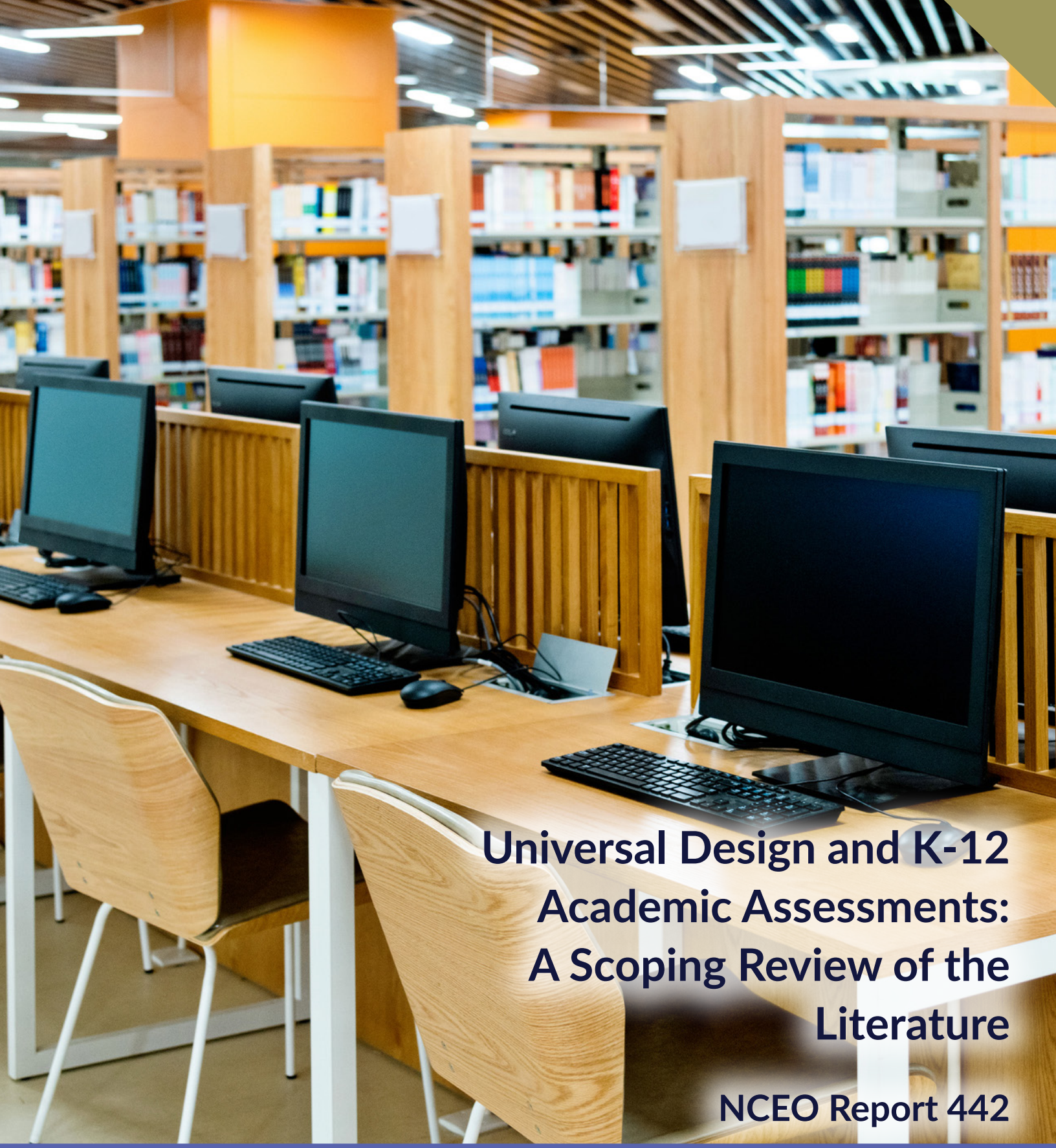


Universal Design and K-12 Academic Assessments: A Scoping Review of the Literature

NCEO Report 442

NCEO
National Center on Educational Outcomes

NCEO Report 442

Universal Design and K-12 Academic Assessments: A Scoping Review of the Literature

Kristin K. Liu, Martha L. Thurlow, Mari Quanbeck, Jessica A. Bowman, and Amy Riegelman

February 2024

The Center is supported through a Cooperative Agreement (#H326G210002) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. Consistent with EDGAR §75.62, the contents of this report were developed under the Cooperative Agreement from the U.S. Department of Education, but do not necessarily represent the policy or opinions of the U.S. Department of Education or Offices within it. Readers should not assume endorsement by the federal government.
Project Officer: David Egnor

In collaboration with:

NCEO Core Staff

Sheryl S. Lazarus, Director
Jessica Bowman
Kascinda Fleming
Gail Ghere
Linda Goldstone
Andrew R. Hinkle
Kristi K. Liu
Charity Funfe Tatah Mentan
Michael L. Moore
Darrell Peterson
Mari Quanbeck
Virginia A. Ressa
Christopher M. Rogers
Jennifer Sommerness
Kathy Strunk
Martha L. Thurlow
Yi-Chen Wu

National Center on Educational Outcomes
University of Minnesota
2025 East River Parkway, Room 1-330 • Minneapolis, MN 55414
Phone 612/626-1530
http://www.nceo.info

Executive Summary

State education agencies, K-12 assessment vendors, teacher trainers, and classroom teachers are increasingly applying the principles of universal design (UD)—including Universal Design for Learning (UDL), Universal Design of Instruction (UDI), Universal Instructional Design (UID), and Universal Design of Assessment (UDA)—to the instruction and assessment of students from special populations. In theory, UD ensures that instruction and assessment are created from the beginning to be accessible to the widest population of students possible.

This report summarizes an investigation of literature published in 1985-2023 on universal design (UD) of large-scale assessments. We included literature on the application of UD to the broad range of K-12 district and state large-scale academic assessments in the U.S. Using a scoping review process, we searched peer-reviewed manuscripts as well as Ph.D. dissertations, white papers, briefs, reports and information published on websites of non-governmental organizations (i.e., gray literature). After a four-stage resource identification and screening process, we included 76 resources in our literature review; 28 of the included resources were research-based.

We found that the resources included in this scoping review, particularly the research-based resources, did not apply a consistent UD framework or definition. In the 76 resources, UDA was represented more often than UDL; sometimes multiple frameworks were referred to, or no framework was mentioned. In the subset of 28 research-based resources, more studies referenced UDA than UDL. Again, these resources often referred to multiple frameworks, without clearly articulating how frameworks with different components related to each other. In some cases, published resources simply referred to UD without articulating any underlying framework.

We found that the UD resources primarily addressed students with disabilities, but a growing number of them examined UD for English learners; a few addressed English learners with disabilities. When students with disabilities were referenced, more resources described the general population of students with disabilities, followed by resources that mentioned students with learning disabilities. Relatively few studies addressed students with significant cognitive disabilities who might be taking alternate assessments. We also found that the term UD in reference to assessments seemed to have shifted to be equated primarily with providing accessibility features and accommodations to students with disabilities, English learners, and English learners with disabilities.

Of the 28 research-based resources, only three (Housh et al., 2020; Johnstone et al., 2006; McMahon et al., 2016) tied the study findings back to concepts associated with an underlying UD framework. Aside from these three studies, the assessment research reviewed was not linked to a set of elements associated with a specific UD framework. Thus, UD applied to K-12 large-scale academic assessments has not yet been consistently implemented and therefore does not yet have a sufficient research base.

Our scoping literature review of the UD literature from 1985 to 2023 revealed a number of gaps, both overall and in the research literature. Perhaps the most glaring gap was the lack of specificity about which UD framework was the focus of the resources. Without this focus, especially in the research resources, it is unclear what is being studied and thus, what the results of the studies mean.

We identified several implications of the literature for educators, state education agencies and test developers. For researchers, it is essential that a clear description of the UD framework that guides the work is provided. Further, the specific elements of the chosen framework should be delineated. Researchers should avoid equating UD with simply the development or provision of accommodations or other accessibility supports. For state education agencies and test developers, the UD framework that is to be implemented in assessments should be defined in requests for proposals for assessments, in considerations of test vendors' proposals, and in evaluating assessments that are developed.

Continued research on the impact of UD on large-scale assessment accessibility will benefit the field and should result in higher-quality assessments for all students. With such research, the evidence provided could more powerfully inform test creation, test administration, and analyses of educational achievement for all.

Table of Contents

Including all students in state and districtwide K-12 academic assessments is required by federal education laws. Both the Elementary and Secondary Education Act (ESEA) and the Individuals with Disabilities Education Act (IDEA) emphasize that the assessments be accessible for all students, not only by providing accommodations and other accessibility supports, but also by being designed from the beginning for all students.

State education agencies, K-12 assessment vendors, teacher trainers, and classroom teachers are increasingly applying the principles of universal design (UD)—including Universal Design for Learning (UDL), Universal Design of Instruction (UDI), Universal Instructional Design (UID), and Universal Design of Assessment (UDA)—to the instruction and assessment of students, including those from special populations. In theory, UD ensures that instruction and assessment are created from the beginning to be accessible to the widest population of students possible. Given these multiple conceptual frameworks, however, the specific elements of UD can be hard to define.

In a systematic review of UD frameworks applied to K-12 classroom instruction research, Ok et al. (2017) found that researchers applied multiple UD frameworks for a range of varied purposes. Ok et al.'s work focused specifically on UDL (CAST, 2018) in instruction. State education agencies and K-12 assessment vendors are now applying UD principles to K-12 large-scale academic assessments (Lazarus et al., 2021). Yet, it is unclear how these UD frameworks and their associated principles and elements are being applied, or even which UD framework is used. This obscurity and lack of defined UD principles in the research literature creates obstacles in research replication and in identification of evidence-based practices. State education agency staff, test vendors, school leaders, and policymakers need to know more as they consider how to incorporate "best practices" in UD into assessment programs. A scoping review of the literature on UD applied to assessment can help address this need.

This review identifies how the concept of UD has been applied to the broad range of U.S. district and state large-scale academic assessments. We had two guiding research questions:

1. How has the concept of UD been described and applied to K-12 large-scale academic assessments?

2. To what extent do research-based publications address the application of UD of large-scale assessments to students with disabilities and diverse students?

Perspectives or Theoretical Frameworks

The term UD was coined by Ron Mace, an architect with a disability (see udinstitute.org). He described how improved building design could address and remove access barriers for people with disabilities to the maximum extent possible. The concept spread to education

where barriers to learning exist for students with disabilities and other students such as English learners. UD first appeared in the Individuals with Disabilities Education Act (IDEA) of 2004. Additional references to UD followed in the Federal Register announcement of the Race-to-the Top Assessment Program of 2009, and to UDL in the Every Student Succeeds Act of 2015. Although UD is most recently often identified in federal legislation using the term "UDL," there are four separate UD conceptual frameworks that have been applied in education: (a) Universal Design for Learning (UDL; CAST, 2018); (b) Universal Instructional Design (UID; Higbee & Goff, 2008); (c) Universal Design of Instruction (UDI; Burghstahler, 2009), and (d) Universal Design of Assessment (UDA; Thompson et al., 2002). Each framework has different components addressing areas such as flexible learning environments, multiple means for student engagement with and responses to educational materials, supports for learning and assessment, and inclusive instructional design. Table 1 summarizes the main elements of the two frameworks most often applied to large-scale assessments (UDA and UDL). Detailed descriptions of these UD frameworks are presented in Appendix A.

Table 1. Elements of Two Primary Universal Design Frameworks—Universal Design of Assessment (UDA) and Universal Design for Learning (UDL)

Universal Design of Assessment or UDA (Thompson, Johnstone, & Thurlow, 2002)	**Universal Design for Learning or UDL** (CAST, 2018)
Elements	***Elements***
Inclusive Assessment Population: Assessment provides opportunities for all students to participate no matter their cognitive abilities, cultural backgrounds, or linguistic backgrounds. **Precisely Designed Constructs:** Assessment constructs are clearly defined and generally accepted. **Accessible, Non-Biased Items:** Items are reviewed for issues such as content quality, clarity, ambiguity, gender sensitivity, and cultural issues. **Amenable to Accommodations:** The assessment facilitates the use of appropriate accommodations and reduces threats to validity and comparability of scores. **Simple, Clear, and Intuitive Instructions and Procedures:** Instructions and procedures are provided in simple, clear, consistent, and understandable language so test takers can respond to tasks as intended. **Maximum Readability and Comprehensibility:** Assessments minimize the linguistic complexity of items while preserving the skills and concepts they are intended to measure. **Maximum Legibility:** Text, tables, figures, illustrations, and response formats can be deciphered easily.	**Provide Multiple Means of Action and Engagement:** Provide options for recruiting interest, sustaining effort and persistence, and self-regulation. **Provide Multiple Means of Representation:** Provide options for perception, language and symbols, and comprehension. **Provide Multiple Means of Action and Expression:** Provide options for physical action, expression and communication, and executive functions.

