning opportunities. To this point, then, it is important to continue documenting important student characteristics beyond simple demographics. Given the importance of student characteristics and their influence on the effects of test adaptations, NAEP would be advised to support small-scale research studies to ensure that any proposed test adaptations are empirically supported.

Finally, options for rethinking standardization were speculated. First was the function and format of testing process to emphasize performance on test items that could be varied in settings, administrations, and responses without compromising interpretations. By strictly using narrow formats, generalizations are inherently limited to more restrictive constructs and student groups. While carefully analyzing the format of the process, many adaptations can nevertheless be made that accomplish the same function and therefore could be adapted to allow flexibility to the test setting, administration, or response. Furthermore, in many NAEP frameworks, the critical constructs being addressed appeared to be quite independent of format and, therefore, pave the way for a more comprehensive view in rethinking standardization. For example, the function of passage content can be considered as a possible influence on student performance (e.g., interest, background knowledge, opportunity to learn). Yet, most reading measurement constructs are broad (as they are in writing), thereby withstanding problems with construct deficiency or construct-irrelevant variance. In the end, adaptations need to be classified as an accommodation, a designated support, or universal design. With more specific criteria (answering the three questions), many adaptations fail to warrant their classification as an accommodation and therefore could be considered a universal design feature within the NAEP testing program.

Recent movements toward increasing equity and access, as well as rapidly expanding applications of technology are challenging traditional notions of standardized testing. For NAEP, standardization of conditions has always been a critical component of maintaining trend, and, therefore, a loosening of standardized testing conditions must be approached with great caution. Nevertheless, this paper has identified areas where, through research and related initiatives, NAEP can act to address the possibility that testing conditions may interact with personal characteristics in ways that hinder construct validity without diminishing its role as the Nation's Report Card.

Note

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues. This panel had reviewed this paper and approved it for publication, but this was canceled under the Trump administration; hence its publication as a BRT Technical Report.

Former Panel Members:

Keena Arbuthnot
Louisiana State University

Peter Behuniak
Criterion Consulting, LLC

Derek Briggs
University of Colorado Boulder

Jack Buckley
American Institutes for Research

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Gerunda Hughes
Howard University

Akisha Osei Sarfo
Council of the Great City Schools

James Pellegrino
University of Illinois at Chicago

Gary Phillips
Cambium Assessment

Jennifer Randall
University of Michigan

Lorrie Shepard
University of Colorado Boulder

David Thissen
University of North Carolina, Chapel Hill

Gerald Tindal
University of Oregon

Sheila Valencia
University of Washington

Denny Way
Consultant

Project Director

Sami Kitmitto
American Institutes for Research

Project Officer

Grady Wilburn
National Center for Education Statistics

REFERENCES

Abedi, J. (1999). *NAEP math test accommodations for students with limited English proficiency*. National Center for Research on Evaluation, Standards, and Student Testing. https://eric.ed.gov/?id=ED431787

Abedi, J. (2009). English language learners with disabilities: Classification, assessment, and accommodation issues. *Journal of Applied Testing Technology*, *10*(2). https://eric.ed.gov/?id=EJ865585

Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, *17*(4), 371–392. https://doi.org/10.1207/s15324818ame1704_3

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (CSE Technical Report 536). National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/newTR536.pdf

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *American Educational Research Journal*, *74*(1), 1–28. https://www.jstor.org/stable/3516059

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*(3), 16–26. https://doi.org/10.1111/j.1745-3992.2000.tb00034.x

Abedi, J., Lord, C., Kim, C., & Miyoshi, J. (2001). *The effects of accommodations on the assessment of LEP students in NAEP* (CSE Technical Report 537). National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/TR537.pdf

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Technical Report 429). National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/TECH429.pdf

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

American Psychological Association. (n.d.). Standardized test. In *APA dictionary of psychology*. https://dictionary.apa.org/standardized-test

Bolt, S. E., & Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential Item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment*, *26*(2), 121–138. https://doi.org/10.1177/0734282907307703

Burzick, H., & Stone, E. (2014). A meta-analysis of research on the read aloud accommodation. *Educational Measurement Issues and Practice*, *33*(3), 17–30. https://doi.org/10.1111/emip.12040

Campbell, J. R., & Donahue, P. L. (1997). *Students selecting stores: The effects of choice in reading assessment.* National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main1994/97491.pdf

