Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2025, Vol. 44, No. 4, pp. 18–32

Universal by Design: Unveiling the Effectiveness of Accommodations and Universal Design Features through Process Data

Burhan Ogut, *American Institutes for Research*
Michelle Yin and Hoa Vu, *Northwestern University*
Juanita Hicks and Ruhan Circi, *American Institutes for Research*

Abstract: *This study examines the relationship between testing accommodations, universal design (UD) features, and standardized test performance within a digital testing framework. Utilizing a novel dataset—the National Assessment of Educational Progress (NAEP) process data—and advanced statistical methodologies such as inverse probability weighting and doubly robust models, we provide compelling evidence on the utility of specific accommodations and UD features. Extended time usage is associated with improved test performance for students with disabilities, emphasizing the crucial role it plays in fostering equitable testing conditions. While scratchwork is associated with improved test performance for students with disabilities, the associations between text-to-speech and equation editors and student performance are mixed, suggesting that these features may act as potential distractions rather than support for some students.*

Keywords: equation editor, extended time accommodation, NAEP, scratchwork, standardized test performance, students with disabilities, text-to-speech, universal design features

Introduction

The implementation of standardized testing in the United States, notably escalated by the No Child Left Behind Act of 2001, sought to uniformly assess student and school performance. However, this approach presents unique challenges for students with disabilities (SWDs) due to the rigid structure of these assessments. The expectation of uniform completion within a fixed format is particularly problematic for SWDs. For instance, a student with a vision impairment facing a paper-and-pencil test, or a student with learning disabilities being expected to perform within the same timeframe as their non-disabled peers, raises questions of equity and fairness.

This scenario has ignited a heated debate on eliminating assessment barriers that prevent SWDs from demonstrating their full potential. Concerns about the equitable distribution of testing accommodations have emerged, highlighting the potential for misuse or unfair advantages, such as students fabricating disabilities to gain access to these accommodations (Balasa et al., 2019; Mitchell, 2012; Tapper et al., 2006). Conversely, proponents of accommodations assert that these measures level the playing field, allowing SWDs to fairly demonstrate their abilities. They argue that accommodations remove construct-irrelevant factors, thus preserving the integrity of the assessment and enabling a true reflection of a student's abilities (Fuchs et al., 2005; Sireci et al., 2005; Thurlow et al., 2003; Thurlow et al., 1993).

In the 2022–2023 academic year, a significant portion of K–12 students in the United States were recognized under legal frameworks designed to support their educational needs due to disabilities. Specifically, 15% of students were recipients of an Individualized Education Program (IEP), as mandated by the Individuals with Disabilities Education Act (IDEA) (National Center for Education Statistics, 2024). The IEP is a customized educational plan designed to meet the unique needs of a student with a disability, ensuring they receive personalized support and accommodations in their learning environment. The accommodations and services provided under an IEP can include specialized instruction, speech therapy, behavioral interventions, and other supports. Additionally, 2–3% of students were covered under a 504 Accommodation Plan, which is provided under Section 504 of the Rehabilitation Act of 1973. The key difference between an IEP and a 504 plan is that an IEP provides specialized instruction and tailored services to students with disabilities who require significant support to access the curriculum, while a 504 plan focuses on providing accommodations to ensure equal access to general education for students with disabilities who can largely participate in regular classes with modifications. Essentially, an IEP is more comprehensive and provides more intensive support than a 504 plan.

These legal provisions are crucial not only for guaranteeing the educational access of SWDs but also for upholding fairness and validity in their assessment experiences. Fairness in this context refers to providing SWDs with equal opportunities to demonstrate their knowledge and skills without being disadvantaged by their disabilities, while validity ensures that test scores accurately reflect the intended constructs rather than irrelevant barriers, such as the lack of ac-

commodations. According to the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), both fairness and validity are essential components in designing and implementing assessments, particularly for populations requiring accommodations.

Accommodations play a critical role in addressing these principles by reducing construct-irrelevant variance—factors unrelated to the knowledge or skills being assessed—that can arise when SWDs face barriers during testing. For example, accommodations such as extended time or alternative formats help mitigate the influence of disabilities on test performance, aligning with the *Test Standards*' emphasis on providing equitable assessment conditions. Despite the critical importance of these accommodations, our understanding of their association with academic performance is often based on small-scale studies that primarily examine test presentation and timing accommodations (Christensen et al., 2011; Sireci et al., 2018). Furthermore, data on the actual utilization of these accommodations are sparse, frequently collected through methods such as surveys or observational studies, which may not fully capture the breadth of their impact (Bone & Bouck, 2018; Elliott & Marquart, 2004; Lang et al., 2008; Sireci et al., 2005). While our data come from a digital testing environment, we acknowledge that the principles of universal design in assessment have roots that predate computer-based delivery. Seminal studies in the early 2000s laid the foundation for applying universal design principles in large-scale assessments, emphasizing inclusive design from the outset (e.g., Thompson et al., 2002).

This paper seeks to advance our understanding of the relationship between the usage of testing accommodations and universal design features—elements intentionally incorporated into assessments to ensure they are accessible, usable, and equitable for the widest range of individuals, regardless of ability or disability—and test performance among students with and without disabilities, utilizing a unique dataset: process data from the National Assessment of Educational Progress (NAEP). NAEP process data include detailed information on how students interact with the digital testing platform, such as response times, the use of UD features, and navigation patterns during the test. This novel dataset offers an unprecedented opportunity to define the utilization of accommodations and universal design features accurately and objectively, enabling a comprehensive examination of their association with test performance.

Our study is guided by three principal questions: (1) Is there a correlation between the use of extended time accommodations, as granted through 504/IEP plans, and improved test performance for SWDs? Extended time is one of the most commonly granted accommodations under IEPs. Given its widespread use, we aim to investigate whether extended time correlates with improved outcomes for SWDs. (2) Does the use of universal design features—specifically, scratchwork, text-to-speech, and the equation editor—correlate with better test performance among both SWDs and their non-disabled counterparts (students without disabilities, SWODs)? These features were selected because they are among the most commonly used universal design features. (3) Does the relationship between UD feature usage and test performance differ for SWDs based on their use of extended time accommodations?

Our findings highlight a significant relationship between the use of extended time accommodations and test performance among SWDs who received such accommodations. On average, SWDs who utilized extended time accommodations outperformed their peers who did not by 3.5 percentage points on test items. This finding corroborates the utility of extended time as a crucial support mechanism, allowing SWDs to thoroughly engage with test content without the constraints of stringent time limitations hampering their ability to perform.

The analysis of universal design features produced mixed results. Scratchwork emerged as a beneficial feature for SWDs, potentially aiding in comprehension and reducing cognitive load. The association of text-to-speech, however, was more complex, not showing a significant benefit for SWDs while unexpectedly revealing a negative correlation with SWODs' performance. This suggests that, although text-to-speech may not universally enhance the performance of SWDs, it could potentially serve as a distraction for SWODs. Similarly, the use of the equation editor presented as a potential source of distraction for both SWDs and SWODs during tests.