Most often, UDI and UID are applied to higher education, while UDL is applied to K-12 instruction, and UDA is applied to K-12 assessments. However, there may not be consistency in the literature in this respect. UDA can be applied to assessments given at any level of the education system. Similarly, UDL can be applied to classroom and other assessments.

Methods and Data Sources

To allow for the incorporation of emergent thinking on the topic of universal design and assessments, both peer-reviewed literature and gray literature searches were conducted for resources published between 1985 and 2023. Three sets of detailed inclusion/exclusion criteria, along with the rationale for each criterion, were created to address the different types of literature included in the study (see Appendix B for tables that provide the full lists of inclusion and exclusion criteria).

Many criteria were the same across types of literature, such as publication years (1985-2023), context addressed (U.S. education context), and referral to "universal design" or a related framework (UDL, UDA, UDI, UID). Other criteria were specific to each type of literature, such as being published in a peer-reviewed education or testing journal, being a Ph.D. or Ed.D. dissertation (gray literature), and being a document, report, or web article with an author and publication date (other gray literature).

The research team created keyword lists and tested searches while finalizing the inclusion and exclusion criteria. The basic forms of the keywords are shown in Table 2. The keywords were tested to refine our approach and were adapted slightly for web searches using a Google site search process.

Table 2. Search Terms

Primary Search Terms	Secondary Search Terms	Tertiary Search Terms
Universal Design	Assess*	School*
Universal Design for Learning	Educational Tests and Measurements	Education
Universal Instructional Design	Examinations	Elementary Education
Universal Design of Instruction	Summative Tests	Middle School Education
Universal Design of Assessment		Secondary Education
UD		
UDL		
UID		
UDI		
UDA		

Note. An asterisk in a Boolean search string indicates to include words that start with the same letters but may have different word endings.

The next step in the process involved drafting a scoping review protocol in adherence with conducting standards and awareness of the PRISMA-ScR extension for scoping review reporting guidelines (Tricco et al., 2018). We registered the protocol in the Open Science Framework Registries. It describes preliminary searches used to plan the research project, search terms, resource inclusion and exclusion criteria and rationale, search strategy, study selection procedures, and initial plans for data analysis and presentation. Throughout the writing stage, the fifth author—a university Social Sciences and Evidence Synthesis librarian who is a member of the Campbell Collaboration (an international science research network)—reviewed drafts and provided feedback to ensure the team followed best practices in conducting scoping reviews.

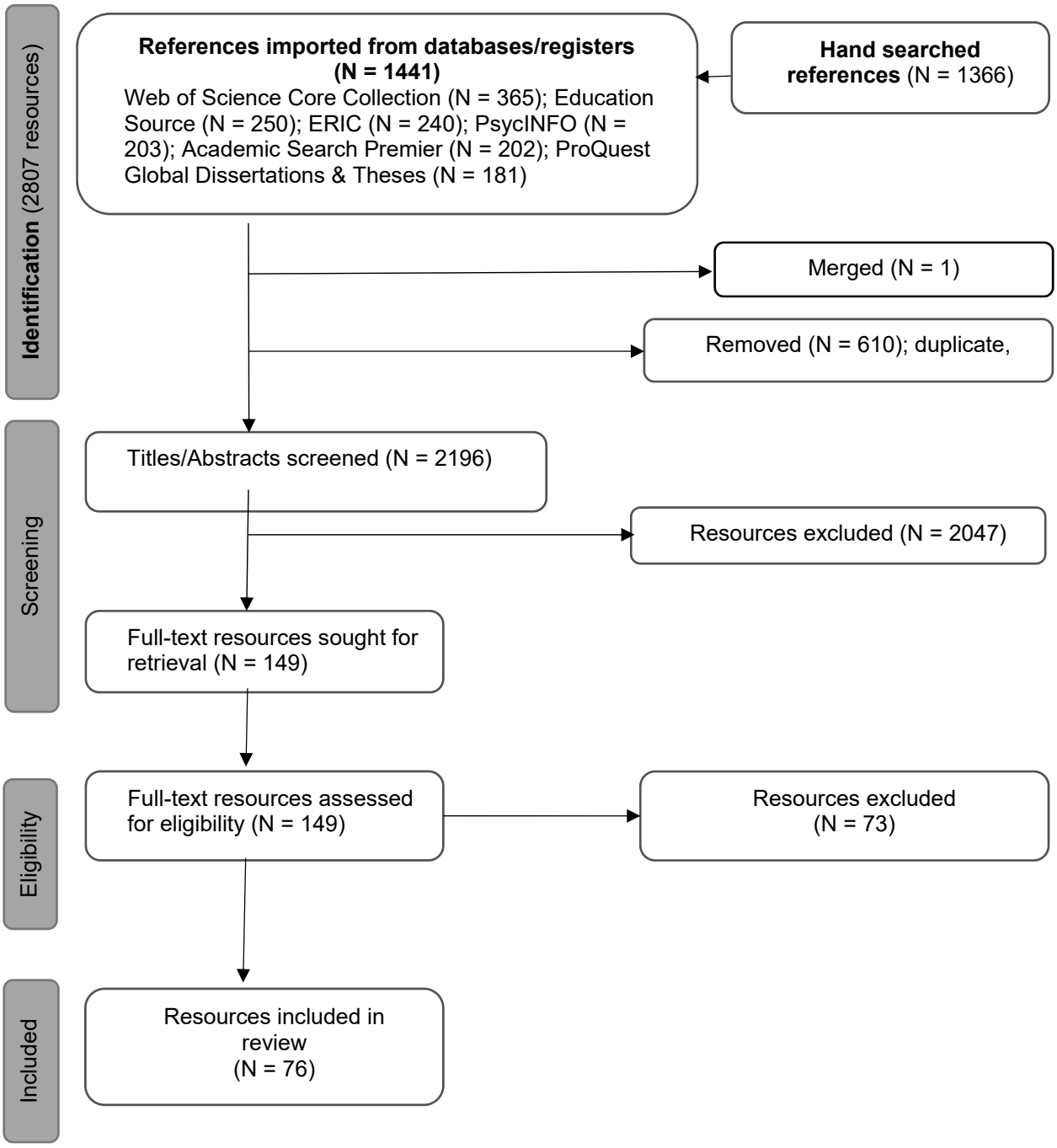
Resource Identification and Screening

Figure 1 shows the four-stage resource identification and screening process. The Figure provides the number of resources included and excluded at each stage. First, the research team searched University-affiliated library databases (where we found 1,441 potential resources) and conducted a hand search of websites (finding 1,366 additional potential resources). In total, we identified 2,807 resources to screen. Database searches conducted by the author who is a librarian, included resources found in ERIC (Ebsco), Education Source, APA PsycInfo via OVID, ProQuest Global Dissertations and Theses, Academic Search Premier, Web of Science Core Collection, and EdArXiv. These database searches used Boolean logic with primary, secondary, and tertiary search terms (see Table 2 for a list). Search terms included words relating to universal design (e.g., *universal design, universal design for learning, UD, UDL*), assessment (e.g., *assess*, educational tests and measurements*) and education (e.g., *school*, education, elementary education*). All database searches were conducted on April 10, 2023. The full, reproducible electronic search strategy is available in Appendix C. Records were exported out of each database and imported into Covidence, an evidence synthesis web application, and duplicate records were removed

The other members of the team conducted Google site searches of targeted assessment vendor and non-governmental organization websites. Websites to search were identified with the assistance of a subject matter expert and included websites of K-12 large-scale educational assessment companies (ACT, Cambium, Cognia, College Board, Curriculum Associates, DRC, ETS, Meridian, NWEA, Pearson VUE, Renaissance) and non-governmental organizations that have published on universal design (CAST, National Center on Educational Outcomes, Project Do-IT). The Google site search process combined specific search terms with website addresses (e.g., *universal design site: www.act.org*). All searches were conducted between February and April 2023.

Resources found in university library databases were imported directly into Covidence, which is software designed for conducting rigorous systematic literature reviews. Resources found through

Figure 1. Resource Identification and Screening Process



website searches were first documented in G Suite (e.g., Google Sheets and Google Docs). Due to the extremely large number of potential resources identified through Google site searches, the unpredictable accuracy of those searches, and a high number of duplicate resources identified, the research team conducted an initial screening of website literature outside of Covidence. The team imported a smaller set of website resources into Zotero, a reference citation manager, and

then into Covidence. In the process of conducting these two types of searches, we found 610 duplicate or incomplete resource records that we deleted. Multiple records for one additional resource were merged to retain the most complete information about that resource. Overall, the team identified 2,196 resources for title and abstract screening.

Second, in the screening stage, four researchers used Covidence to screen titles and abstracts of these 2,196 resources using the criteria found in Appendix B. Either the first or second author screened each of the resources. Covidence randomly assigned a second researcher to review each one. In the case of differing opinions, a third team member was randomly chosen to settle them. At this stage, we excluded 2,047 resources because they did not meet the title and abstract screening criteria. We selected 149 resources to move on to full text screening.

In the third step, we located the full text versions of those 149 resources chosen during the screening stage to ensure that the entire resource met the inclusion criteria described in Appendix B. Either the first or second author completed a full text review of each resource and Covidence randomly assigned a second researcher. In the case of a disagreement, Covidence assigned a third, randomly chosen, researcher to settle it. The team excluded 73 resources of the 149 resources. The remaining 76 resources are included in this review. For a complete list of these 76 resources, see the References.

Data Extraction

Research team members conducted a content review and analysis of the 76 resources. Information from the resources was extracted into a Google Form. Extracted information included the year of publication, type of resource, aims and purpose, universal design conceptual framework, type of large-scale assessment, and student groups addressed. Additionally, researchers considered whether each resource presented a history of universal design, cited additional research on universal design, advocated for universal design, cited policy related to universal design, and explicitly described the use of universal design in assessments. Two researchers identified data to extract from each resource, discussed their results, and settled any disagreements. We compiled Google Forms data into a Google spreadsheet and used the spreadsheet to create figures and tables for the overall results sections of this report. The team then conducted further analysis of the 28 studies identified as research resources. (These studies are marked with an asterisk in the Reference list.) Appendix D tables contain a summary of these data. Data from the tables are summarized in figures in the research resource sections of this report. Some of the research-based articles report original systematic research and some represent scholarly thinking that is typically published in a peer-reviewed journal.

Results

This section of the report is divided into two sections. First, it presents general characteristics of the total set of 76 resources. See the References for a full list of included resources. Next it presents findings by research question. Findings are first presented for the total set of 76 resources and then are followed by findings specifically for the subset of resources that contained research findings (N=28). These research resources are marked in the Reference list with an asterisk.

General Characteristics of Resources

Figure 2 shows that the years with the largest number of resources explicitly referencing UD occurred between 2005 and 2009. The 32 resources published during these five years represent 42% of all the resources included in this review.

Figure 2. Publication Year of Resources Reviewed (Number)