Cawthon, S. W., & Leppo, R. (2013). Assessment accommodations on tests of academic achievement for students who are deaf or hard of hearing: A qualitative meta-analysis of the research literature. *American Annals of the Deaf*, *158*(3), 363–376. https://doi.org/10.1353/aad.2013.0023

Chiu, C. W. T., & Pearson, P. D. (1999). *Synthesizing the effects of test accommodations for special education and limited English proficient students* [Paper presentation]. Council of Chief State School Officers' National Conference on Large-Scale Assessment, Snowbird, UT, United States. https://eric.ed.gov/?id=ED433362

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, D., Tracy, R., & Cohen, J. (2017). On the effectiveness of pop-up English language glossary accommodations for EL students in large-scale assessments. Applied Measurement in Education, 30(4), 259–272. https://doi.org/10.1080/08957347.2017.1353986

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155Cohen, D., Tracy, R., & Cohen, J. (2017). On the effectiveness of pop-up English language glossary accommodations for EL students in large-scale assessments. *Applied Measurement in Education*, *30*(4), 259–272. https://doi.org/10.1080/08957347.2017.1353986

Cormier, D. C., Altman, J., Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007-2008* (Technical Report 56). University of Minnesota, National Center on Educational Outcomes. https://files.eric.ed.gov/fulltext/ED511744.pdf

Dabbs, P. (2003). *NAEP validity studies: An agenda for NAEP validity research.* American Institutes for Research. https://www.air.org/sites/default/files/2022-11/NVS-Agenda-NAEP-Validity-Research-Stancavage-2002-508.pdf

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Research-based recommendations for the use of accommodations in large-scale assessments.* University of Houston, Center on Instruction. https://files.eric.ed.gov/fulltext/ED517792.pdf

Fuchs, L. S., & Fuchs, D. (1999). Fair and unfair testing accommodations. *School Administrator*, *56*(10), 24–27, 29. https://eric.ed.gov/?id=EJ597145

Gallagher, C. J. (2003). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review*, *15*(1), 83–99. https://www.jstor.org/stable/23361535

Gregg, N., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, *45*(2), 128–138. https://doi.org/10.1177/0022219409355484

Harrison, J. R., Bunford, N., Evans, S. W., & Owens, J. S. (2013). Educational accommodations for students with behavioral challenges: A systematic review of the literature. *Review of Educational Research*, *83*(4), 551–597. https://doi.org/10.3102/0034654313497517

Hughes, G. B. (2023). *Improving equitable measurement and reporting in NAEP*. American Institutes for Research. https://www.air.org/sites/default/files/2023-09/EquitMeasReportNVS-2023-508.pdf

Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., & Rust, K. (2020). *2017 NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study* [White paper]. https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf

Johnstone, C. J., Altman, J., Thurlow, M., & Thompson, S. J. (2006). *A summary of research on the effects of test accommodations: 2002 through 2004* (Technical Report 45). University of Minnesota, National Center on Educational Outcomes. https://files.eric.ed.gov/fulltext/ED495886.pdf

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*(3), 1168–1201. https://doi.org/10.3102/0034654309332490

Li, H. (2014). The effects of read-aloud accommodations for students with and without disabilities: A meta-analysis. *Educational Measurement Issues and Practice*, *33*(3), 3–16. https://doi.org/10.1111/emip.12027

Li, H., & Suen, H. K. (2012). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education*, *25*(4), 327–346. https://doi.org/10.1080/08957347.2012.714690

Liu, K. K., Lazarus, S., Thurlow, M. L., Stewart, J., & Larson, E. (2020). *A summary of the research on test accommodations for English learners and English learners with disabilities: 2010-2018*. University of Minnesota, National Center on Educational Outcomes. https://files.eric.ed.gov/fulltext/ED605768.pdf

Lutkus, A. D., & Mazzeo, J. (2003). *Including special-needs students in the NAEP 1998 reading assessment. Part I: Comparison of overall results with and without accommodations—A report on 1998 NAEP research activities* [Statistical Analysis Report]. U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard//pdf/main1998/2003467.pdf

Mazany, T., Davy, L. E., Bushaw, W. J., & Stooksberry, L. (2017). *Writing Framework for the 2017 National Assessment of Educational Progress*. Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297–334. https://doi.org/10.1177/014662169301700401

National Academies of Sciences, Engineering, and Medicine. (2022). *A pragmatic future for NAEP: Containing costs and updating technologies*. The National Academies Press. https://doi.org/10.17226/26427

National Assessment Governing Board. (2014). *NAEP testing and reporting on students with disabilities and English language learners*. U.S. Department of Education. https://www.nagb.gov/content/dam/nagb/en/documents/policies/naep_testandreport_studentswithdisabilities.pdf