This study leverages the unique insights provided by NAEP process data, offering a fresh perspective on the efficacy of accommodations and universal design within the realm of digital standardized testing. Our findings contribute valuable knowledge to the existing body of literature and lay the groundwork for informed future policies and educational practices aimed at achieving equitable assessment for all students. The rest of the manuscript is organized as follows: we begin by discussing the rationale behind testing accommodations and reviewing relevant literature. This is followed by an examination of the data, measures, and methodologies employed in our study. We then present and discuss our findings, concluding with a summary of the implications of our research for educational policy and practice.

Background and Relevant Literature

Rationale for Testing Accommodations

Ensuring fairness in testing hinges on the validity of the inferences drawn from test scores. Validity focuses on the accuracy of conclusions drawn from test results and the decisions made based on these conclusions. If a test misrepresents the abilities of a particular group of students, it can lead to unintended and possibly detrimental outcomes. This issue is relevant even in low-stakes assessments like NAEP, which, despite seeming less critical, is used for various significant purposes. For instance, NAEP results are used to assess the effectiveness of various federal or state educational interventions (Dee & Jacob, 2011; Wong et al., 2015). In high-stakes assessments, the implications are even more profound, affecting decisions like grade promotion or repetition, which have lasting impacts on a student's life. Therefore, it is essential to use assessment features that yield valid results to accurately measure students' knowledge and abilities (Fuchs et al., 2005).

Two primary threats can compromise the validity of test scores and their interpretations. The first is construct underrepresentation, where a test fails to encompass all aspects of the subject it is intended to measure. For instance, a general mathematics test focusing solely on algebra items is an example of this threat. The second threat, which is the focal point of this study, is construct-irrelevant variance. This occurs when extraneous factors in the test, such as complex language in a mathematics test, unfairly disadvantage certain

Table 1

Type of Accommodations Provided to Students

	Granted Accommodation Features				
	Extended Time		None		Total
	N	%	*N*	%	*N*
Students with disabilities	1,670	59.9	1,120	40.1	2,790
Students without disabilities	440	1.7	24,930	98.3	25,370

Note: Data are from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment. Unweighted number of observations may not sum to total due to the requirement that the reported sample sizes must be rounded to the nearest 10.

student groups, thereby affecting their performance and the interpretation of their scores.

To address this second threat, testing accommodations are provided to remove barriers related to students' disabilities, enabling them to demonstrate their true competencies. The 2014 *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) specify that accommodations should eliminate irrelevant obstacles. This involves adjustments in the test administration setting, presentation style, user interface, engagement methods, response requirements, and possibly the inclusion of additional personnel, like a human reader (Standard 3.9, p. 67). These standards also outline the responsibilities of those developing and implementing test accommodations. These responsibilities include ensuring accommodations for students with documented needs, providing these accommodations in testing situations, offering adequate training and information to those involved in the accommodation process, and documenting the use of these accommodations (Standard 3.10, pp. 65–66).

Historically, the provision of accommodations for students with disabilities, including those with learning disabilities, physical impairments, and hearing or visual impairments, has involved adjustments in how test content is presented. These accommodations have traditionally included the use of a scribe, read-aloud accommodations, or Braille (Clapper et al., 2005; Pitoniak & Royer, 2001; Thurlow et al., 1993). Moreover, accommodations like oral administration, whether through teacher delivery, student reading, or screen-reading software, and extended time for test completion have been widely studied and implemented (Buzick & Stone, 2014; Hicks et al., 2019; Sireci et al., 2005). However, the landscape of educational assessments is evolving with the increasing digitization of tests and the advancements in assistive technologies. This evolution is prompting a shift from traditional accommodations to the integration of universal design (UD) features, which are not only more cost-effective but also offer enhanced flexibility and inclusivity.

Importantly, the conceptual foundation of universal design in assessment predates the digital transition. Thompson et al. (2002) outlined core UD principles including inclusive test populations, accessible item formats, and legibility, all of which informed assessment policy throughout the 2000s. These principles remain relevant regardless of test modality.

UD in assessment aims to make tests accessible to all students right from the start, embodying principles that ensure inclusivity, precision in construct definition, unbiased test items, and flexibility in accommodations. Moreover, UD emphasizes the importance of straightforward instructions, optimal readability and comprehensibility of test materials, and maximum legibility. Thus, the digital delivery of NAEP provides a modern lens through which to evaluate the legacy of these longstanding design principles. The shift toward universal design represents a fundamental rethinking of how educational assessments are developed and implemented, with the goal of ensuring equitable access and opportunity for all students to demonstrate their knowledge and skills.[1]

Existing Evidence on the Effectiveness of Accommodations and Universal Design Features

Extended time. Extended Time (ET) is a critical accommodation for SWDs, allowing for additional test-taking time or flexible scheduling. Studies have shown that ET is widely used and often paired with other accessibility features (AFs) to support diverse needs (Christensen et al., 2011; Elliott et al.2015, Sireci et al., 2005; Sireci et al., 2018). Elliott and Marquart (2004) noted a slight performance increase in Grade 8 mathematics for both SWDs and SWODs under ET conditions, although differences between the groups were not significant. Conversely, Sireci et al. (2005) found that ET particularly benefits SWDs, highlighting its role in leveling the playing field. Recent studies leveraging NAEP process data offer nuanced insights: Kim and Circi (2018) reported that only 51% of students with ET accommodations utilized them, and those who did demonstrated higher ability estimates. Witmer et al. (2023) further confirmed that the utilization of ET among eligible students correlated with improved academic performance. Wei and Zhang (2024) also used the NAEP process data and showed that students with learning disabilities who used ET scored better than students with learning disabilities who did not use ET. Although this study employed propensity score weighting, the model relied on a limited set of variables, primarily focusing on demographics. Suk et al. (2022) and Suk and Kim (2024) also utilized NAEP 2017 process data and response data to apply a new regression discontinuity approach to study the impact of ET on student performance. However, their results did not reveal any significant impact of ET on student performance for English language students or SWDs. One potential reason is the small sample, resulting in large standard errors. Another reason is probably the use of English proficiency as a forcing variable. The provision of ET in NAEP is not necessarily determined by students' EL proficiency but rather by many factors, as discussed earlier, including input from parents, teachers, and administrators. Therefore, the integrity of the forcing variable in the regression discontinuity approach may not be strong.