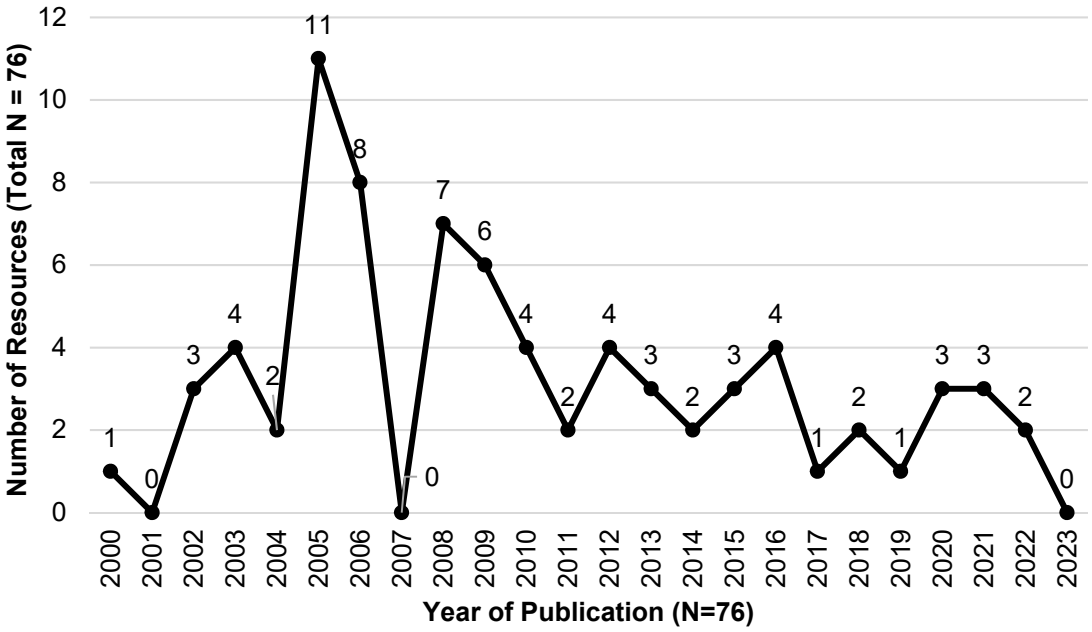
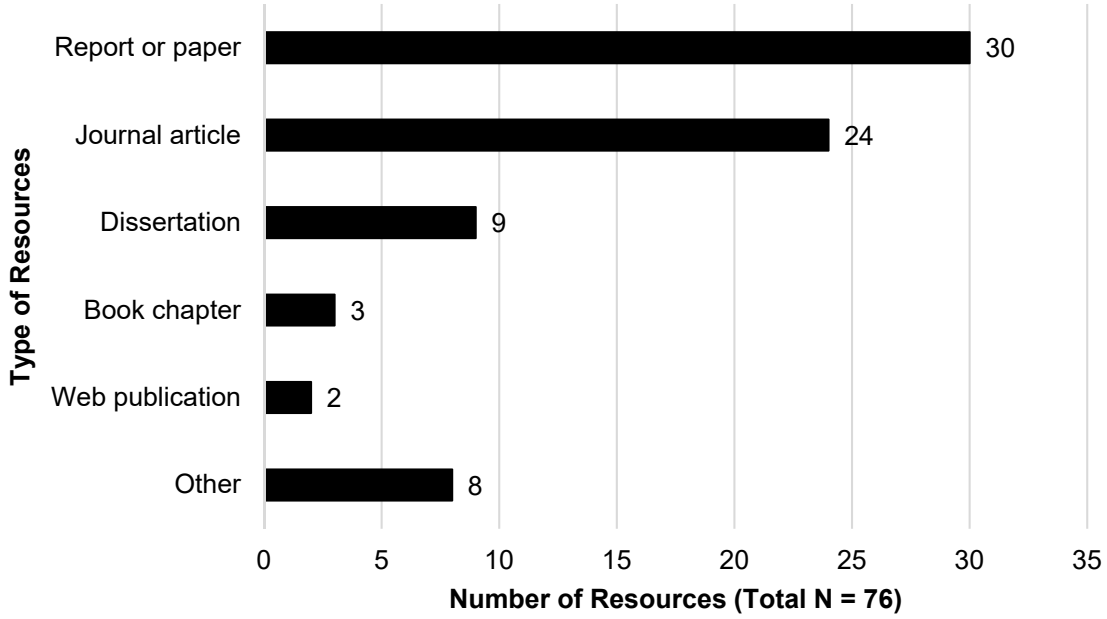


Figure 3 presents the types of resources reviewed. Reports or papers from non-governmental organizations made up the largest category of resources (N=30), followed by journal articles (N=24), dissertations (N=9), book chapters (N=3), and web-based publications (N=2). "Other" (e.g., newspaper articles, blogs) accounted for eight of the resources. Most resources were gray literature rather than peer-reviewed literature.

Figure 3. Types of Resources Reviewed (Number)

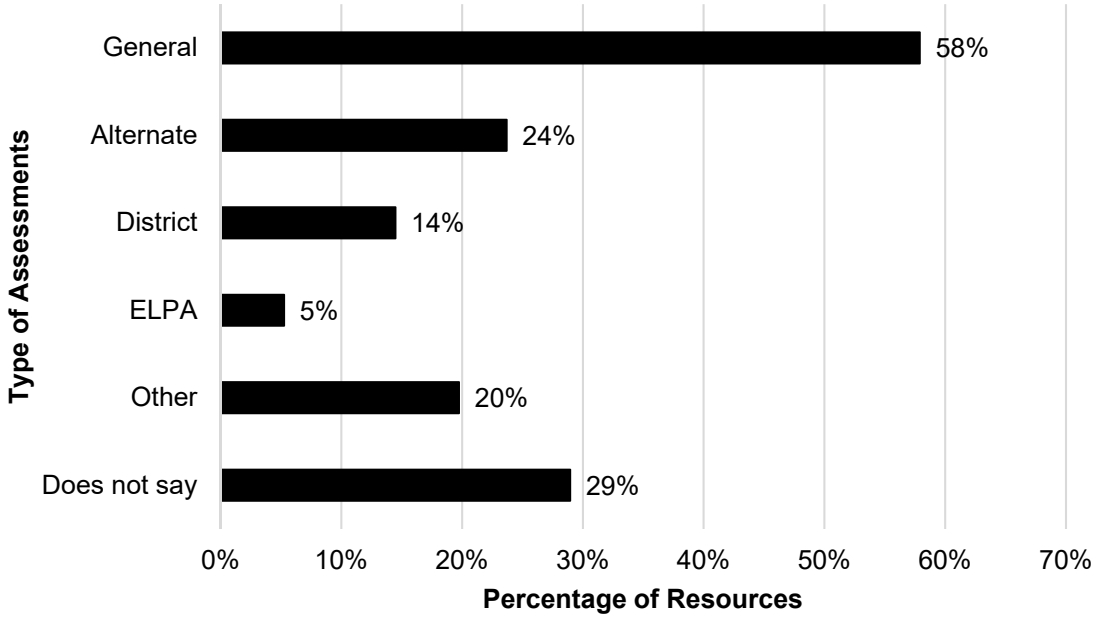


Note. Each resource is represented in only one category.

UD resources were created for a wide variety of purposes. One common purpose was to support the development and implementation of accessible assessments, particularly those given on computer. Another was to examine the effect of testing accommodations either on students' test performance or on the validity of the assessment. More than half of the included resources (63%; N=48) did not describe systematic research. Instead, the majority of resources described the concept and benefits of UD. Only 37% (N=28) of the resources contained original research.

Figure 4 identifies the types of large-scale academic assessments mentioned most frequently in resources. More than half of resources addressed a state general assessment, 24% addressed an alternate assessment, 14% addressed a district-wide assessment, and 5% addressed a state English language proficiency assessment (ELPA). Twenty percent addressed another type of large-scale assessment (e.g., a college entrance assessment, the National Assessment of Educational Progress, a researcher-developed assessment), and 29% did not identify the type of assessment. Fifty resources (66%) specified an assessment content area at some point in the writing, while 26 (34%) never specified the content area at any point. The majority of the resources addressed mathematics, followed by English language arts/Reading/Writing, Science, or "other" (e.g., social studies).

Figure 4. Types of Assessments Mentioned in Resources Reviewed (Percentage)



Notes. Each resource could include more than one type of assessment. Nearly two-thirds of resources fell into just one category. The remaining resources fell into two or more of the categories. ELPA = English Language Proficiency Assessment

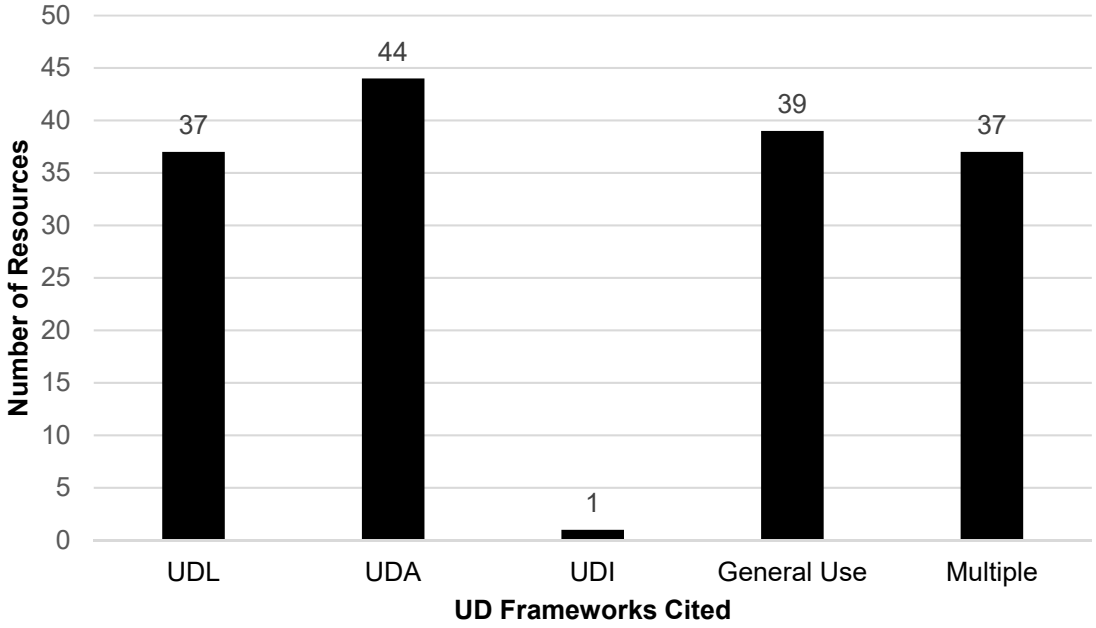
Research Question Results: Question 1

How has the concept of UD been described and applied to K-12 large-scale academic assessments?

Overall Resources

Our search looked for resources on any of the four major UD frameworks: UDL, UDA, UDI, and UID. Figure 5 shows that more than half of the 76 resources (58%; N=44) addressed UDA, followed by UDL (49%; N=37) and UDI (1%; N=1). No resources addressed UID. Many resources (52%; N=39) discussed UD generally; some of these also named a specific framework. Slightly less than half of the resources (49%; N=37) mentioned multiple UD frameworks, including making general references to UD.

Figure 5. Conceptual Frameworks Represented in Overall Resources (Number)



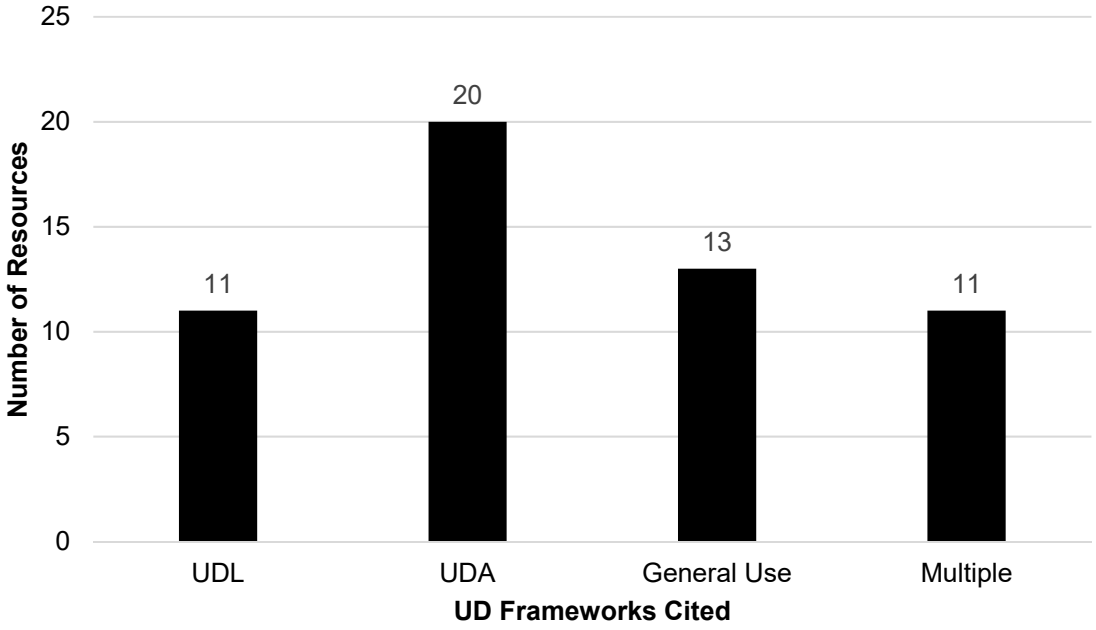
Notes. Each resource could include more than one conceptual framework. UDA = Universal Design of Assessment; UDL = Universal Design for Learning; UDI = Universal Design of Instruction; UID = Universal Instructional Design.

Research Resources

The concept of UD was described in various ways within the subset of research-based resources, as summarized in Figure 6. (For a detailed analysis, see Appendix E.) Eleven resources (39%) addressed CAST's UDL model (Abell & Lewis, 2005; Andersen & Nash, 2016; Beddow, 2011; Dolan et al., 2005; Fleming et al., 2006; Guerreiro et al., 2020; Housh et al., 2020; Kavanaugh, 2017; McMahon et al., 2016; Shelton, 2012; Wilson, 2015). Twenty research-based resources (68%) addressed Thompson et al.'s (2002) UDA model (Baker, 2008; Beddow, 2011; Bernstein, 2021; Cohen et al., 2019; Dembitzer, 2016; Dolan et al., 2005; Fleming et al., 2006; Jamgochian, 2010; Johnstone, 2003; Johnstone et al., 2005; Johnstone et al., 2006; Kavanaugh, 2017; Ketterlin-Geller et al., 2004; Liu & Anderson, 2008; Shelton, 2012; Shobe, 2020; Thompson, Johnstone, Anderson, & Miller, 2005; Thompson, Johnstone, Thurlow, & Altman, 2005; Thompson & Thurlow, 2003; Wilson, 2015). Thirteen research-based resources (43%) described UD in a general way, either alone or in combination with other frameworks (Baker, 2008; Beddow, 2011; Bernstein, 2021; Fleming et al., 2006; Guerreiro et al., 2020; Hansen et al., 2008; Jamgochian, 2010; Ketterlin-Geller et al., 2004; Rieke et al., 2013; Shelton, 2012; Shobe, 2020; Shyyan et al., 2015; Thompson, Johnstone, Thurlow, & Altman, 2005). Eleven of the research-based resources (39%) referred to multiple frameworks including UDL, UDA, or general UD (Beddow, 2011; Bernstein, 2021; Dolan et al., 2005; Fleming et al., 2006; Guerreiro

et al., 2020; Jamgochian, 2010; Kavanaugh, 2017; Shelton, 2012; Shobe, 2020; Thompson, Johnstone, Thurlow, & Altman, 2005; Wilson, 2015).