National Center for Education Statistics. (n.d.). *NAEP accommodations increase inclusiveness*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/about/accom_table.aspx

National Center for Education Statistics. (2000). *Increasing the participation of special needs students: A report on 1996 research activities*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/pdf/main1996/2000473.pdf

National Center for Education Statistics. (2005). *National Assessment of Educational Progress*. U.S. Department of Education, Institute of Education Sciences. http://nces.ed.gov/nationsreportcard/

National Center for Education Statistics. (2011). *Measuring status and change in NAEP inclusion rates of students with disabilities: Results 2007-09*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/pdf/studies/Inclusion_Highlights_2009.pdf

National Center on Educational Outcomes. (2021). *Accommodations toolkit*. University of Minnesota. https://publications.ici.umn.edu/nceo/accommodations-toolkit/introduction

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, *30*(3), 10–28. https://10.1111/j.1745-3992.2011.00207

Ricci, N. N. (2015). *The effect of the read-aloud testing accommodation on the 2011 fourth-grade National Assessment of Educational Progress in Reading, for students with disabilities in the New York State metropolitan, tri-state area* [Unpublished doctoral dissertation]. St. John's University, School of Education.

Rios, J. A., Ihlenfeldt, S. D., & Chavez, C. (2020). Are accommodations for English learners on state accountability assessments evidence-based? A multi-study systematic review and meta-analysis. *Educational Measurement: Issues and Practice*, *39*(4), 65–75. https://doi.org/10.1111/emip.12337

Rose, D. H. (2006). *A practical reader in universal design for learning*. Harvard Education Press.

Ross, S. M. (2020). Technology infusion in K-12 classrooms: A retrospective look at three decades of challenges and advancements in research and practice. *Educational Technology Research and Development*, *68*, 2003–2020. https://doi.org/10.1007/s11423-020-09756-7

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis*, *5*(3), 1–20. https://eric.ed.gov/?id=EJ580763

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research, and development series* (NCES 2005–457). U.S. Department of Education, National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pubs/studies/2005457.aspx

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105. https://doi.org/10.1111/emip.12377

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457–490. https://doi.org/10.3102/00346543075004457

Smarter Balanced. (2021). *Usability, accessibility, and accommodations guidelines*. https://files.eric.ed.gov/fulltext/ED617375.pdf

Srikanth, A. (2022). *Three essays on the equity and adequacy of K-12 school funding for English language learners: A national analysis* [Doctoral dissertation, Rutgers, The State University of New Jersey, School of Graduate Studies]. https://rucore.libraries.rutgers.edu/rutgers-lib/67597/PDF/1/play/

Tam, I. (2020). *The effect of the read aloud and extended time accommodations on NAEP fourth and eighth grade reading and mathematics for students with disabilities* [Doctoral dissertation, St. John's University, School of Education]. https://scholar.stjohns.edu/theses_dissertations/151

Tavani, C. M. (2007). *The impact of testing accommodations on students with learning disabilities: An investigation of the 2000 NAEP mathematics assessment* (Publication No. 3137495) [Doctoral dissertation, Florida State University]. ProQuest Dissertations and Theses Global. https://www.proquest.com/docview/305182301

Thompson, S. J., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001*. University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/TechReport34.pdf

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations*. University of Kentucky, Mid-South Regional Resource Center. https://eric.ed.gov/?id=ED442245

Tindal, G., & Ketterlin-Geller, L. R. (2004). *Research on mathematics test accommodations relevant to NAEP testing*. https://files.eric.ed.gov/fulltext/ED500433.pdf

Tindal, G., Nese, J. F. T., & Stevens, J. J. (2017). Estimating school effects with a state testing program using transition matrices. *Educational Assessment*, *22*(3), 189–204. https://doi.org/10.1080/10627197.2017.1344093

Vanchu-Orosco, M. (2012). *A meta-analysis of testing accommodations for students with disabilities: Implications for high-stakes testing*. https://digitalcommons.du.edu/etd/668/

Way, D., & Strain-Seymour, E. (2021). *A framework for considering device and interface features that may affect student performance on the National Assessment of Educational Progress*. American Institutes for Research. https://www.air.org/sites/default/files/Framework-for-Considering-Device-and-Interface-Features-NAEP-NVS-Panel-March-2021.pdf