To evaluate whether accommodations or universal design features enhance assessment fairness, researchers have often relied on the interaction hypothesis. This framework

Table 2

Summary Statistics for Students with Disabilities

	Students with Disabilities Who Are Not Eligible for Extended Time		Students with Disabilities Who Are Eligible for Extended Time				All Students with Disabilities	
			Used Extended Time		Not Used Extended Time			
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Female	.31	.46	.36	.48	.36	.48	.34	.47
Black	.14	.34	.18	.38	.15	.36	.15	.36
White	.51	.5	.41	.49	.47	.5	.47	.5
Hispanic	.22	.41	.31	.46	.29	.45	.26	.44
Asian American or Pacific Islander	.03	.18	.02	.15	.02	.13	.03	.16
American Indian or Alaska Native	.04	.2	.02	.15	.02	.12	.03	.16
More than one race	.06	.24	.05	.23	.06	.24	.06	.24
English language learners	.09	.29	.11	.32	.09	.28	.09	.29
School lunch program eligibility	.57	.49	.59	.49	.6	.49	.59	.49
Total raw score	5.84	3.71	6.57	3.80	5.22	3.41	5.73	3.64
Observations	1,120		530		1,140		2,790	

Note: This table presents summary statistics for students with disabilities, separated by their eligibility and use of extended time accommodations. The demographic variables provide context on the sample composition, while the total raw score indicates average performance on the assessment. The test block had a maximum possible raw score of 25. Seven items were scored dichotomously, with 1 for correct and 0 for incorrect responses. Additionally, seven items had a maximum score of 2, while one item had a maximum score of 4. Unweighted number of observations are shown and rounded to the nearest 10 for reporting requirements. Data are from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

Table 3

Extended Time Use and Performance among Students with Disabilities

	Outcome: Item Correct		
	Logit†	Inverse Probability Weighting	Doubly Robust
	(1)	(2)	(3)
Extended time use	.03451***	.0417***	.0419***
	(.006)	(.006)	(.007)
Observations	24,420	24,420	24,420
Student characteristics	√	√	√
Parental controls	√	√	√
Item characteristic controls	√	√	√
School controls	√	√	√

Note: Standard errors in parentheses. †Marginal effects are reported for logistic regression models. Unweighted number of observations are shown and rounded to the nearest 10 for reporting requirements. Data are from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment. *$p < .1$, **$p < .05$, ***$p < .01$

posits that accommodations should benefit SWODs without unfairly advantaging them over peers. Studies employing this hypothesis compare performance across accommodated and non-accommodated conditions, often using experimental or quasi-experimental designs (Fuchs et al., 2005; Sireci et al., 2005). This logic extends to UD evaluation, although the inherently non-targeted nature of UD features pose methodological challenges for attribution.

Read-aloud or text-to-speech. Transitioning from read-aloud to text-to-speech (TTS) technology reflects advancements in digital assessments and assistive technology. Traditional read-aloud accommodations, involving test content delivery by humans or through audio recordings, have been effective, especially in non-reading assessments like mathematics (Buzick & Stone, 2014). These accommodations are applied within the specific context of the assessment to address accessibility barriers without compromising the validity of the test. Read-aloud accommodations are typically used in assessments that do not directly measure reading skills, like mathematics or science tests, ensuring the construct being assessed remains intact. Research supports the efficacy of such accommodations. For instance, Weston (2002) found that oral accommodation improved mathematics performance for both SWDs and SWODs, with SWDs benefiting more. This finding supports the interaction hypothesis that SWDs gain more from certain accommodations (Calhoon et al., 2000; Huynh et al., 2004; Sireci et al., 2005). TTS technology, as a more autonomous and adaptable approach, has shown that while its usage is widespread among SWODs and accommodated SWDs, its relationship with performance varies, indicating a potential distraction for non-accommodated students but a benefit for accommodated ones (Lee et al., 2020).

In a recent study, Wei (2024) explored the predictors of TTS use using data from the 2017 NAEP Mathematics process data and identified mathematics proficiency and time pressure felt by students as factors associated with TTS use for both SWDs and SWODs. Wei (2024) also reported that test performance was positively associated with TTS use for SWDs (only the frequency of use) and English learner students (both use and frequency) with ET accommodation. However, there were a few limitations to this study. First, the results were mostly descriptive that relied on regression models without attempting to control for selection bias. Moreover, the set of covariates omitted several important predictors of student performance such as parental education or socioeconomic status indicators, or school level characteristics. It is not clear if the observed relationships would still hold when these variables are included in the models. Second, there was no differentiation between a meaningful use of TTS versus non-meaningful use where a student may just turn on and off the TTS without using it. Not distinguishing between these two uses may also explain some of the results obtained only for the frequency of use but not for the binary indicator of use.

Scratchwork and equation editor. In NAEP, new on-screen features like scratchwork and equation editor features have been introduced as part of its UD features. While these developments align with broader shifts toward digital assessments, it is important to acknowledge that the principles of universal design for assessments were established prior to the advent of digital testing (Thompson et al., 2002). Scratchwork includes modes for pencil and highlighter, allowing students to annotate, calculate, sketch diagrams, and highlight sections across both multiple-choice and short response items. Equation editor, accessible for selected items, enables students to input mathematical symbols and numbers, with interactive tutorials provided at the start of the assessment for familiarization.

Research on the utilization of these UD features is still in nascent stages. Hicks et al. (2019) conducted a study on their use in the NAEP 2017 Grade 8 mathematics assessment, but did not specifically focus on SWDs. Their findings indicated that 34% of students used the scratchwork feature, 19% utilized the highlighter, and 38% engaged with the equation editor. The scratchwork feature was predominantly used in fill-in-the-blank items, whereas the highlight feature saw less than 12% usage across all item types. The equation editor, limited to fill-in-the-blank and composite constructed-response items, was more frequently used in fill-in-the-blank questions.

Gaps in existing research. The existing research broadly indicates that certain accommodations are advantageous for SWDs, while others benefit both SWDs and SWODs (Römhild & Hollederer, 2024). Nonetheless, the literature has notable

Table 4

Universal Design Usage and Demographic Characteristics for All Students

	Student with Disability				Student without Disability			
	Used Universal Design		Not Used Universal Design		Used Universal Design		Not Used Universal Design	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Female	.34	.47	.34	.47	.48	.5	.53	.5
Black	.16	.37	.14	.35	.14	.34	.14	.35
White	.46	.5	.48	.5	.48	.5	.47	.5
Hispanic	.26	.44	.26	.44	.25	.43	.25	.43
Asian American or Pacific Islander	.03	.16	.03	.16	.05	.22	.06	.23
American Indian or Alaska Native	.03	.17	.03	.16	.02	.14	.02	.13
More than one race	.06	.24	.06	.24	.06	.24	.06	.25
English language learners	.1	.3	.09	.29	.06	.23	.05	.22
School lunch program eligibility	.59	.49	.59	.49	.49	.5	.46	.5
Total raw score	5.78	3.52	5.68	3.73	9.24	4.6	9.01	4.72
Observations	1,260		1,530		8,810		16,600	

Note: The test block had a maximum raw score of 25. Seven items were scored dichotomously, with 1 for correct and 0 for incorrect responses. Additionally, seven items had a maximum score of 2, while one item had a maximum score of 4. The utilization of universal design is determined by whether the student used any of the following features: Text-to-Speech, Scratchwork, or Equation Editor. Conversely, a student is considered to have not used universal design if they did not utilize any of the three mentioned features. Unweighted number of observations are shown and rounded to the nearest 10 for reporting requirements. Data are from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

Table 5

Scratchwork Use and Performance for Students with and without Disabilities by Extended Time Use

	Outcome: Item Correct		
	Logit†	Inverse Probability Weighting	Doubly Robust
	(1)	(2)	(3)
Panel A: Students with Disabilities			
Scratchwork Use	.008	.0345***	.010
	(.007)	(.008)	(.008)
Observations	40,080	40,080	40,080
Panel B: Students without Disabilities			
Scratchwork Use	.004*	.022***	.003
	(.002)	(.002)	(.002)
Observations	363,640	363,640	363,640
Panel C: Students with Disabilities Who Used Extended Time			
Scratchwork Use	.068	.027*	.019
	(.013)	(.014)	(.014)
Observations	7,770	7,770	7,770
Panel D: Students with Disabilities Who DID Not use Extended Time			
Scratchwork not use	.008	.035**	.006
	(.012)	(.014)	(.014)
	16,590	16,590	16,590
Student characteristics	√	√	√
Parental controls	√	√	√
Item characteristic controls	√	√	√
School controls	√	√	√