Figure 6. UD Frameworks Cited by Research Resources (Number)



Notes. Each resource could include more than one conceptual framework. UDA = Universal Design of Assessment; UDL = Universal Design for Learning

Despite frequent references to UD, only three of the research resources (11% of this subset of resources) explicitly tied study findings back to a UD framework (Housh et al., 2020; Johnstone et al., 2006; McMahon et al., 2016; see Appendix E for details). Housh et al. (2020) examined the extent to which computer-based science assessments with simulations catered to UDL principles. They argued that universally-designed assessments should have alternative formats, such as being embedded in immersive interfaces designed to be appealing to students. In the study, 20 Midwestern students in grades 6 and 9 completed three simulation-based science assessments, with each assessment composed of a variety of item types including matching and ranking items, videos, simulations, and multiple-choice questions. After taking the assessments, students participated in a 20-minute semi-structured interview to discuss their experiences. Students appreciated the variety of assessment tasks that helped engage them and helped them better understand the science constructs in the assessment. Analysis of interview transcripts identified three themes (relevance, interest, variety) that were mapped onto the UDL principles of representation, engagement, and expression and their attributing tasks. Multiple forms of representation such as videos, simulations, and model building mapped on to all three themes. Multiple forms of engagement, such as model building, simulation, and videos also mapped onto all three themes. Multiple forms of action or expression, such as model building, constructed

response, multiple choice questions, and matching and ranking items mapped onto interest and variety. The authors indicated that the assessments in their study were an example of how UDL considerations can be applied, but they also indicated that UDL concepts must be present in both the curriculum and assessments.

Johnstone et al. (2006) conducted think-alouds with 4th and 8th-grade students with disabilities, English learners, and students without a disability who were proficient in English. The researchers used student outcomes data to select six problematic statewide assessment items per grade. They sat down with students individually, demonstrated how to conduct a think aloud, and had the students practice verbalizing while solving problems before recording the students' verbalizations. They coded individual student results through a product analysis to determine how many items the student solved correctly and through a qualitative analysis of student verbalizations. After organizing the results of the analysis by UDA element (see Table 1 and Appendix A), Johnstone and colleagues found that the think aloud data could directly address needed changes to five of the seven UDA elements. The changes were associated with the following elements: 2 (Precisely Defined Constructs), 3 (Accessible, Non-biased Items), 5 (Simple, Clear, and Intuitive Items), 6 (Maximum Readability and Comprehensibility), and 7 (Maximum Legibility). Think aloud data could not address Element 1 (Inclusive Testing Population) or 4 (Amenable to Accommodations; there was no braille form of the test so this could not be studied). Think aloud data were also useful for determining that construct-relevant content was too challenging on one item. Based on the think aloud results, the researchers determined which item design features needed to be adjusted to provide greater accessibility to students.

McMahon et al. (2016) examined the effect of a digitized podcast to deliver read aloud accommodations on mobile devices to students with disabilities and students with reading difficulties on a researcher-created science assessment. The authors stated that test accommodations provide greater test accessibility, which relates to UD. Forty-six 6th graders with reading difficulties from largely diverse backgrounds, including 16 students with disabilities, were randomly assigned to one of three testing conditions: (a) standard paper test with no accommodations; (b) group administered assessment with teacher-delivered read aloud; and (c) researcher-created podcast read aloud delivered on a mobile device. Study findings indicated podcast delivery of read aloud testing accommodations increased scores of both students with disabilities and general education students with reading difficulties compared to the standard paper assessment with no accommodations condition. The standard condition did not alter student achievement significantly compared to the teacher read-aloud condition. There was a greater increase in the percentage of items correct for students with disabilities who had the podcast read aloud than for students without disabilities (with reading difficulties) who had the podcast read aloud format.

The authors found that the use of read aloud via podcast delivery made science assessments more accessible for students who struggle with reading fluency. According to McMahon et al.

(2016), the content knowledge of students with disabilities in particular may have been better measured when they were provided with both written and audio versions of science items. They argued that the podcast read aloud accommodation may have also increased scores because it helped students focus their attention and have more time to complete items as a result. The authors directly linked the use of a read aloud accommodation to UDL principles. Both the teacher-delivered read aloud and the podcast-delivered read aloud provided multiple means of representation (*Principle 1.3*—provide an alternate to visual information). In addition, the podcast-delivered read aloud provided multiple means of engagement (*Principles 7.1*—optimize individual choice and autonomy and *8.2*—vary demands and resources to optimize challenge). It allowed students to be more independent in accessing the read aloud accommodation, and they could choose which words to have read aloud, reflecting multiple means of action and expression (*Principle 6.2*—options that support planning and strategy development).

Research Question Results: Question 2

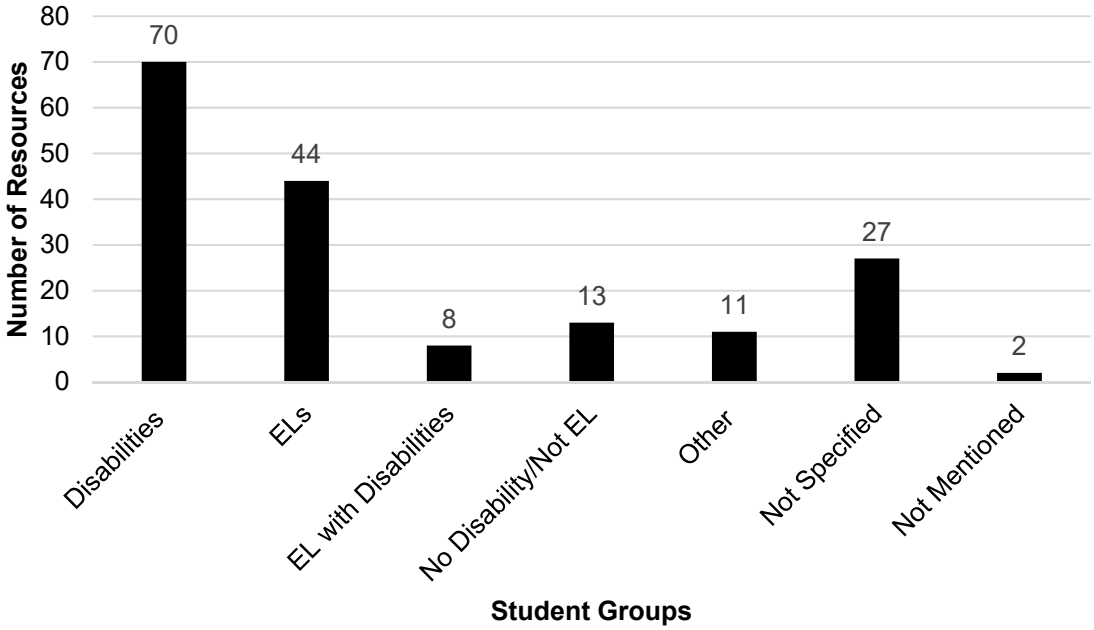
To what extent do publications address the application of universal design of large-scale assessments to students with disabilities and diverse students?

Overall

Historically, UD has been applied to instruction and assessment of students with disabilities, but more recently, UD has also been applied to English learners as well as other student groups. Figure 7 shows the student populations that resources addressed. Nearly all of the 76 resources addressed students with disabilities (92%; $N = 70$), and just over half addressed English learners (58%; $N = 44$). Other groups mentioned include English learners with disabilities (11%; $N = 8$) and students who did not have disabilities and were also not English learners (17%; $N = 13$). Eleven resources (14%) addressed students from other subgroups such as racial/ethnic groups or students with reading difficulties. More than a third of the resources (36%; $N = 27$) had at least some text that discussed UD in reference to "students" generally, without specifying what populations were included. Two resources (3%) discussed assessment concepts or obtained input from experts and educators and did not mention students at all.

Most resources addressed more than one student population, with 77% of resources including multiple student populations. If resources mentioned the grade level of the students described, it was typically a mixture of elementary and secondary students (23 resources). Only 14 resources addressed secondary students alone, followed by four that mentioned only elementary students. Still, just under half of the resources (34) did not specify the grade levels of students described in the text.

Figure 7. Student Groups Mentioned in Resources Reviewed (Number)



Notes. EL=English learner. Each resource could include more than one student group.

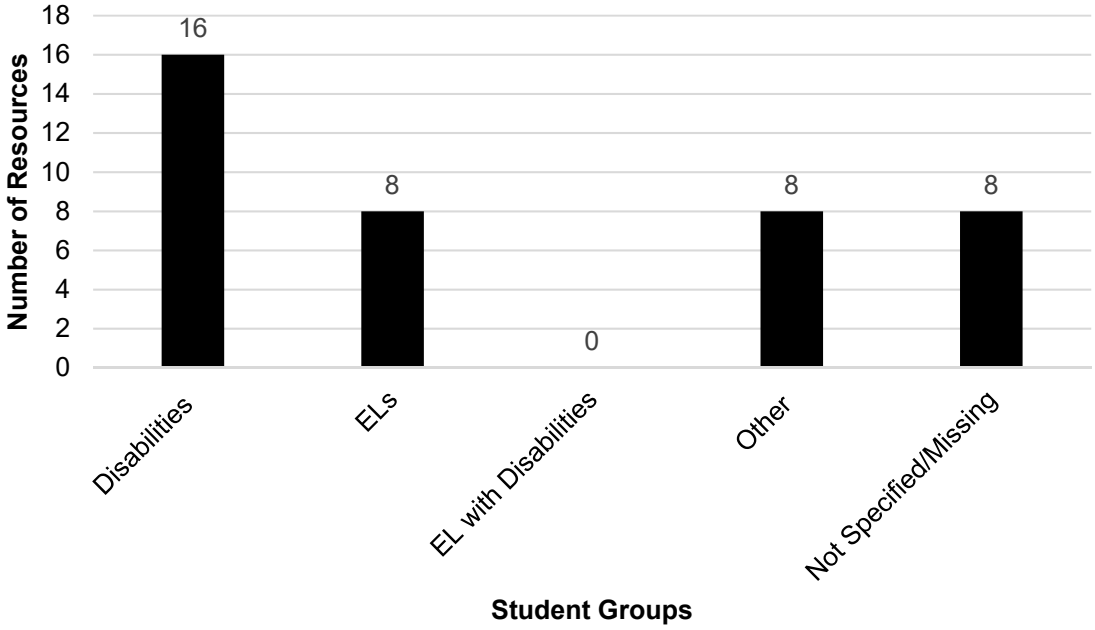
Research-based Resources

Several general observations emerged from our examination of the 28 research-based resources. For example, 13 of the resources seemed to relate UD to the use of accessibility features and accommodations by either evaluating their impact or examining individuals' perceptions of those features (Abell & Lewis, 2005; Beddow, 2011; Bernstein, 2021; Cohen et al., 2019; Dembitzer, 2016; Dolan et al., 2005; Hansen et al., 2008; Kavanaugh, 2017; Ketterlin-Geller et al., 2004; Liu et al., 2008; McMahon et al., 2016; Shelton, 2012; Wilson, 2015). Twelve resources evaluated design issues in existing large-scale assessments (Baker, 2008; Beddow, 2011; Bernstein, 2021; Cohen et al., 2019; Fleming et al., 2006; Housh et al, 2020; Johnstone et al., 2005; Kavanaugh, 2017; Ketterlin-Geller et al., 2004; McMahon et al., 2016; Shelton, 2012; Shobe, 2020). In addition, three research-based resources used released items from existing assessments to explore aspects of UD (Dolan et al., 2005; Johnstone, 2003; Johnstone et al., 2006). One other resource described using UD explicitly to inform the development of content standards and related assessments (Andersen & Nash, 2016), and one provided considerations for test development based on UD (Thompson, Johnstone, Anderson, & Miller, 2005). One resource identified prerequisite skills needed for students to participate in accessible computer-based assessments (Fleming et al., 2006).