Zenisky, A. L., & Sireci, S. G. (2007). *A summary of the research on the effects of test accommodations: 2005-2006* (Technical Report 47). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/Tech47/TechReport47.pdf

APPENDIX A. NATIONAL CENTER ON EDUCATION OUTCOMES RESEARCH SYNTHESIS AND POLICY ANALYSIS

Exhibit A1. Research Synthesis and Policy Analysis on Accommodations

Accommodation	*N* of studies	Main finding quotation	Reference	*N* of states allowed
Assistive technology	8	"In general, some students with disabilities who are in many different disability categories (e.g., blind or low vision, specific learning disabilities, TBI, autism, deaf or hard of hearing, emotional and behavioral disorders, speech or language disorders, intellectual disabilities, multiple disabilities, orthopedic impairment, OHI), including some English learners with disabilities, may benefit from an array of assistive technology devices."	Fleming, K., Ressa, V., Lazarus, S. S., Rogers, C. M., & Goldstone, L. (2022). *Assistive technology: Research* (NCEO Accommodations Toolkit #26a). University of Minnesota, National Center on Educational Outcomes.	27 AC 2 DS 0 UD
Braille	7	"Students with visual disabilities have benefited from reading braille versions of assessments."	Rogers, C., Hinkle, A. R., Ressa, V., Goldstone, L., & Lazarus, S. S. (2021). *Braille: Research* (NCEO Accommodations Toolkit #2a). University of Minnesota, National Center on Educational Outcomes.	40–46 AC 2 DS 0 UD
Calculator	15	"Overall, the performance of students with disabilities across all grade levels increased when a calculator was used regardless of the type of calculator (e.g., four-function, graphing, etc.) used."	Goldstone, L., Hendrickson, K., Lazarus, S. S., Ressa, V. A., & Hinkle, A. R. (2021). *Calculator use: Research* (NCEO Accommodation Toolkit #13a). University of Minnesota, National Center on Educational Outcomes.	23–42 AC 2–3 DS 2–41 UD
Clarify/simplify/ repeat directions	4	"The limited number of studies and the lack of recent data make it difficult to draw conclusions about the usefulness of this accommodation."	Goldstone, L., Hendrickson, K., Lazarus, S., Rogers, C. M., & Ressa, V. (2021). *Clarify/simplify/repeat directions: Research* (NCEO Accommodations Toolkit #11a). University of Minnesota, National Center on Educational Outcomes.	6 AC 10–12 DS 14 UD

Accommodation	*N* of studies	Main finding quotation	Reference	*N* of states allowed
Color contrast	5	"The research findings are mixed…The limited research suggests that color contrast enhancements may be more effective when used for math assessments than for reading assessments. The research also suggests that some students with visual impairments and some students with attention-related disabilities (e.g., ADHD) may find this accommodation useful."	Rogers, C. M., Lazarus, S., S, Ressa, V. A., Goldstone, L., & Fleming, K. (2022). *Color contrast: Research* (NCEO Accommodations Toolkit #25a). University of Minnesota, National Center on Educational Outcomes.	7 AC 22 DS 21 UD
Extended time	21	"The research on the effects of extended time is inconclusive. Some studies found benefit while others found mixed effects, or no effect."	Goldstone, L., Ressa, V., Lazarus, S. S., Hinkle, A. R., & Rogers, C. (2021). *Extended time: Research* (NCEO Accommodations Toolkit #6a). University of Minnesota, National Center on Educational Outcomes.	19–21 AC 1 DS 4–5 UD
Familiar proctor/test administrator	5	"Student performance has been shown to improve when familiar proctors have been used as an accommodation for young students with ASD."	Goldstone, L., Lazarus, S. S., Hendrickson, K., Rogers, C. M., Hinkle, A. R., & Ressa, V. (2021). *Familiar proctor/test administrator: Research* (NCEO Accommodations Toolkit #8a). University of Minnesota, National Center on Educational Outcomes.	4 AC 3 DS 6 UD
Highlighting	2	"Some students may benefit from the use of highlighting. The limited research suggests that digital highlighting may be more useful than the use of a physical marker to highlight text."	Ressa, V. A., Lazarus, S. S., Rogers, C. M., Hinkle, A. R., & Fleming, K. (2022). *Highlighting: Research* (NCEO Accommodations Toolkit #19a). University of Minnesota, National Center on Educational Outcomes.	4 AC 4 DS 41–44 UD
Human read aloud	18	"Some students may benefit from this accommodation—though many may not."	Goldstone, L., Lazarus, S. S., Hinkle, A. R., Rogers, C. M., & Ressa, V. A. (2022). *Human read aloud: Research* (NCEO Accommodations Toolkit #18a). University of Minnesota, National Center on Educational Outcomes.	23–36 AC 5–17 DS 1–3 UD
Large print	5	"Research has shown that there is some benefit for students with visual impairments when they use large print over standard print on paper exams."	Goldstone, L., Lazarus, S. S., Hendrickson, K., Rogers, C. M., & Hinkle, A. R. (2022). *Large print: Research* (NCEO Accommodations Toolkit #20a). University of Minnesota, National Center on Educational Outcomes.	36–37 AC 4 DS 0 UD