Note: Standard errors in parentheses. † Marginal effects are reported for logistic regression models. Unweighted number of observations are shown and rounded to the nearest 10 for reporting requirements. Data are from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment. *$p < .1$, **$p < .05$, ***$p < .01$

limitations, particularly in scale and generalizability (Buzick & Stone, 2014; Sireci et al., 2005). This study aims to fill these gaps by employing a large, nationally representative dataset and analyzing objective process data from NAEP. Such an approach contrasts with earlier research that often relied on subjective methods like surveys or observations, offering a more detailed and accurate picture of how accommodations and UD influence test performance (Bone & Bouck, 2018; Elliott & Marquart, 2004; Lang et al., 2008).

Data and Measures

Data

Our analysis leverages data from the 2017 administration of the National Assessment of Educational Progress (NAEP), focusing on the Grade 8 mathematics test. These data allow us to explore the relationship between the use of ET and universal design features (including scratchwork, text-to-speech, and the equation editor) and student performance outcomes. The NAEP, often dubbed "The Nation's Report Card," systematically selects students for assessment, ensuring a nationally representative sample. In 2017, approximately 148,100 Grade 8 students participated in the mathematics assessment, reflecting a broad demographic spectrum across the United States.

The National Assessment Governing Board (NAGB) sets a target inclusion rate of at least 85% for SWDs and English language learners (ELLs) in the National Assessment of Educational Progress (NAEP). This inclusion rate is intended to ensure that the assessment does not disproportionately exclude these students and that the results accurately reflect the performance of these student populations. The 2017 assessment successfully exceeded these benchmarks, with inclusion rates of 89% for SWDs and 90% for ELLs. Among the students assessed, 13% were identified as SWDs, with 12% actually participating in the assessment. Within this subgroup, 2% were assessed without any accommodations, while 10% received various accommodations. This differentiation underscores that our study's findings are primarily generalizable to SWDs who participated in the NAEP with accommodations.

The NAEP mathematics assessment is structured into three main sections: a tutorial section, a survey questionnaire, and an assessment component comprising two cognitive blocks, each initially allotted 30 minutes but extendable to 90 minutes for students with ET accommodations. The tutorial segment is designed to familiarize students with the digital platform, mitigating potential disadvantages across diverse student groups (Kikis-Papadakis & Kollias, 2009; Sandene et al., 2005). The survey questionnaire collects demographic information and self-reported

Table 6

Text-to-Speech Use and Performance for Students with and without Disabilities by Extended Time Use

	Outcome: Item Correct		
	Logit†	Inverse Probability Weighting	Doubly Robust
	(1)	(2)	(3)
Panel A: Students with Disabilities			
Text-to-Speech Use	.009	–.002	.002
	(.006)	(.007)	(.008)
Observations	40,080	40,080	40,080
Panel B: Students without Disabilities			
Text-to-Speech Use	–.0002	–.016***	–.009*
	(.003)	(.005)	(.004)
Observations	363,640	363,640	363,640
Panel C: Students with Disabilities Who Used Extended Time			
Text-to-Speech Use	–.015	–.023	–.018
	(.012)	(.017)	(.015)
Observations	7,770	7,770	7,770
Panel D: Students with Disabilities Who Did Not Use Extended Time			
Text-to-Speech Use	.024**	.025*	.027**
	(.009)	(.012)	(.011)
Observations	16,590	16,590	16,590
Student characteristics	√	√	√
Parental controls	√	√	√
Item characteristic controls	√	√	√
School controls	√	√	√

Note: Standard errors in parentheses. † Marginal effects are reported for logistic regression models. Unweighted number of observations are shown and rounded to the nearest 10 for reporting requirements. Data are from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$

data on test-taking experiences and familiarity with digital assessments.[2] The cognitive portion consists of ten assessment blocks, from which each student is assigned two, with each block including 15 test items, integrated into 50 different forms through a matrix sampling design.[3] This approach ensures the comparability and random equivalence of the sample groups.

This study harnesses the NAEP's release of response process data, offering a nuanced glimpse into student interactions during a portion of the assessment. Specifically, it examines data from one of ten assessment blocks, encompassing around 28,000 students, alongside more detailed data from one of fifty forms involving approximately 2,800 students. The block data, incorporating a substantial cohort of students with disabilities, capture only a segment of the full assessment experience. In contrast, the form data deliver a comprehensive overview for a smaller group of participants.[4] Given its broader scope, our analysis primarily leverages the block data. This choice allows us to focus on a larger dataset to explore how the utilization of accommodations and universal design features correlates with academic outcomes across a more representative sample of students.

Measures and Variables

Outcome variable. Our primary outcome variable measures a student's performance on an individual item, assigned a value of 1 if the student answers a test item correctly, and 0 otherwise. This particular block comprised 15 mathematics items.

Students with disabilities. Students with disabilities are defined as those with either an IEP plan or a 504 plan.

Use of extended time. To assess the utilization of ET accommodations, we calculated the total time spent on the assessment block, measured from the first to the last action within the block. Students eligible for ET who spent over 30 minutes on the block were considered to have utilized the ET accommodation provided to them, as such ET use is a student-level variable in our analysis.

Use of scratchwork. Scratchwork utilization was classified based on the total active time using the feature. Students who engaged with the scratchwork feature (e.g., highlight-

Table 7

Equation Editor Use and Performance for Students with and without Disabilities by Extended Time Use

	Outcome: Item Correct		
	Logit†	Inverse Probability Weighting	Doubly Robust
	(1)	(2)	(3)
Panel A: Students with Disabilities			
Equation editor use	–.320***	–.225***	–.213***
	(.022)	(.010)	(.009)
Observations	40,080	40,080	40,080
Panel B: Students without Disabilities			
Equation editor use	–.305***	–.281***	–.341***
	(.007)	(.004)	(.008)
Observations	363,640	363,640	363,640
Panel C: Students with Disabilities Who Used Extended Time			
Equation editor use	–.145***	–.235***	–.244***
	(.033)	(.019)	(.021)
Observations	7,770	7,770	7,770
Panel D: Students with Disabilities Who Did Not Use Extended Time			
Equation editor use	–.210***	–.223***	–.190***
	(.038)	(.059)	(.011)
Observations	16,590	16,590	16,590
Student characteristics	√	√	√
Parental controls	√	√	√
Item characteristic controls	√	√	√
School controls	√	√	√

Note: Standard errors in parentheses. † Marginal effects are reported for logistic regression models. Unweighted number of observations are shown and rounded to the nearest 10 for reporting requirements. Data are from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment. ***$p < .01$.

ing, drawing) for at least 2 seconds after opening it were defined as active users (additional information can be found in the supplementary file). The 2-second threshold was used to identify meaningful interactions, balancing the need to exclude accidental openings while capturing brief but intentional use. This approach for identifying active use of UD features aligns with practices in the process data literature, where minimal time thresholds are often used to distinguish active engagement from passive or unintentional actions (e.g., Michaelides & Ivanova, 2022; Wise et al., 2010). In contrast to ET, scratchwork and the other UD features in this study are student-by-item variables (i.e., their use varies by student and item)

Use of text-to-speech. Text-to-Speech (TTS) usage is defined as having at least one sentence read aloud to the student after activating TTS. This threshold ensures that TTS was meaningfully employed as a support feature, rather than being activated unintentionally or out of curiosity without actual use.