The 28 research-based publications addressed the application of UD to large-scale assessments for a variety of students. Figure 8 summarizes the 20 resources that specified various groups of students as the research participants. (For a detailed table of findings for this research question, see Appendix D.) The majority of research-based resources included students with disabilities

(*N* = 16 or 84%; Abell & Lewis, 2005; Andersen & Nash, 2016; Beddow, 2011; Cohen et al., 2019; Dembitzer, 2016; Dolan et al., 2005; Fleming et al., 2006; Guerreiro et al., 2020; Johnstone, 2003; Johnstone et al., 2005; Johnstone et al., 2006; Kavanaugh, 2017; Ketterlin-Geller et al., 2004; McMahon et al., 2016; Shelton, 2012; Wilson, 2015). In addition, eight resources (42%; Abell & Lewis, 2005; Baker, 2008; Beddow, 2011; Guerreiro et al., 2020; Johnstone, 2003; Johnstone et al., 2006; Ketterlin-Geller et al, 2004; Shelton, 2012) included English learners. No resources included English learners with disabilities, while eight (25%) included other student groups, such as students from various racial or ethnic groups, those who were gifted and talented or higher performing, and those who were reading below grade level (Bernstein, 2021; Dembitzer, 2016; Guerreiro et al., 2020; Johnstone, 2003; Johnstone et al., 2006; McMahon et al., 2016; Shelton, 2012; Wilson, 2015). Eight research resources (29%; Baker, 2008; Beddow, 2011; Bernstein, 2021; Guerreiro et al., 2020; Housh et al., 2020; Johnstone et al., 2003, 2006; Shelton, 2012) either did not specify characteristics of students involved in the studies or were missing information on student characteristics for at least some students. Figure 8 does not include students without disabilities who were not English learners because it was often difficult to identify these students in study samples. In total, 15 research-based resources that specified participating student subgroups included more than one group of students.

Figure 8. Research Resources Involving Students from Various Groups as Participants (Number)



Note. Each resource could include more than one student group.

Table 3 provides data on the 16 research resources that included participants who were students with disabilities. Thirteen of these addressed students with disabilities generally; sometimes in addition to mentioning a specific group. Six studies included students with learning disabilities

(38%; Beddow, 2011; Dolan et al., 2005; Johnstone, 2003; Johnstone et al., 2006; Shelton, 2012; Wilson, 2015). One study included students with a significant cognitive disability (6%; Andersen & Nash, 2016), one included students with mild intellectual disabilities (6%; Wilson, 2015), one included students with disabilities who had a 504 plan (6%; Guerreiro et al., 2020), and one included students from the 13 federal disability categories (6%; Johnstone et al., 2005).

Table 3. UD Research Resources That Included Students from Various Disability Categories

Research Resources	Learning Disabilities	Significant Cognitive Disabilities	Mild Intellectual Disabilities	504 Plan	Multiple Categories	Students with Disabilities or IEP Generally
Abell & Lewis (2005)						X
Andersen & Nash (2016)		X				
Beddow (2011)	X					X
Cohen et al. (2019)						X
Dolan et al. (2005)	X					X
*Dembitzer (2016)						X
**Fleming et al. (2006)						X
Guerreiro et al. (2020)				X		X
Johnstone (2003)	X					X
Johnstone, et al. (2005)					X	
Johnstone et al. (2006)	X					X
Kavanaugh (2017)						X
Ketterlin-Geller et al. (2004)						X
*McMahon et al. (2016)						X
Shelton (2012)	X					
Wilson (2015)	X		X			X
Total (16)	6	1	1	1		13

Notes. *Author indicated some participants with disabilities had reading difficulties but did not provide disability categories for these students. **Study included students with disabilities who took an online assessment incorporating screen reader or text reader software. Disability categories not specified.

An additional nine research-based resources, shown in Table 4, included adult research participants such as researchers, educators, school administrators, disability advocates, or parents. These adults provided input or recommendations about ways to best assess students with disabilities, English learners, and English learners with disabilities. Almost all of these studies (89%; N=8) addressed students with disabilities who were not English learners in relation to UD. Six studies addressed English learners in relation to UD (67%; Liu & Anderson, 2008; Rieke et al., 2013; Shyyan et al., 2015; Thompson, Johnstone, Anderson, & Miller, 2005; Thompson, Johnstone, Thurlow, & Altman, 2005; Thompson & Thurlow, 2003), and three studies (33%) specifically addressed English learners with disabilities in relation to UD (Liu & Anderson, 2008; Rieke et al, 2013; Shyyan et al., 2015). Notably, none of these studies with adult research participants addressed students from other groups.

Table 4. Research Resources That Included Adult Participants Providing Information about Assessing Students from Various Backgrounds

Research Resources	Students with Disabilities	ELs	ELs with Disabilities
Hansen et al. (2008)	X		
Jamgochian (2010)	X		
Liu & Anderson (2008)		X	X
Rieke et al. (2013)	X	X	X
Shobe (2020)	X		
Shyyan et al. (2015).	X	X	X
Thompson, Johnstone, Anderson… (2005)	X	X	
Thompson, Johnstone, Thurlow… (2005)	X	X	
Thompson & Thurlow (2003)	X	X	
Total (9)	8	6	3

Notes. English Learners = ELs. Resources may have addressed more than one group.

Discussion

Of the 76 resources identified for this review, the majority were gray literature ($N = 52$; 68%). Reports or white papers published by non-governmental organizations made up the largest group of gray literature resources. This finding speaks to the importance of including resources published outside of academic journals to ensure emerging knowledge and knowledge of test vendors and state education agencies has been incorporated. Only about one-third ($N = 28$; 37%) of the 76 resources that met inclusion criteria were research-based studies.

In the resources included in this scoping review, particularly among the research-based resources, there was not a consistent UD framework or definition applied. In the total set of resources (N = 76), UDA was represented more often than UDL, but sometimes multiple frameworks were referred to, or no framework was mentioned at all. Likewise, in the subset of 28 research resources, more studies referenced UDA than UDL. However, again, resources often referred to multiple frameworks, without clearly articulating how frameworks with different components relate to each other. In some cases, published resources simply referred to UD without articulating any underlying framework.

Further, out of the 28 research resources, only three (Housh et al., 2020; Johnstone et al., 2006; McMahon et al., 2016) tied the study findings back to concepts associated with an underlying UD framework. The study authors argued for removing assessment barriers with various assessment components. The assessment components they promoted were more engaging assessments with more diverse item types (Housh et al., 2020) and more accessible read aloud accommodation (McMahon et al., 2016). Authors argued that these assessment strategies provided multiple means of representation, multiple means of action and engagement, and multiple means of action and expression. These concepts are elements found in the UDL framework developed by CAST (2018). The study by Johnstone et al. (2006) examined assessment item design through the lens of student think-aloud data and found that the data could point the researchers toward needed changes aligned with five of the seven UDA elements. Aside from these three studies, the assessment research to date is largely not linked to a set of elements associated with a specific UD framework. Thus, UD applied to K-12 large-scale academic assessments does not yet have a sufficient research base and has not been applied consistently. This is true despite the fact that UD is mentioned in federal law addressing K-12 statewide assessments, publicized by assessment vendors as part of their test development process, and named in many state and district assessment policies and requests for proposals.

UD resources examined in this review primarily addressed students with disabilities, but a growing number of studies examined UD for English learners, and a few studies addressed English learners with disabilities. When students with disabilities were referenced, more resources described the general population of students with disabilities, followed by resources that mentioned students with learning disabilities. Relatively few studies addressed students with significant cognitive disabilities who might be taking alternate assessments.

Throughout this review, the authors found that the term "Universal Design" in reference to assessments seems to have shifted to be equated primarily with providing accessibility features and accommodations to students with disabilities, English learners, and English learners with disabilities. This narrowing of the definition reduces UD to a single element according to UDA (Accessible, non-biased items), rather than including the full set of seven elements of universally designed assessments proposed by Thompson et al. (2002; UDA; see Table 1) or encompassing

the detailed sub-elements associated with CAST's (2018) three-part UDL conceptual framework (see Table 1). The one exception was Johnstone et al. (2006), which attempted to use student think-aloud data to address as many of the UDA components as possible. In addition, the idea that providing more accessibility features and accommodations makes a test more universally-designed contradicts Thompson et al.'s (2002) suggestion that fewer accommodations may be needed for universally designed assessments.

Limitations

This scoping review had a clearly defined and comprehensive search strategy but ultimately could have missed items that were irregularly indexed, such as book chapters in an edited volume. Future research should consider forward and backward citation searching and hand searching of book chapters.

In addition, gray literature such as white papers and reports on websites were included in this review in order to avoid publication bias and to capture emerging thinking that might be shared on test vendor or technical assistance center websites. It was important for our purposes to include everything that might be relevant to inform our audience of state education agency staff. However, the search tools available to us, such as Google site search, often produced large numbers of irrelevant or duplicate results that could not be imported directly into Covidence. In future reviews, it may be sufficient to include ERIC documents in the gray literature search without conducting additional website searches.

Conclusions

Our scoping literature review of the UD literature from 1985 to 2023 revealed a number of gaps, both overall and in the research literature. Perhaps the most glaring gap was the lack of specificity about which UD framework was the focus of the resources. Without this focus, especially in the research resources, it is unclear what is being studied and thus, what the results of the studies mean.

For researchers, the implications of our scoping review are straightforward. First, it is essential that researchers provide a clear description of the UD framework that is the focus of their studies. This must be done regardless of whether the research involves examining the effects of UD or the perceptions of UD. Further, the specific elements of the chosen framework should be delineated. Researchers should avoid equating UD with simply the development or provision of accommodations or other accessibility supports.

Our findings also have implications for state education agencies and test developers. To meet the federal requirements for UD applied to assessments, they must define the UD framework that is to be implemented. State education agencies should do this in requests for proposals for the development of assessments, in considerations of test vendors' proposals, and in evaluating assessments that are developed. Even if not required by funds, test developers should clearly define the UD framework that they are using when developing an assessment, as well as include an indication of the specific elements they will be using.

Continued research on the impact of UD on large-scale assessment accessibility will benefit the field and should result in higher-quality assessments for all students. With continued high-quality research, UD could have a solid research base, more powerfully informing test creation, test administration, and analyses of educational achievement for all.

References

Resources Cited in Text or Appendix Tables That Were Not in the Scoping Review

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009a). *Accessibility rating matrix (ARM)*. Vanderbilt University.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009b). *Test accessibility and modification inventory (TAMI) technical supplement.* Vanderbilt University.

Burghstahler, S. (2009). Universal design of instruction (UDI): Definition, principles, guidelines, and examples (ED506547). University of Washington, Project Do-IT. ERIC. https://files.eric.ed.gov/fulltext/ED506547.pdf

CAST. (2018). *The UDL guidelines version 2.2*. https://udlguidelines.cast.org/

Higbee, J. L., & Goff, E. (2008). *Pedagogy and student services for institutional transformation: Implementing Universal Design in higher education* (ED503835). University of Minnesota, College of Education and Human Development, Center for Research on Developmental Education and Urban Literacy. ERIC. https://files.eric.ed.gov/fulltext/ED503835.pdf

Lazarus, S., Goldstone, L., Wheeler, T., Paul, J., Prestridge, S., Sharp, T. Hochstetter, A., & Warren, S. (2021). *CCSSO accessibility manual: How to select, administer, and evaluate use of accessibility supports for instruction and assessment of all students.* Council of Chief State School Officers. www.ccsso.org

Ok, M. W., Rao, K., Bryant, B. R., & McDougall, D. (2017). Universal Design for Learning in pre-K to grade 12 classrooms: A systematic review of research. *Exceptionality*, *25*(2), 116–138. https://doi.org/10.1080/09362835.2016.1196450

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Resources Included in Scoping Review (N=76)

*Research-based Resource cited in Appendix D

*Abell, M., & Lewis, P. (2005). Universal design for learning: A statewide improvement model for academic success. *Information Technology & Disabilities Journal*, *XI*(1). http://itd.athenpro.org/volume11/number1/abell.html

AEL. (2005). *Universal design for assessment: An e-mail-based Soapbox discussion by IAETE, March 2005* (ED486655). ERIC. https://files.eric.ed.gov/fulltext/ED485655.pdf

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., Torres, C., Haertel, G., Dolan, R., Beddow, P., & Lazarus, S. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. *Journal of Technology, Learning, and Assessment*, *10*(5).

*Andersen, L., & Nash, B. (2016). Making science accessible to students with significant cognitive disabilities. *Journal of Science Education for Students with Disabilities*, *19*(1), 17–38. https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=1055&context=jsesd

Axelson, M. (2005). *Maximizing the effectiveness of online accountability assessments for students with disabilities* (ED489140). Appalachia Educational Laboratory at Edvantia. ERIC. https://files.eric.ed.gov/fulltext/ED489140.pdf

*Baker, H. E. (2008). *Effects of plain language revision on item difficulty, discrimination and DIF*. [Doctoral dissertation, University of Denver]. Electronic Theses and Dissertations.

Barker, E. (2016, May 10). *What does equity and accessibility look like within assessment?* Teach. Learn. Grow. https://www.nwea.org/blog/2016/what-does-equity-and-accessibility-look-like-within-assessment/

*Beddow, P. A. (2011). *Effects of testing accommodations and item modifications on students' performance: An experimental investigation of test accessibility strategies* (Publication No. 898546345; Issue 3479839) [Doctoral dissertation, Vanderbilt University]. ProQuest LLC.