Accommodation	*N* of studies	Main finding quotation	Reference	*N* of states allowed
Magnification	0	"No research was found that examined whether magnification improved student performance, but several studies looked at student perceptions."	Goldstone, L., Lazarus, S. S., Hendrickson, K., Rogers, C. M., & Hinkle, A. R. (2022). *Magnification: Research* (NCEO Accommodations Toolkit #21a). University of Minnesota, National Center on Educational Outcomes.	14 AC 18–21 DS 25 UD
Manipulatives	7	"The research showed that the use of either physical or virtual manipulatives improved mathematics performance. This accommodation may be especially helpful for students with LD, ASD, and mild intellectual disabilities."	Goldstone, L., Hendrickson, K., Lazarus, S., & Fleming, K. (2021). *Manipulatives: Research* (NCEO Accommodation Toolkit #12a). University of Minnesota, National Center on Educational Outcomes.	32–46 AC 3–5 DS 2–3 UD
Mathematics charts/tables	2	"The research suggests that math charts benefit students with various disabilities at different grade levels, though it is inconclusive to what extent they are beneficial."	Goldstone, L., Lazarus, S. S., Hendrickson, K., Hinkle, A., & Rogers, C. (2022). *Math charts: Research* (NCEO Accommodations Toolkit #22a). University of Minnesota, National Center on Educational Outcomes.	15–31 AC 2–3 DS 1–2 UD
Multiple days	5	"However, testing over multiple days may benefit elementary students who struggle with reading. They are likely to experience fatigue while reading. The benefits of testing over multiple days appears to diminish as students move into middle school."	Ressa, V., Rogers, C., Lazarus, S. S., Hinkle, A. R., & Goldstone, L. (2021). *Multiple days: Research* (NCEO Accommodations Toolkit #3a). University of Minnesota, National Center on Educational Outcomes.	10–11 AC 2 DS 2 UD
Noise reduction	4	"There is some evidence noise reduction is a useful testing accommodation, though only a few studies have analyzed the use of this accommodation."	Goldstone, L., Lazarus, S. S., Olson, R., & Ressa, V. A. (2021). *Noise reduction: Research* (NCEO Accommodation Toolkit #15a). University of Minnesota, National Center on Educational Outcomes.	10 AC 22 DS 16 UD
Paper format	3	"Findings suggest that for students with disabilities the benefits of completing assessments on paper is dependent on individual characteristics and needs… However, research found that a majority of students allowed the print on demand option chose not to use it or used it sparingly."	Ressa, V. A., Lazarus, S. S., Hinkle, A. R., & Fleming, K. (2022). *Paper format: Research* (NCEO Accommodations Toolkit #28a). University of Minnesota, National Center on Educational Outcomes.	24 AC 1 DS 3 UD
Preferential seating	0	"No identified studies examined the effects of preferential seating on student performance, so there is a particular need for research in this area."	Goldstone, L., Hendrickson, K., Lazarus, S. S., Ressa, V., & Rogers, C. M. (2021). *Preferential seating: Research* (NCEO Accommodations Toolkit #10a). University of Minnesota, National Center on Educational Outcomes.	9 AC 18–21 DS 8 UD