Use of equation editor. Similarly, students who engaged with the editor by pressing keys for inputs after activating it for more than 2 seconds are considered users. The 2-second threshold was chosen to exclude incidental activation while capturing deliberate engagement, consistent with the rationale used for scratchwork.

Item variables. Item characteristics are defined by NAEP and include item difficulty (easy, medium, or hard), average word length, average sentence length, Coleman–Liau readability index, Flesch–Kincaid readability grade level, Linsear Write readability score, and Spache grade. Descriptive statistics for items can be found in the supplementary file.

Student variables. Our study incorporates a range of student demographic characteristics, including student age, sex, race/ethnicity, disability status, eligibility for free school lunch, and a dummy for students classified as English language learners. Descriptive statistics for student characteristics can be found in Supplementary File.

Parental variables and school variables. We also include both parental and school-level variables in our models. Parental variables include parental education, while school-level variables include public school status and the percentage of students by race/ethnicity (White, Black, Hispanic, and Asian).

Analysis

Extended Time Usage and Performance among Student with Disabilities

To investigate the first research question concerning the correlation between ET accommodations and test performance among SWDs, we employed a multi-faceted analytical approach. First, we utilized a generalized linear model (GLM) to establish a baseline understanding:

$$P(Y_{ij} = 1) = \Lambda\left(\beta_0 + \beta_1 ET\_Use_{ij} + \beta_2 X_{ij}\right), \quad (1)$$

where Y_{ij} represents a binary variable equal to 1 if student (i) gets correct score in test item j. ET_Use_{ij} refers to extended time usage. X_{ij} is a vector comprising: (i) student characteristics such as age, sex, race/ethnicity, eligibility for free school lunch, and a dummy for students classified as English language learners; (ii) parental educational attainment; and (iii) test item characteristics such as item difficulty, number of sentences, average word length, average sentence length, Coleman–Liau readability index, Flesch–Kincaid readability grade level, Linsear Write readability score, and Spache grade. The term, $\mathbf{\Lambda}(\cdot)$, is the standard logistic function used for the GLM.

This logit model compared the performance of students who used Extended Time with those who are granted extended time but did not use them at all. The parameter of interest, β_1, captures the association between extended time usage and test performance.

Although we anticipate that adjusting for observable student, parental, and test item characteristics might help reduce the selection bias into extended time usage, we opted to conduct two additional methods: inverse probability weighting models and doubly robust models. These additional analyses are intended to provide further insight into the selection bias, aiming to better understand the associations between accommodation and universal designs usage and performance.

Inverse probability weighting models. We employ inverse probability weighting (Austin & Stewart, 2015) to mitigate potential selection bias that may arise because students who choose to use extended time might differ systematically from those who do not. The inverse probability weighting (IPW) models can be implemented in several steps.

We first developed a logistic regression model to estimate the propensity score for each student:

$$P_i = \Lambda\left(\beta' X_i\right), \quad (2)$$

where P_i is the probability of using extended time for student i, X_i is a vector of student, parental, and item characteristics same as in Equation (1). These variables are selected due to their theoretical and empirical relationships with both the outcome variable and the use of extended time accommodation and UD features (Brookhart et al., 2006).[5]

We then calculated the inverse probability weights based on the propensity score. Specifically, for students who used extended time, the weights are defined as the inverse of the propensity score ($\frac{1}{P_i}$), while for students who did not use extended time, the weights are the inverse of 1 minus the propensity score ($\frac{1}{1-P_i}$). Finally, we used the inverse probability weights as weights to estimate Equation (1) employing a weighted GLM.

Under the assumption that there is no misspecification of the model used to estimate the weights, the IPW models yield an unbiased estimate of the average difference in test performance between students who used extended time/universal design features and those who did not. However, in the context of model misspecification where the above assumption does not hold, doubly robust models have been shown to be more robust (Bang & Robins, 2005; Robins et al., 2007).

Doubly robust models. These models combine both regression adjustment and propensity score methods. Individually, both regression adjustment and propensity score methods yield unbiased estimations only when their respective statistical models are accurately specified. However, the doubly robust (DR) estimator combines these two approaches, ensuring that unbiased estimation is achieved even if only one of the two models is correctly specified (Robins et al., 2000; Scharfstein et al., 1999). This method enhances the robustness of our analyses, enabling us to draw more reliable conclusions regarding the relationship between usage and students' test performance.

The doubly robust estimator was implemented using built in Stata command (e.g., teffects) which estimates it using the following formula (Emsley et al, 2008):

$$\hat{\tau}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{T_i\left(Y_i - \hat{\mu}_1\left(X_i\right)\right)}{\hat{p}_i}\right) + \hat{\mu}_1\left(X_i\right)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{T_i\left(Y_i - \hat{\mu}_0\left(X_i\right)\right)}{1-\hat{p}_i}\right) + \hat{\mu}_0\left(X_i\right)\right], \quad (3)$$

where $\hat{p}_i$ denotes the propensity scores computed from the propensity score model for each student as in Equation (2). The predicted outcome under treatment, $\hat{\mu}_1(X_i)$, is computed using a logistic regression model based on X_i for students who used extended time exclusively, similarly $\hat{\mu}_0(X_i)$, the predicted outcome under no treatment (i.e., students who did not use ET) is computed through a model estimated only on no treatment students.

Universal Design Usage and Performance among All Students

To answer our second research question, we estimated GLM, inverse probability weighting, and doubly robust models as described above for each UD feature (scratchwork, text-to-speech, and equation editor). These models were estimated separately for SWDs and SWOD. In all models ET_Use_{ij} variable was replaced with UD_Use_{ij}, which equals 1 if the student used that particular UD feature, and equals 0 otherwise.

Universal Design Usage and Performance among SWDs by Extended Time Use

To answer our third research question, the GLM, inverse probability weighting, and doubly robust models described above is replicated separately for the intersection of each UD feature (scratchwork, text-to-speech, and equation editor) with ET usage, resulting in six models for each combination of UD feature and ET usage. These models were estimated only for SWDs.

Results

Extended Time Usage and Performance among Students with Disabilities

Summary statistics. Table 1 presents the numbers and percentages of SWDs and SWODs identified as eligible for extended time.[6] Within SWDs, 62% are eligible for extended time, in contrast to just 1.7% of SWODs. Therefore, our examination of the relationship between extended time use and test performance will be confined to SWDs.