Beddow, P. A., Kurz, A., & Frey, J. R. (2011). Accessibility theory: Guiding the science and practice of test item design with the test taker in mind. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests* (pp. 163-182). Springer.

*Bernstein, E. E. (2021). *Using the think aloud lab method to gather student input on statewide science assessments* (Publication No. 28323763) [Doctoral dissertation, Rutgers, The State

University of New Jersey]. ProQuest Dissertations Publishing. https://doi.org/doi:10.7282/t3-5nx3-sf34

Bowman, T., Wiener, D., Reavis, T., & Griswold, D. (2013). *PARCC accessibility features and accommodations manual: Guidance for districts and decision-making teams to ensure that PARCC mid-year, performance-based, and end-of-year assessments produce valid results for all students* (2nd ed.). Partnership for Assessment of Readiness for College and Careers. https://www.isbe.net/Documents/acc-features-accomm-manual.pdf

Bowman, T., Wiener, D., Reavis, T., & Griswold, D. (2014). *PARCC accessibility features and accommodations manual: Guidance for districts and decision-making teams to ensure that PARCC mid-year, performance-based, and end-of-year assessments produce valid results for all students* (3rd ed.) (ED561902). Partnership for Assessment of Readiness for College and Careers [PARCC]. ERIC. https://files.eric.ed.gov/fulltext/ED561902.pdf

Browder, D. M., Wakeman, S. Y., & Flowers, C. (2006). Assessment of progress in the general curriculum for students with disabilities. *Theory into Practice*, *45*(3), 249–259. https://doi.org/10.1207/s15430421tip4503_7

Case, B. J. (2008). *Universal design*. Policy Report. Pearson Education, Inc.

*Cohen, D. J., Zhang, J., & Wothke, W. (2019). Effects of item modifications on test accessibility for persistently low-performing students with disabilities. *Applied Measurement in Education*, *32*(4), 269–280. https://doi.org/10.1080/08957347.2019.1660343

College Board. (2021). *Fairness on the SAT suite of assessments: January 2021*. College Board. https://satsuite.collegeboard.org/media/pdf/sat-suite-fairness-standards-process.pdf

*Dembitzer, L. (2016). *Universal design and accommodations: Accessibility, reliability, and validity* (Publication No. 10297479). [Doctoral dissertation, Rutgers, The State University of New Jersey]. ProQuest Dissertations Publishing.

Dillon, N. (2006). Multiple choice: When it comes to assessment, special populations might need accommodations. *American School Board Journal*, *193*(1), 22–25.

*Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment*, *3*(7).

*Fleming, J., Kearns, J., Dethloff, A., Lewis, P., & Dolan, R. (2006). Technology skills checklist for online assessment. *Special Education Technology Practice*, *8*(1), 19–32.

Gordon, D. (2015). UDL's personal approach. *ESchool News*, *18*(4), 1–9. https://www.eschool-news.com/files/2016/05/eSchoolNews2015June.pdf

*Guerreiro, M. A., Barker, E., & Johnson, J. L. (2020, April 19). *Exploring student reading comprehension performance, equity, and engagement by embedding comprehension items within reading passages.* Paper presented at the 2020 annual meeting of the American Educational Research Association. https://doi.org/10.3102/1568124

Haertel, G. D., Cheng, B. H., Cameto, R., Fujii, R., Sanford, C., Rutstein, D., & Morrison, K. (2012). *Design and development of technology enhanced assessment tasks: Integrating evidence-centered design and universal design for learning frameworks to assess hard-to-measure science constructs and increase student accessibility*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/session1-cameto-cheng-haertel-paper-tea2012.pdf

Hanna, E. I. (2005). *Inclusive design for maximum accessibility: A practical approach to Universal Design* (PEM Research Report 05-04). Pearson Educational Measurement.

*Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2008). *Evidence-centered assessment design for reasoning about accommodations for individuals with disabilities in NAEP reading and mathematics.* ETS Research Report Series. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02124.x

Hasselbring, T. S., Lewis, P., Bausch, M. E., Axelson, M., Kay, K., & Honey, M. (2005). *Assessing students with disabilities: Moving forward through universal design* (ED489135). ERIC. https://files.eric.ed.gov/fulltext/ED489135.pdf

*Housh, K., Rehmat, A. P., Hmelo-Silver, C., Cisterna-Alburquerque, D., & Liu, L. (2020). *Evaluating computer-based science assessments for universal design learning principles*. International Society of the Learning Sciences. https://repository.isls.org/bitstream/1/6418/1/1739-1740.pdf

Jackson, R. M. (2005). *Curriculum access for students with low-incidence disabilities: The promise of universal design for learning.* National Center on Accessing the General Curriculum. (Links updated 2011). http://www.cast.org/products-services/resources/2005/ncac-curriculum-access-low-incidence-udl

*Jamgochian, E. M. (2010). *Designing and validating a measure of teacher knowledge of Universal Design for Assessment (UDA)* (Publication No. 749937191; Issue 3420440) [Doctoral dissertation, University of Oregon]. ProQuest LLC.

*Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). National Center on Educational Outcomes. https://nceo.info/Resources/publications/Onlinepubs/Technical37.htm

Johnstone, C., Altman, J., & Thurlow, M. (2006*). A state guide to the development of universally designed assessments*. National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/StateGuideUD/UDmanual.pdf

Johnstone, C., Altman, J., Thurlow, M., & Moore, M. (2006). *Universal design online manual*. National Center on Educational Outcomes. https://nceo.info/Resources/publications/UDmanual/default.html

Johnstone, C. J., Anderson, M. E., & Thompson, S. J. (2006). Universally designed assessments for ELLs with disabilities: What we've learned so far. *Journal of Special Education Leadership*, *19*(1), 27–33.

*Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/onlinepubs/Tech44/default.html

Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. L. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues & Practice*, *27*(1), 25–36. https://doi.org/10.1111/j.1745-3992.2008.00112.x

*Johnstone, C. J., Thompson, S. J., Moen, R. E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design* (Technical Report 41). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/TechReport41.pdf

Johnstone, C., Thurlow, M., Moore, M., & Altman, J. (2006). *Using systematic item selection methods to improve universal design of assessments* (Policy Directions 18). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/Policy18/PolicyDirections18.pdf

Karger, J., & Currie-Rubin, R. (2013). Addressing the educational needs of incarcerated youth: Universal Design for Learning as a transformative framework. *Journal of Special Education Leadership*, *26*(2), 106–116.

*Kavanaugh, M. (2017). *Examining the impact of accommodations and universal design on test accessibility and validity* (Publication No. 10264765). [Doctoral dissertation, Boston College]. ProQuest LLC.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing Universal Design for assessment. *Journal of Technology, Learning, and Assessment*, *4*(2).

Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice*, *27*(3), 3–16. https://doi.org/10.1111/j.1745-3992.2008.00124.x

*Ketterlin-Geller, L., Alonzo, J., & Tindal, G. (2004). *Use of focus groups to inform the construction of a universally designed mathematics test* (Technical Report 29) (ED531453). University of Oregon, Behavioral Research and Teaching. ERIC. https://files.eric.ed.gov/fulltext/ED531453.pdf

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, *84*(4), 529–551. https://doi.org/10.1080/01619560903240996

Lazarus, S. S., Johnstone, C. J., Liu, K. K., Thurlow, M. L., Hinkle, A. R., & Burden, K. (2022). *An updated state guide to universally designed assessments* (NCEO Report 431). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/NCEOReport431.pdf

*Liu, K. K., & Anderson, M. (2008). Universal design considerations for improving student achievement on English language proficiency tests. *Assessment for Effective Intervention*, *33*(3), 167–176. https://doi.org/10.1177/1534508407313242

Lovett, B. J., & Lewandowski, L. J. (2015). Universal design for assessment. *Testing Accommodations for Students with Disabilities: Research-Based Practice,* 207–223. https://doi.org/10.1037/14468-010

*McMahon, D., Wright, R., Cihak, D. F., Moore, T. C., & Lamb, R. (2016). Podcasts on mobile devices as a read-aloud testing accommodation in middle school science assessment. *Journal of Science Education and Technology*, *25*(2), 263–273. https://doi.org/10.1007/s10956-015-9591-3

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., & Vendlinski, T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation*, *19*(2–3), 121–140. https://doi.org/10.1080/13803611.2013.767614

Muller, E., & Tschantz, J. (2003). *Universal design for learning: Four state initiatives. Quick Turn Around (QTA)* (ED478563). National Association of State Directors of Special Education. ERIC. https://files.eric.ed.gov/fulltext/ED478563.pdf

NCEO. (2012). *Including students with disabilities in common non-summative assessments* (NCEO Brief 6). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/briefs/brief06/NCEOBrief6.pdf

New Mexico Public Education Department. (2021). *NM-ASR and NM-MSSA Spring 2021: Accessibility features and accommodations manual.* https://newmexico.onlinehelp.cognia.org/wp-content/uploads/sites/10/2021/03/NM_20-21_Summative_AFAM_ADA.pdf

Pearson & CAST. (2010). *Universal design for computer-based testing (UD-CBT) guidelines* (Revision B). https://static1.squarespace.com/static/540fa848e4b0da480523125e/t/5b34e1d0aa4a993c6225dbf4/1530192342500/TMRS_RR_UDCBTGuidelinesrevB.pdf

*Rieke, R. L., Lazarus, S. S., Thurlow, M. L., & Dominguez, L. M. (2013). *2012 survey of states: Successes and challenges during a time of change.* University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/StateReports/2012StateSurvey.pdf

Riley, J. C. (2018). *Appendix C: Accessibility and accommodations manual for the 2018-2019 MCAS tests/retests*. Massachusetts Department of Elementary and Secondary Education. http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2019/NextGen%20ADA/2019_MCAS_NG_TechReport_Appendix%20C.pdf

Rose, D. H., Hall, T. E., & Murray, E. (2008). Accurate for all: Universal design for learning and the assessment of students with learning disabilities. *Perspectives on Language & Literacy*, *34*(4), 23–28.

Russell, M. (2012). Accessible next generation assessments. In M. C. Mayrath (Ed.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 371-386). Information Age Publishing.

Russell, M., Hoffmann, T., & Higgins, J. (2009). Meeting the needs of all students: A universal design approach to computer-based testing. *Innovate: Journal of Online Education*, *5*(4). https://www.learntechlib.org/p/104243/article_104243.pdf

Russell, M., Hoffmann, T., & Higgins, J. (2009b). Nimble tools: A universally designed test delivery system. *Teaching Exceptional Children*, *42*(2), 6–12.

Salend, S. (2009). Using technology to create and administer accessible tests. *Teaching Exceptional Children*, *41*(3), 40–51.

Samuels, C. A. (2008). Researchers piloting 'accessible' guidelines. *Education Week*, *28*(9), 10–11.

Scalise, K., Irvin, P. S., Alresheed, F., Zvoch, K., Yim-Dockery, H., Park, S., Landis, B., Meng, P., Kleinfelder, B., Halladay, L., & Partsafas, A. (2018). Accommodations in digital interactive STEM assessment tasks: Current accommodations and promising practices for enhancing accessibility for students with disabilities. *Journal of Special Education Technology*, *33*(4), 219–236. https://doi.org/10.1177/0162643418759340

*Shelton, A. (2012). *Comparing the performance and preference of students experiencing a reading aloud accommodation to those who do not on a virtual science assessment* (Publication No. 3510444) [Doctoral dissertation, Temple University]. ProQuest LLC.

*Shobe, R. E. (2020). *Striving for the ideal: Technology's role in creating accessible tests for students with disabilities* (Publication No. 28027579) [Doctoral dissertation, Pepperdine University]. ProQuest LLC.

*Shyyan, V., Lazarus, S. S., & Thurlow, M. L. (2015). *2014 survey of states: Initiatives, trends, and accomplishments.* University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/StateReports/2014StateSurveyReport.pdf

Strain-Seymour, E., Way, D., & Dolan, R. P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. Pearson.

*Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/TechReport42.pdf

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/OnlinePubs/Synthesis44.html

*Thompson, S., Johnstone, C., Thurlow, M., & Altman, J. (2005). *2005 state special education outcomes: Steps forward in a decade of change*. University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/OnlinePubs/2005StateReport.htm

Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions 14). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/PolicyDirections14.pdf

*Thompson, S., & Thurlow, M. (2003). *2003 State special education outcomes: Marching on*. University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/OnlinePubs/2003StateReport.htm

Thompson, S., Thurlow, M., & Malouf, D. B. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology*, *6*(1), 1–15.