Accommodation	*N* of studies	Main finding quotation	Reference	*N* of states allowed
Recorded oral delivery	12	"Overall, research on the recorded oral delivery of assessments provides mixed results on its effectiveness for students with various disabilities across grade levels on the ELA and mathematics assessments."	Goldstone, L., Hendrickson, K., Lazarus, S. S., & Hinkle, A. R. (2022). *Recorded oral delivery: Research* (NCEO Accommodations Toolkit #17a). University of Minnesota, National Center on Educational Outcomes.	6–8 AC 1 DS 0 UD
Scribe	3	"In general, research shows that students with disabilities who have difficulty with writing mechanics or the physical act of writing may benefit from the use of a scribe on both writing assessments and assessments of other content."	Ressa, V. A., Lazarus, S. S., Hinkle, A. R., & Rogers, C. M. (2021). *Scribe: Research* (NCEO Accommodation Toolkit #14a). University of Minnesota, National Center on Educational Outcomes.	38–49 AC 11–13 DS 0 UD
Signed administration	0	"The research on the effect of signed administration on performance is limited but suggests that the signed administration accommodation is beneficial."	Goldstone, L., Lazarus, S. S., Hendrickson, K., Rogers, C., & Fleming, K. (2022). *Signed administration: Research* (NCEO Accommodations Toolkit #24a). University of Minnesota, National Center on Educational Outcomes.	37–45 AC 3 DS 3–6 UD
Small group and individual administration	5	"Students who use these accommodations often use them in conjunction with other accommodations."	Fleming, K., Ressa, V. A., Lazarus, S. S., Rogers, C., & Hinkle, A. (2022). *Small group and individual administration: Research* (NCEO Accommodations Toolkit #23a). University of Minnesota, National Center on Educational Outcomes.	12 AC 15–16 DS 18 UD
Speech-to-text	6	"Speech-to-text is beneficial for some students with disabilities, including those with fine motor impairments that affect handwriting, across grade levels. Overall, students produced longer written text with fewer errors."	Goldstone, L., Lazarus, S. S., Olson, R., Hinkle, A. R., & Ressa, V. A. (2021). *Speech-to-text: Research* (NCEO Accommodations Toolkit #16a). University of Minnesota, National Center on Educational Outcomes.	24–34 AC 1 DS 0 UD
Spell check	2	"Research studies found that spell check is one of least assigned assessment accommodations… No studies were identified that examined the effect of spell check on student performance."	Goldstone, L., Lazarus, S. S., Hendrickson, K., Rogers, C., & Hinkle, A. R. (2022). *Spell check: Research* (NCEO Accommodations Toolkit #27a). University of Minnesota, National Center on Educational Outcomes.	2–5 AC 1 DS 10–20 UD

Accommodation	*N* of studies	Main finding quotation	Reference	*N* of states allowed
Student reads aloud to self	4	"There is mixed evidence regarding the usefulness of the student reads aloud to self-accommodation. Research suggests it may be helpful for some primary grade students. However, other factors such as type of text and setting may be responsible for a beneficial effect on student scores."	Goldstone, L., Lazarus, S. S., Olson, R., Hinkle, A. R., Ressa, V., & Rogers, C. M. (2021). *Student reads aloud to self: Research* (NCEO Accommodations Toolkit #9a). University of Minnesota, National Center on Educational Outcomes.	6–7 AC 19–20 DS 5 UD
Tactile graphics	10	"In general, teachers feel that they need more training on how to use tactile graphics, especially for newer tactile graphic technology options."	Lazarus, S. S., Hochstetter, A., Rogers, C. M., Ressa, V., Thurlow, M. L., & Liu, K. K. (2021). *Tactile graphics: Research* (NCEO Accommodations Toolkit #1a). University of Minnesota, National Center on Educational Outcomes.	19–26 AC 1–2 DS 0 UD
Test breaks	12	"Test breaks comprise one of the most frequently included accommodations on student IEPs, yet research on test breaks as an assessment accommodation on its own, and not bundled with other accommodations, is limited."	Ressa, V., Lazarus, S. S., Rogers, C. M., & Goldstone, L. (2021). *Test breaks: Research* (NCEO Accommodations Toolkit #7a). University of Minnesota, National Center on Educational Outcomes.	16 AC 5 DS 22–25 UD
Text-to-speech (computer generated voice)	9	"Across studies, students with disabilities have experienced positive effects, no effects, and negative effects from using this accommodation."	Ressa, V., Rogers, C., Lazarus, S. S., Hinkle, A. R., & Goldstone, L. (2021). *Text-to-speech (computer generated voice): Research* (NCEO Accommodations Toolkit #4a). University of Minnesota, National Center on Educational Outcomes.	26–33 AC 15–16 DS 1–5 UD
Word prediction	4	"There is a need for additional research on the use of word prediction."	Goldstone, L., Lazarus, S. S., Ressa, V., Rogers, C., & Hinkle, A. R. (2021). *Word prediction: Research* (NCEO Accommodations Toolkit #5a). University of Minnesota, National Center on Educational Outcomes.	5–16 AC 1 DS 0–2 UD

Note. AC = accommodations; DS = designated support; UD = universal design.