Table 2 divides SWDs into three groups: those ineligible for extended time, those eligible but not utilizing it, and those eligible and actively using it. This categorization reveals demographic variances, notably a higher percentage of female, Black, and Hispanic students within the eligible group, suggesting broader eligibility criteria might correlate with certain demographic factors. Conversely, White, Asian American, Pacific Islander, American Indian, or Alaska Native students are less likely to be eligible. Additionally, eligibility for extended time is more common among students participating in the school lunch program, indicating socio-economic factors might also influence eligibility. Given these demographic disparities, subsequent analyses will adjust for these characteristics to ensure a comprehensive evaluation. Interestingly, only 31.2% of SWDs (calculated as $\frac{530}{530+1140}$) who are eligible for extended time actually utilize this accommodation, underscoring the need to explore the factors influencing this decision.

Extended time usage and test performance. Table 3 examines the correlation between extended time use and test performance among SWDs. Column 1 shows the marginal effects from a logit estimate, accounting for student demographics, parental background, school, and test item characteristics. Columns 2 and 3 present the results using IPW and doubly robust models, respectively. All three models yield very similar results, suggesting a positive association between extended time usage and test performance among SWDs, with the use of extended time showing a 3.5 to 4.2 percentage point increase in the likelihood of correctly answering a test item. This finding highlights the significant influence that access to and utilization of extended time can have on enhancing the academic outcomes of SWDs.

Universal Design Usage and Performance among All Students

Summary statistics. Table 4 highlights the engagement with UD features—scratchwork, text-to-speech (TTS), or the equation editor—across students, showing that 66% of SWDs (calculated as $\frac{1260}{1260+660}$) and 56% of SWODs (calculated as $\frac{8810}{8810+7060}$) utilized at least one UD feature. Summary statistics for each UD feature can be found in the supplementary file.

The use of UD features among SWDs and SWODs reveals distinct demographic patterns and performance outcomes. SWDs users of UD features are predominantly male, with a significant representation among Black and Asian American or Pacific Islander students, whereas White and Hispanic SWDs are less likely to engage with these features. This trend is mirrored among SWODs, where UD usage similarly skews away from female, White, and Asian American or Pacific Islander students toward Black, Hispanic students, and those facing socio-economic challenges, such as English language learners and students eligible for free or reduced-price lunches. Across both groups, UD users tend to score lower on assessments compared to their peers who did not utilize these features, suggesting an intricate relationship among UD utilization, demographic factors, and academic performance.

Universal design usage and test performance. We now examine the relationships between UD usage and test performance among SWDs and SWODs. This analysis is conducted separately for scratchwork, text-to-speech, and equation editor usage.

Scratchwork usage. Table 5 presents the relationship between scratchwork usage and test performance among SWDs in Panel A and SWODs in Panel B. The logit results are presented in Column 1, followed by the results using IPW in Column 2 and doubly robust results in Column 3. In Panel A, the coefficients on scratchwork use show a positive association with IPW estimates, suggesting that the use of scratchwork is associated with 3.5 percentage points higher in answering an item correctly among SWDs. Similarly, the coefficients in Panel B indicate that scratchwork use is positively associated with performance under IPW estimates for SWODs.

Text-to-speech usage. Our analysis of the relationship between text-to-speech usage and test performance is presented in Table 6 for SWDs (Panel A) and SWODs (Panel B). In Panel A, across all three models (logit, IPW, and DR), the results consistently show no statistically significant association between text-to-speech use and test performance among SWDs. In Panel B, while the logit regression also suggests no correlation, the IPW and DR models reveal a negative correlation between text-to-speech use and test performance among SWODs. This difference arises because IPW and DR models place greater weight on students with similar characteristics in the control variables.

Equation editor usage. Table 7 presents the relationship between equation editor usage and test performance. The findings in Panel A indicate a negative relationship between the use of equation editors and test performance among SWDs. Notably, the magnitudes of the coefficients remain consistent across all three models. Similarly, in Panel B, the coefficients on equation editor use show a negative correlation with test performance among SWODs, with similar magnitudes observed across the three models. This suggests that the observed negative relationships between the use of equation editors and test performance among both SWDs and SWODs are robust and persist regardless of the statistical methodology employed.

Universal Design Usage and Performance among SWDs by Extended Time Use

To address our third research question, we examined whether the relationship between UD feature usage and test performance differed for SWDs based on their use of extended time accommodations. Panels C and D of Tables 5–7 present the subgroup analyses for SWDs who used extended time (ET users) and those who did not (non-ET users), respectively.

Scratchwork usage. As shown in Table 5, Panel C, among SWDs who used extended time, the use of scratchwork was positively associated with test performance in the IPW model (2.7 percentage point increase, $p < .1$), but this association was not statistically significant in the Logit or doubly robust models. In contrast, among SWDs who did not use extended time (Panel D), scratchwork use was associated with a statistically significant increase in test performance in the IPW model (3.5 percentage points, $p < .05$), though results were again not significant in the Logit and DR specifications. These findings suggest that the benefits of scratchwork may be more pronounced among students who did not use ET, potentially serving as a compensatory strategy for students navigating time constraints.

Text-to-speech usage. Table 6 reveals divergent patterns in the association between text-to-speech (TTS) usage and test performance across extended time usage groups. Among SWDs who used extended time (Panel C), no statistically significant association was found in any of the three models. However, among SWDs who did not use ET (Panel D), TTS usage was consistently associated with higher performance across all three models, with differences ranging from 2.4 to 2.7 percentage points. These findings indicate that TTS may be particularly beneficial for SWDs who do not receive additional time, potentially offsetting other barriers they face during the assessment. One possibility is that SWDs who did not use extended time differ in meaningful ways from those who did. They may vary in disability type, severity, or other unobserved characteristics, which could help explain the variation in the observed associations.

Equation editor usage. As shown in Table 7, the use of the equation editor was negatively associated with test performance across both groups. Among SWDs who used extended time (Panel C), all three models reported statistically significant negative coefficients, ranging from –.145 to –.244. Similarly, among SWDs who did not use ET (Panel D), the equation editor was also negatively associated with test performance, with coefficients ranging from –.190 to –.223. These consistent findings across models and subgroups suggest that the equation editor may function as a source of distraction or cognitive overload rather than as a support mechanism. The negative association between equation editor use and performance may reflect increased cognitive load, challenges with navigating the interface, or a mismatch between the tool's design and the demands of the test items. Unfortunately, these mechanisms cannot be definitively tested with the available data; future research is needed to further explore them.

Discussion and Conclusion

Our study provides a comprehensive analysis of the association between testing accommodations, the use of universal design features, and the performance of students with and without disabilities on standardized tests. NAEP process data present a significant opportunity to understand the dynamics of test-taking behavior in a digital environment. Unlike traditional test data, which focus solely on students' final responses, process data capture insights into how students interact with the assessment, allowing us to accurately measure students' use of testing accommodations and universal design features. Through advanced analytical techniques, including inverse probability weighting and doubly robust models, we provide rigorous evidence of the utility of specific accommodations and UD features. These methods strengthen the reliability of our findings and establish a foundation for future research and informed policy development in educational assessment.