Thompson, S. J., Thurlow, M. L., Quenemoen, R. F., & Lehr, C. A. (2002). *Access to computer-based testing for students with disabilities* (Synthesis Report 45). University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/OnlinePubs/Synthesis45.html

Thurlow, M. L. (2010). Steps toward creating fully accessible reading assessments. *Applied Measurement in Education*, *23*(2), 121–131. https://doi.org/10.1080/08957341003673765

Thurlow, M. L., Moen, R. E., Liu, K. K., Scullin, S., Hausmann, K. E., & Shyyan, V. (2009). *Disabilities and reading: Understanding the effects of disabilities and their relationship to reading instruction and assessment*. University of Minnesota, Partnership for Accessible Reading Assessment. https://rtc3.umn.edu/docs/OnlinePubs/PARA/DisabilitiesReadingReport/PARADisabilitiesReadingReport.pdf

*Wilson, C. L. (2015). *Students' perspectives on the impact universally designed assessments have on mathematics achievement* (Publication No. 1730285127; Issue 3729375) [Doctoral dissertation, Southern University and Agricultural and Mechanical College]. ProQuest LLC.

Witmer, S. E., Roschmann, S., Timmermans, R., & Los, J. (2022). Promoting student engagement during large-scale testing. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *95*(1), 42–53. https://doi.org/10.1080/00098655.2022.2028713

Appendix A

Table A1. Detailed Description of UDA and UDL Frameworks

Universal Design of Assessment (Thompson, Johnstone, & Thurlow, 2002)	**Universal Design for Learning** (CAST, 2018)
Principle and Elements	Principle and Elements
Principle One—Equitable Use: The design is useful and marketable to people with diverse abilities. 1a. Provide the same means of use for all users: identical whenever possible; equivalent when not. 1b. Avoid segregating or stigmatizing any users. 1c. Provisions for privacy, security, and safety should be equally available to all users. 1d. Make the design appealing to all users. **Principle Two—Flexibility in Use:** The design accommodates a wide range of individual preferences and abilities. 2a. Provide choice in methods of use. 2b. Accommodation right- or left-handed access and use. 2c. Facilitate the user's accuracy and precision. 2d. Provide adaptability to the user's pace. **Principle Three—Simple and Intuitive Use:** Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level. 3a. Eliminate unnecessary complexity. 3b. Be consistent with user expectations and intuition. 3c. Accommodate a wide range of literacy and language skills. 3d. Arrange information consistent with its importance. 3e. Provide effective prompting and feedback during and after task completion. **Principle Four—Perceptible Information**: The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.	**Provide Multiple Means of Action and Engagement** Provide Options for: *Recruiting Interest (7)* • Optimize individual choice and autonomy (7.1) • Optimize relevance, value and authenticity (7.2) • Minimize threats and distractions (7.3) *Sustaining Effort & Persistence (8)* • Heighten salience of goals and objectives (8.1) • Vary demands and resources to optimize challenge (8.2) • Foster collaboration and community (8.3) • Increase mastery-oriented feedback (8.4) *Self-Regulation (9)* • Promote expectations and beliefs that optimize motivation (9.1) • Facilitate personal coping skills and strategies (9.2) • Develop self-assessment and reflection (9.3) **Provide Multiple Means of Representation** Provide options for: *Perception (1)* • Offer ways of customizing the display of information (1.1) • Offer alternatives for auditory information (1.2) • Offer alternatives for visual information (1.3) *Language & Symbols (2)* • Clarify vocabulary and symbols (2.1) • Clarify syntax and structure (2.2) • Support decoding of text, mathematical notation, and symbols (2.3)

Universal Design of Assessment (Thompson, Johnstone, & Thurlow, 2002)	**Universal Design for Learning** (CAST, 2018)
Principle and Elements	Principle and Elements
4a. Use different modes (pictorial, verbal, tactile) for redundant presentation of essential information. 4b. Provide adequate contrast between essential information and its surroundings. 4c. Maximize "legibility" of essential information. 4d. Differentiate elements in ways that can be described… 4e. Provide compatibility with a variety of techniques or devices used by people with sensory limitations. **Principle Five—Tolerance for Error:** The design minimizes hazards and the adverse consequences of accidental or unintended actions. 5a. Arrange elements to minimize hazards and errors: most used elements, most accessible; hazardous elements eliminated, isolated, or shielded. 5b. Provide warnings of hazards and errors. 5c. Provide fail safe features. 5d. Discourage unconscious action in tasks that require vigilance. **Principle Six—Low Physical Effort:** The design can be used efficiently and comfortably and with a minimum of fatigue. 6a. Allow user to maintain a neutral body position. 6b. Use reasonable operating forces. 6c. Minimize repetitive actions. 6d. Minimize sustained physical effort. **Principle Seven—Size and Space for Approach and Use:** Appropriate size and space is provided for approach, reach, manipulation, and use regardless of user's body size, posture, or mobility. 7a. Provide a clear line of sight to important elements for any seated or standing user. 7b. Make reach to all components comfortable for any seated or standing user. 7c. Accommodate variations in hand and grip size. 7d. Provide adequate space for the use of assistive devices or personal assistance.	• Promote understanding across languages (2.4) • Illustrate through multiple media (2.5) *Comprehension (3)* • Activate or supply background knowledge (3.1) • Highlight patterns, critical features, big ideas, and relationships (3.2) • Guide information processing and visualization (3.3) • Maximize transfer and generalization (3.4) **Provide Multiple Means of Action and Expression** Provide options for: *Physical Action (4)* • Vary the methods for response and navigation (4.1) • Optimize access to tools and assistive technologies (4.2) *Expression and Communication (5)* • Use multiple media for communication (5.1) • Use multiple tools for construction and composition (5.2) • Build fluencies with graduated levels of support for practice and performance (5.3) *Executive Functions (6)* • Guide appropriate goal setting (6.1) • Support planning and strategy development (6.2) • Facilitate managing information and resources (6.3) • Enhance capacity for monitoring progress (6.4)

Appendix B

Table B1. Inclusion and Exclusion Criteria for Peer-Reviewed Literature

Inclusion	Exclusion
Published 1985–2023. *Rationale:* 1985 is the year Ron Mace coined term "universal design."	Published before 1985.
Available in English.	Not available in English.
Addresses a U.S. education context OR does not explicitly address a non-U.S. context. *Rationale:* To narrow the scope of this review in order to make it manageable, searches will only include literature published about U.S. educational contexts.	Explicitly addresses a non-U.S. context. *Rationale:* Excluded to make the search more manageable.
Journals that relate to education or testing.	Journals that do not relate to education or testing.
Full text available.	Full text not available.
Addresses K-12 large-scale academic assessments (i.e., state and district assessments). *Rationale:* To keep the lit review narrow, we chose to exclude classroom instruction and assessment (e.g., formative) or diagnostic assessment for programs and services.	Addresses academic assessments at pre-school or post-secondary levels; OR Does not address education; OR Addresses classroom or diagnostic assessments that are not large-scale academic assessments.
Refers to "universal design" or one of the models of universal design (UDL, UDA, UDI, UID) in the title, abstract or keywords. *Rationale:* Since there are different models of universal design applied in education and they have different components, this review will only examine the literature that authors specifically associate with "universal design" rather than trying to make a judgment of which elements of which model should be addressed.	Does not use "universal design" or a reference to one of the models of UD (e.g., UDL, UDA, UDI, UID) in the title, abstract or keywords. *Rationale:* If there is no reference to UD or one of the models in the title, abstract or keywords that will be taken as an indication that UD is not central to the theme of the publication.

Notes. UDL = Universal Design for Learning; UDA = Universal Design of Assessment; UDI = Universal Design of Instruction; UID = Universal Instructional Design

Table B2. Inclusion and Exclusion Criteria for Doctoral Dissertations (Gray Literature)

Inclusion	Exclusion
Published 1985–2023. *Rationale:* 1985 is the year Ron Mace coined term "universal design."	Published before 1985.
Ph.D. or Ed.D. Dissertation. *Rationale:* Ph.D. or Ed.D. dissertations are more likely to represent new knowledge based on research.	Master's Thesis or Capstone not at Ph.D. or Ed.D level. *Rationale:* Master's theses or capstones are less likely to represent new knowledge based on research.
Available in English.	Not available in English.
Addresses a U.S. education context OR does not explicitly address a non-U.S. context. *Rationale:* To narrow the scope of this review in order to make it manageable, searches will only include literature published about U.S. educational contexts.	Explicitly addresses a non-U.S. context. *Rationale:* Excluded to make the search more manageable.
Full text available.	Full text not available.
Addresses K-12 large-scale academic assessments (i.e., state and district assessments). *Rationale:* To keep the lit review narrow, we chose to exclude classroom instruction and assessment (e.g., formative) or diagnostic assessment for programs and services.	Addresses academic assessments at pre-school or post-secondary levels; OR Does not address education; OR Addresses classroom or diagnostic assessments that are not large-scale academic assessments.
Refers to "universal design" or one of the models of universal design (UDL, UDA, UDI, UID) in the title, abstract or keywords. *Rationale:* Since there are different models of universal design applied in education and they have different components, this review will only examine the literature that authors specifically associate with "universal design" rather than trying to make a judgment of which elements of which model should be addressed.	Does not use "universal design" or a reference to one of the models of UD (e.g., UDL, UDA, UDI, UID) in the title, abstract or keywords. *Rationale:* If there is no reference to UD or one of the models in the title, abstract or keywords that will be taken as an indication that UD is not central to the theme of the publication.

Notes. UDL = Universal Design for Learning; UDA = Universal Design of Assessment; UDI = Universal Design of Instruction; UID = Universal Instructional Design

Table B3. Inclusion and Exclusion Criteria for Book Chapters, Websites, and Other Publications (Gray Literature)

Inclusion	Exclusion
Published 1985–2023. *Rationale:* 1985 is the year Ron Mace coined term "universal design."	Published before 1985.
Published by government agency, education-related non-governmental organization (NGO), university, testing company, an association of test publishers, or education-related professional association. *Rationale:* Publications from these sources may represent newer, emerging research and writing on the topic which will help to give a fuller picture of the application of UD to assessment. Will also include information on policies.	Not published by government agency, education-related NGO, university, testing company, association of test publishers, or education-related professional organization. *Rationale:* To streamline the review process we will focus on those organizations most likely to have information on the topic of universal design and K-12 assessment.
Addresses large-scale K-12 educational assessment.	Addresses academic assessments at pre-school or post-secondary levels; OR Does not address education; OR Addresses classroom or diagnostic assessments that are not large-scale academic assessments.
Document, report, web article with author and publication date. *Rationale:* A publication with an author and date is more likely to include original research or thinking on large-scale assessment and universal design (i.e., new knowledge) compared to a publication without authors or a date.	Web page without authors or publication date, videos, document summary, list of resources, abstract, newsletter, news release or memorandum, master's theses. *Rationale:* These resources are less likely to include original research or thinking on large-scale assessment and universal design.
Full text available on internet.	Full text not available on internet.
Available in English.	Not available in English.
Most current version of document. *Rationale:* Thinking might have changed over time with new versions of a document. Take the newest thinking.	Document was a draft, summary version or has been replaced with another version. *Rationale:* Draft versions typically represent older thinking compared to the most current version of a document.

Inclusion	Exclusion
Mentions universal design in title, abstract, executive summary, introduction, or first chapter if there is no executive summary or abstract. a. If no abstract or keywords available, look at overview, executive summary or first chapter. *Rationale:* Since conceptions of what comprises Universal Design may vary, this review will only examine the literature that authors specifically associate with the concept rather than trying to make a judgment of which constructs in publications are or are not part of Universal Design.	Does not mention universal design in title, abstract, executive summary, introduction or first chapter. *Rationale:* For this review, we are focusing on literature that makes Universal Design central to the theme of the publication and describes the application of UD to assessment. Most reports contain critical information about the topic in an executive summary or abstract, or in the opening paragraphs of the document if there is no executive summary or abstract.