Our findings offer valuable insights into the relationship between accommodations, universal design usage, and students' performance. However, they should be interpreted with caution due to two limitations. First, our analysis relied on NAEP's definition of students with disabilities, which encompasses both students with Individualized Education Plans (IEPs) and those with Section 504 Plans. We recognize that these two groups differ significantly in the level of support they receive, as discussed in the limitations section. Second, while we examined the interaction between extended time and UD features such as text-to-speech and scratchwork, small subgroup sizes constrain the statistical power. Additionally, although our methods mitigated selection bias by employing inverse probability weighting and doubly robust models, residual endogeneity cannot be entirely ruled out.

Despite these limitations, our findings underscore the critical role of extended time in promoting more equitable testing conditions for SWDs. The positive correlation between extended time usage and improved test performance highlights the importance of maintaining this accommodation as a core support. Our subgroup analyses further reveal that the effectiveness of universal design features may depend on whether students also utilize extended time. Supports such as scratchwork and text-to-speech appear to benefit SWDs without extended time more consistently, suggesting that these features may play a compensatory role when time accommodations are not in use. Conversely, the equation editor shows a negative association with performance across all subgroups, raising questions about its utility and design.

These findings build upon two decades of work advocating for universally designed assessments. Rather than being a product of digital testing alone, UD has long emphasized inclusive design principles applicable to both analog and digital assessments (e.g., Thompson et al., 2002). Our results reinforce the need for an individualized and context-aware approach to accessibility supports. UD features should not be assumed benefit to all students equally or to be effective across every accommodation contexts.

At the same time, our investigation revealed the complexity of these UD features in practice. While scratchwork appears to support performance among SWDs, text-to-speech showed no benefit for SWDs overall and a negative relationship with performance for SWODs. The equation editor, in particular, was negatively associated with performance regardless of disability status or accommodation use.

Our findings for TTS align well with established assistive-technology research and cognitive load theory. Cognitive load theory suggests that splitting information across visual and auditory channels can free up working memory (Keelor et al., 2020). In other words, bypassing the decoding process via an audio channel can reduce cognitive strain. Consistent with this, prior studies report that TTS "is a useful compensatory reading aid." For example, one experiment found that reading comprehension was significantly higher when students used TTS rather than reading silently (Keelor et al., 2020). Similarly, a meta-analysis concluded that read-aloud tools (including synthesized TTS) have a clear pos-

itive effect on comprehension for students with reading disabilities (Wood et al., 2018). These results make sense under cognitive load theory: if struggling readers are relieved of decoding text themselves, "more cognitive resources can be allocated to comprehending the text" (Keelor et al., 2020; Wood et al., 2018).

By contrast, the negative association we found for equation editor use echoes cautionary insights from cognitive load theory. If a support tool adds extraneous complexity, it can actually overload working memory (Keelor et al., 2020). Writing math with an on-screen equation editor may require complex formatting steps or unfamiliar navigation that distract from problem solving. This suggests that rather than simplifying tasks, a poorly matched digital tool can function as a cognitive burden. In line with our results, assistive-technology theory predicts that any gadget misaligned with a student's processing needs will introduce extraneous load and impair outcomes. The consistent negative coefficients for the equation editor in both subgroups imply it acted more as a distraction than a help. Together, these findings underline that UD features must be well designed and tailored to cognitive demands: a feature intended to aid (like an equation editor) might only be beneficial when its interface and use match the learner's needs, otherwise it adds to cognitive load and undermines performance.

Taken together, it is clear that a one-size-fits-all approach to accommodations and UD implementation is insufficient. The varied correlations between the usage of different accommodations and UD features and test performance underscore the need for a better understanding of how these interventions affect diverse groups of students. Ongoing research is needed to examine the interaction between student characteristics, specific supports, and item-level demands. Moreover, continuous refinement of digital testing platforms and accommodation policies is critical to ensure that they truly serve the diverse needs of today's learners.

For assessment designers and educators, these results suggest the value of refining and carefully implementing digital testing features. Digital assessment platforms should be designed to minimize extraneous cognitive load and provide clear, user-friendly guidance on tools such as equation editors and text-to-speech, ensuring that their intended benefits are fully realized. Incorporating scaffolding, such as embedded tutorials or guided practice opportunities, may help students build familiarity with these tools prior to high-stakes testing. Additionally, training and instructional resources for educators are essential to support effective implementation, as educators play a key role in preparing students to use digital UD features with confidence.

Future studies should investigate the long-term use of accommodations and UD features across a broader range of subjects and assessment types. There is also a need to develop and test new UD features that can enhance test accessibility and fairness without introducing unintended distractions or barriers. As policymakers and practitioners seek to create more inclusive assessment systems, this study provides timely evidence to inform those efforts. It underscores the importance of iterative research, design, and policy refinement in achieving truly equitable assessment practices—practices that accommodate diversity in learning needs and advance educational opportunities for all students.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324P210002 to the American Institutes for Research (AIR). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Notes

[1]To distinguish accommodations from modifications, it is worth noting that accommodations maintain the original intent and rigor of the assessment, while modifications may alter the content or expectations. Universal design features, on the other hand, aim to reduce barriers for all test-takers by integrating accessible design elements from the outset.

[2]Examples of the questions include: Which of the following best describes you? Select one or more answer choices: "White," "Black or African American," "Asian," "American Indian or Alaska Native," "Native Hawaiian or other Pacific Islander."In this school year, how often have you used a computer or other digital device (excluding handheld calculators) to take an online practice test? "Never," "Once," "Two or three times," "Four or five times," "More than five times."

[3]More information on the matrix sampling used by NCES can be found here https://nces.ed.gov/training/datauser/NAEP_04/assets/NAEP_04_Slides.pdf and here https://nces.ed.gov/statprog/handbook/naep_surveydesign.asp

[4]Students completed a single form consisting of two 30-minute blocks. Each block includes 15 test items.

[5]While we recognize that including UD usage in the ET propensity model (and vice versa) could help account for joint influences, we chose not to do so in order to maintain interpretability and avoid adjusting for variables that may be downstream or partially determined during testing. Instead, our models condition on a broad set of pre-treatment covariates—student, parent, item, and school characteristics—along with disability status in the UD models. We explore heterogeneity in UD effects by ET usage in subgroup analyses (Tables 5–7, Panels C and D), and suggest more integrated modeling as a direction for future research.

[6]Only .6% of SWDs are eligible for calculator usage, while none of the SWODs are eligible for this accommodation. Due to the limited number of students eligible for calculator usage, our analyses concentrate solely on the extended time accommodation.

References

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

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, *34*(28), pp. 3661–3679. https://doi.org/10.1002/sim.6607

Balasa, D., Case, H., Gonthiere, I., Greenberg, M., Incrocci, M., Julian, E., …Suhr, J. (2019). Testing accommodations: The perils of the "approve everything" model. *Journal of the National College Testing Association*, *3*(2).

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), pp. 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bouck, E., & Bone, E. (2018). Interventions for students with intellectual disabilities. In F. E. Obiakor & J. P. Bakken (Eds.). *Viewpoints on interventions for learners with disabilities* (1st ed., pp. 55–73). Bingley, West Yorkshire: Emerald Group Publishing Limited. https://doi.org/10.1108/S0270-401320180000033004

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score

models. *American Journal of Epidemiology*, *163*(12), pp. 1149–1156. https://doi.org/10.1093/aje/kwj149

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

Calhoon, M. B., Fuchs, L. S., & Hamlett, C. L. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly*, *23*(4), pp. 271–282. https://doi.org/10.2307/1511349

Christensen, L. L., Braam, M., Scullin, S., & Thurlow, M. L. (rep.). (2011). *2009 State policies on assessment participation and accommodations for students with disabilities*. Minneapolis, Minnesota: National Center on Educational Outcomes.

Clapper, A. T., Morse, A. B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. L. (rep.). (2005). 2003 State policies on assessment participation and accommodations for students with disabilities (Synthesis Report 56). *Minneapolis*, Minnesota: National Center on Educational Outcomes.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), pp. 418–446. https://doi.org/10.1002/pam.20586

Elliott, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children*, *70*(3), pp. 349–367. https://doi.org/10.1177/001440290407000306

Emsley, R., Lunt, M., Pickles, A., & Dunn, G. (2008). Implementing double-robust estimators of causal effects. *The Stata Journal*, *8*(3), pp. 334–353. https://doi.org/10.1177/1536867X0800800302

Fuchs, L. S., Fuchs, D., & Capizzi, A. M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children*, *37*(6), pp. 1–8. https://doi.org/10.17161/foec.v37i6.6812

Hicks, J., Circi, R., & Li, M. (2019). Students' use of support functions in DBAs: Analysis of NAEP grade 8 mathematics process data. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.). *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, pp. 568–571.

Huynh, H., Meyer, J. P., & Gallant, D. J. (2004). Comparability of student performance between regular and oral administrations for a high-stakes mathematics test. *Applied Measurement in Education*, *17*(1), pp. 39–57. https://doi.org/10.1207/s15324818ame1701_3

Keelor, J. L., Creaghead, N., Silbert, N., & Horowitz-Kraus, T. (2020). Text-to-speech technology: Enhancing reading comprehension for students with reading difficulty. *Assistive Technology Outcomes & Benefits*, *14*(1), pp. 19–35.

Kikis-Papadakis, K., & Kollias, A. (2009). Reflections on paper-and-pencil tests to eAssessments: Narrow and broadband paths to 21st century challenges. In F. Scheuermann & J. Bjórnsson (Eds.). *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing*. Luxembourg: Office for Official Publications of the European Communities.

Kim, Y. Y., & Circi, R. (2018). *The extended time accommodation (ETA) and performance of students with ETA*. New York, USA: National Council on Measurement in Education.

Lang, S. C., Elliott, S. N., Bolt, D. M., & Kratochwill, T. R. (2008). The effects of testing accommodations on students' performances and reactions to testing. *School Psychology Quarterly*, *23*(1), pp. 107–124. https://doi.org/10.1037/1045-3830.23.1.107

Lee, S. Y., Hicks, J., & Circi, L. (2020). *Insights on text-to-speech as a universal design feature: NAEP mathematics process data*. NCME Virtual Conference.

National Center for Education Statistics. (2024). *Students with disabilities. Condition of Education*. U.S. Department of Education, Institute of Education Sciences.

Michaelides, M. P., & Ivanova, M. (2022). Detecting rapid-guessing behavior using response time and accuracy in PISA items. *Psychological Test and Assessment Modeling*, *64*(3), pp. 304–338.

Mitchell, R., & Qi, S. (2012). Large-scale academic achievement testing of deaf and hard-of-hearing students: Past, present, and future. *The Journal of Deaf Studies and Deaf Education*, *17*(1), pp. 1–18. https://doi.org/10.1093/deafed/enr028

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research*, *71*(1), pp. 53–104. https://doi.org/10.3102/00346543071001053

Robins, J. M., Rotnitzky, A., & van der Laan, M. (2000). On profile likelihood: Comment. *Journal of the American Statistical Association*, *95*(450), pp. 477–482. https://doi.org/10.2307/2669391

Robins, J., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, *22*(4). https://doi.org/10.1214/07-sts227d

Römhild, A., & Hollederer, A. (2024). Effects of disability-related services, accommodations, and integration on academic success of students with disabilities in higher education. A scoping review. *European Journal of Special Needs Education*, *39*(1), pp. 143–166.

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)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics, U.S. Government Printing Office.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), pp. 1096–1120. https://doi.org/10.2307/2669923

Sireci, S. G., Banda, E., & Wells, C. S. (2018). Promoting valid assessment of students with disabilities and English learners. In Elliott, S., Kettler, R., Beddow, P., & Kurz, A. (Eds), *Handbook of accessible instruction and testing practices* (pp. 231–246). https://doi.org/10.1007/978-3-319-71126-3_15

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), pp. 457–490. https://doi.org/10.3102/00346543075004457

Suk, Y., Steiner, P. M., Kim, J.-S., & Kang, H. (2022). Regression discontinuity designs with an ordinal running variable: Evaluating the effects of extended time accommodations for english-language learners. *Journal of Educational and Behavioral Statistics*, *47*(4), pp. 459–484. https://doi.org/10.3102/10769986221090275

Suk, Y., & Kim, Y. (2024). Fuzzy regression discontinuity designs with multiple control groups under one-sided noncompliance: Evaluating extended time accommodations. *Journal of Educational and Behavioral Statistics*, *0*(0). https://doi.org/10.3102/10769986241268902

Tapper, J., Morris, D., & Setrakian, L. (March 30). (2006) *Does loophole give rich kids more time on SAT?* ABC NEWS.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from https://nceo.umn.edu/docs/OnlinePubs/Synth44.pdf

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (rep.). (1993). *Testing accommodations for students with disabilities: A review of the literature* (pp. 1–67). Minneapolis, Minnesota: National Center on Educational Outcomes.

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (2003). *Testing students with disabilities practical strategies for complying with district and state requirements* (2nd ed.). Thousand Oaks, California: Corwin Press.

Wei, X., & Zhang, S. (2024). Extended time accommodation and the academic, behavioral, and psychological outcomes of students with learning disabilities. *Journal of Learning Disabilities*, *57*(4), pp. 242–254. https://doi.org/10.1177/00222194231195624

Wei, X. (2024). Text-to-speech technology and math performance: A comparative study of students with disabilities, English Language Learners, and Their General Education Peers. *Educational Researcher*, *53*(5), pp. 285–295. https://doi.org/10.3102/0013189X241232995

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). *An investigation of the relationship between time of testing and test-taking effort*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.
Weston, T. J. (rep.). (2002). *The validity of oral accommodation in testing*. Palo Alto, California: NAEP Validity Studies (NVS) American Institutes for Research.
Witmer, S. E., Lovett, B. J., & Buzick, H. M. (2023). Extended time accommodations on the 2017 NAEP grade 8 mathematics test: Eligibility, use, and benefit. *Journal of Psychoeducational Assessment*, *41*(2), pp. 123–135. https://doi.org/10.1177/07342829221130457
Wood, S. G., Moxley, J. H., Tighe, E. L., & Wagner, R. K. (2018). Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *Journal of learning disabilities*, *51*(1), pp. 73–84.
Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No child left behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness*, *8*(2), pp. 245–279. https://doi.org/10.1080/19345747.2013.878011

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Supporting information