Notes. UDL = Universal Design for Learning; UDA = Universal Design of Assessment; UDI = Universal Design of Instruction; UID = Universal Instructional Design

Appendix C

Electronic Search Strategy

ERIC

1. TI (test OR tests OR testing OR assess* OR evaluat* OR "educational measurement" OR exam OR exams OR examination* OR quiz*) OR AB (test OR tests OR testing OR assess* OR evaluat* OR "educational measurement" OR exam OR exams OR examination* OR quiz*) OR KW (test OR tests OR testing OR assess* OR evaluat* OR "educational measurement" OR exam OR exams OR examination* OR quiz*)
2. DE ("Test scoring" OR "Test design" OR "Achievement Tests" OR "Standardized Tests" OR "Test Items" OR "Test Construction" OR "Testing Programs" OR "Test Validity" OR "Measurement Techniques")
3. TI ("Universal design") OR AB ("Universal design") OR KW ("Universal design")
4. (S1 OR S2) AND S3
5. (DE "Postsecondary Education" OR DE "Higher Education" OR DE "Higher Education" OR DE "Graduate Study" OR DE "Postdoctoral Education" OR DE "Undergraduate Study" OR DE "Adult Education" OR DE "Colleges" OR DE "Community Colleges" OR DE "Two Year Colleges" OR DE "Community Colleges" OR DE "Technical Institutes" OR DE "Universities" OR DE "Land Grant Universities" OR DE "Open Universities" OR DE "Research Universities" OR DE "State Universities" OR DE "Urban Universities") NOT DE ("Elementary School Students" OR "Middle School Students" OR "Elementary Schools" OR "Middle Schools" OR "High schools" OR "Junior high schools" OR "secondary schools" OR "Secondary School Students" OR "Secondary School Teachers" OR "Grade 1" OR "Grade 10" OR "Grade 11" OR "Grade 12" OR "Grade 2" OR "Grade 3" OR "Grade 4" OR "Grade 5" OR "Grade 6" OR "Grade 7" OR "Grade 8" OR "Grade 9" OR "Intermediate Grades" OR "Kindergarten" OR "Special Needs Students" OR "Special Education")
6. S4 NOT S5

Education Source

1. TI (School* OR academ* OR educat* OR classroom* OR learn* OR instruct*) OR AB (School* OR academ* OR educat* OR classroom* OR learn* OR instruct*) OR KW (School* OR academ* OR educat* OR classroom* OR learn* OR instruct*)
2. DE ("Preschool Children" OR "Kindergarten children" OR "preschool education" OR "school children" OR "Kindergarten" OR "elementary schools" OR schools OR "elementary education" OR "Fifth grade (Education)" OR "Fourth grade (Education)" OR "First grade (Education)" OR "middle schools" OR "primary education" OR "Second grade (Education)" OR "Sixth grade (Education)" OR "Seventh grade (Education)" OR "Third grade (Education)" OR " Middle school education" OR "Middle school students" OR "secondary education" OR

“high schools” OR “junior high schools” OR “Eighth grade (Education)” OR “Eleventh grade (Education)” OR “Ninth grade (Education)” OR “Tenth grade (Education)” OR “Twelfth grade (Education)”)
3. TI (test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*) OR AB (test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*) OR KW (test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*)
4. DE (“Test scoring” OR “Test design” OR “Testing” OR “Test design” OR “Educational tests & measurements” OR “Achievement tests” OR “Examination item analysis” OR “educational tests & measurements”)
5. TI (“Universal design”) OR AB (“Universal design”) OR KW (“Universal design”)
6. DE “Universal design”
7. (DE “Higher education” OR DE “Higher education of people with disabilities” OR DE “ Postsecondary education”) NOT S2
8. (S1 OR S2) AND (S3 OR S4) AND (S5 OR S6)
9. S8 NOT S7

PsycINFO (Ovid)
1. (School* OR academ* OR educat* OR classroom* OR learn* OR instruct*).ti,ab,id
2. exp elementary school students/ or high school students/ or junior high school students/ or kindergarten students/ or middle school students/ or exp preschool students/
(test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*).ti,ab,id
3. exp educational measurement/ OR standardized tests/ or test types/ or test norms/ or test standardization/
4. (“Universal design”).ti,ab,id
5. (exp higher education/ or exp colleges/) NOT 2
6. (1 OR 2) AND (3 OR 4) AND 5
8. 7 NOT 6

Academic Search Premier
1. TI (School* OR academ* OR educat* OR classroom* OR learn* OR instruct*) OR AB (School* OR academ* OR educat* OR classroom* OR learn* OR instruct*) OR KW (School* OR academ* OR educat* OR classroom* OR learn* OR instruct*)
2. TI (test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*) OR AB (test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*) OR KW (test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*)

3. TI (“Universal design”) OR AB (“Universal design”) OR KW (“Universal design”)
(S1 AND S2 AND S3)

Web of Science Core Collection
1. TOPIC(School* OR academ* OR educat* OR classroom* OR learn* OR instruct*) AND TOPIC(test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*) AND TOPIC(“Universal design”)
2. TOPIC(“higher education” OR universit*)
3. #1 NOT #2

Proquest Global Dissertations & Theses
1. noft(School* OR academ* OR educat* OR classroom* OR learn* OR instruct*) AND noft(test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*) AND noft(“Universal design”)
2. abstract((“higher education” OR universit*)) OR title((“higher education” OR universit*))
3. 1 NOT 2

EdArXiv
(test OR tests OR testing OR assess* OR evaluat* OR “educational measurement” OR exam OR exams OR examination* OR quiz*) AND (“Universal design”)
n=0

Appendix D

Table D1. UD Research-based Resources—Purpose, Research Questions, UD Connection, Research Method and Analysis

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Abell, M., & Lewis, P. (2005). *Universal design for learning: A statewide improvement model for academic success.*	**Purpose:** To describe a Kentucky initiative involving accessible digital curriculum materials, an infrastructure of software tools (e.g., text reader), and an online accessible version of the KY CATs state assessment with an embedded text reader. **Research Questions:** None given.	Digital curriculum and comprehension reading supports described as a UD principle. Work was part of statewide UD initiative. **Frameworks:** UDL	**Research Method:** Mixed methods. Teacher survey data on use of E-Text software during 2003-2004. Students who participated in the Spring 2004 CATS Online Assessment gave comments. **Data Analysis:** Not described.
Andersen, L., & Nash, B. (2016). *Making science accessible to students with significant cognitive disabilities.*	**Purpose:** To use Evidence-Centered Design (ECD) and Universal Design for Learning (UDL) to develop DLM alternate science content standards and assessments**. Research Questions:** 1. How can the science disciplinary core ideas, crosscutting concepts, and science and engineering practices described in *A Framework for K-12 Science Education* be made accessible to students with significant cognitive disabilities? 2. How can the new alternate standards be assessed? 3. How do we know the newly created alternate standards and assessments are accessible?	UD incorporated via consideration of student characteristics in drafting: (a) alternate standards and linkage levels, (b) essential elements and levels of cognitive complexity, (c) content and design of test items, (d) accessibility modifications to testlets based on assessment pilot data. **Frameworks:** UDL	**Research Method:** Mixed methods. ***RQ1***—Creation of universally designed alternate science content standards via content analysis of extant standards. Internal, external, and state education agency reviews. ***RQ2***—Development of DLM alternate assessment science test specifications for three grade levels using principles of ECD and prior DLM test development processes. Teacher-drafted items/testlets. Expert and SEA review of procedures and drafts. ***RQ3***—Pilot test. Each student took 9 testlets at one linkage level chosen based on information from their First Contact Survey. **RQ 3 Data Analysis:** Calculated percentage of students correctly answering pilot test items. P-values of items and testlets were reviewed to evaluate item quality and accessibility. Percentage of students selecting other answer options was examined to determine which other answer options were attractive to students and the cause of the attraction. Possible causes for low p-values were considered.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Baker, H. E. (2008). *Effects of plain language revision on item difficulty, discrimination, and DIF.*	**Purpose:** To examine whether a plain language revision of Colorado CSAP assessment items (grades 3-10) allowed better access for English learners with varying levels of proficiency. **Research Questions:** 1. Does the cognitive complexity (DOK) of items change after plain language revisions? 2. Is item difficulty impacted by plain language revision? 3. Is item discrimination impacted by plain language revision? 4. Do CSAP math items revised for plain language provide increased access to assessment items by reducing differences on individual item performance as measured by Differential Item Functioning (DIF) for all groups of ELLs (NEP, LEP and FEP)? 5. Are there certain plain language revisions (context, graphics, vocabulary/wording, sentence structure, and format/style) of math CSAP items that reduce DIF and affect difference in item difficulty?	Author describes applying UD and plain language separately in assessment item edits but in other places describes plain language as falling under "the larger umbrella of universal design" (p. 7). **Frameworks:** UDA, General use	**Research Method:** Quantitative. Post-hoc analysis of 2005 and 2007 CSAP math data. In 2006 after new math items were developed and reviewed, the Colorado Department of Education and the test vendor incorporated both UD and plain language into math test items. UD adjustments: (a) direct match between an assessment item and a single objective, (b) simplified directions, (c) consistency across items and assessments, (d) simplified fonts, (e) increased white space. Plain language adjustments: (a) accessible vocabulary, (b) clear style of discourse, (c) clear sentence structure, (d) clear item format. **Data Analysis:** Evaluated plain language revision for effects on item difficulty, item discrimination, DIF and DOK. DOK ratings from 2005 and 2007 were compared using Chi-square. One sample t-tests used to examine differences in DIF and item difficulty (p-value), as well as item discrimination for ELs and Non-ELs comparing the 2005 and 2007 revised items. Revised items were classified by type of linguistic change. Each type of change was examined using ANOVA to determine if a specific type of change affected differences in DIF and item difficulty.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Beddow, P. A. (2011). *Effects of testing accommodations and item modifications on students' performance: An experimental investigation of test accessibility strategies.*	**Purpose:** To examine: (1) the relative and additive effects of testing accommodations and test item modifications on computer-based test performance of a diverse sample of 7th graders, (2) relations of test accessibility with common psychometric indices, and (3) students' perspectives about access strategies and related issues. **Research Questions:** 1. What are the effects of testing accommodations, item modifications, and a combination of the two on test performance for students with different abilities? 2. What are the relations between item accessibility and other psychometric indices used to characterize items? 3. How do students with different abilities perceive the accessibility of items, their cognitive demand, their teachers' coverage of the content, and their own predicted performance?	Assessment item modification teams were trained on universal design principles. They used the *Test Accessibility and Modifications Inventory Accessibility Rating Matrix* (*TAMI-ARM;* Beddow et al., 2009a) to modify items to increase accessibility. TAMI-ARM contains some elements that might be associated with UD (e.g., reducing answer choices, simplifying language in item stimuli and stems, reducing complexity of visuals, etc.). **Frameworks:** UDL, UDA, General use	**Research Method:** Quantitative. 2 x 4 experimental design used to determine the effects of accommodations and modifications on test performance on two groups of 7th graders (with and without an IEP) who were randomly assigned to one of four groups (1) original test items, no accommodations; (2) original test items, accommodations; (3) modified test items, no accommodations; or (4) modified test items, accommodations. Random assignment of students to one of two 34-item computer-based researcher-developed math test forms. Form A = items from the Discovery Education Assessment grade 6 test item bank delivered in original form. Form B = Form A items modified using the TAMI-ARM. If student was assigned to an accommodated condition, their teacher completed an accommodations checklist before testing to determine what the student would use. Accommodations from AZ's list were offered. From that list, students with disabilities could use what was in their IEP. Teachers could recommend accommodations for other students. Following each test, student participants completed a questionnaire about some of the items they saw on the test. **Data Analysis:** Descriptive statistics, significance tests and correlational analysis.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Bernstein, E. E. (2021). *Using the think aloud lab method to gather student input on state-wide science assessments.*	**Purpose:** To gather think aloud evidence to evaluate whether computerized TNReady science test items, on which students could use embedded tools, evoked the intended cognitive and problem-solving processes, barriers that may have impeded their access (e.g., high reading loads), and whether valid inferences about students' science content knowledge could be drawn from the results. **Research Questions:** 1. What does the validity evidence based on cognitive processes indicate with regard to the new Tennessee state science assessments? 2. How do the questions on state science assessments evoke intended cognitive processes? 3. What is the relationship between Maze scores and number of correct responses to science items?	Author describes online assessment with embedded tools (e.g., highlighters, answer eliminators, bookmarks) as universally designed. **Frameworks**: UDA, General use	**Research Method:** Students took science test and completed think alouds using Johnstone et al. (2006) protocol. Observers coded students' verbal and nonverbal behaviors. Audio recordings were transcribed and coded. Following think alouds students completed a survey, and a Maze reading assessment. **Data Analysis:** Basic descriptive statistics (percentages of students engaged in each strategy), coding of item transcripts. One-tailed Pearson correlations examined relationships between Maze scores and total raw scores on assessment items for each grade. Two-tailed, independent sample T-tests compared mean word counts and Flesch-Kincaid grade reading levels of items. Coded transcriptions analyzed to determine which cognitive strategies were used to complete each item. Strategies recorded for each item were compared to item specifications to determine whether the cognitive processes students used were what the developers intended.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Cohen, D. J., Zhang, J., & Wothke, W. (2019). *Effects of item modifications on test accessibility for persistently low-performing students with disabilities.*	**Purpose:** To investigate effects of item modifications on test accessibility by reducing construct-irrelevant cognitive barriers for persistently low-performing fifth grade students with cognitive disabilities. **Research Questions:** 1. Do the item modification methods (bolding/underlining, relevant pictures, and scaffolding) reduce construct-irrelevant cognitive complexity, thus improving the accessibility of the item for the target population? 2. Do the item modification methods improve item accessibility for students with specific cognitive deficits? 3. Do item modification methods reduce construct-irrelevant cognitive complexity, thus improving the accessibility of the item for the target population? 4. Do item modification methods improve item accessibility for students with specific cognitive deficits?	Authors relate test item accessibility to universal design. **Frameworks:** UDA	**Research Method:** Quantitative. Statistical analyses of item response data from a statewide standardized achievement assessment with modified and unmodified 5th grade math and reading items. Item modification methods: *Math:* bolding/underlining, relevant pictures, and scaffolding. *Reading:* bolding/underlining, graphic organizers, and scaffolding. **Data Analysis:** Descriptive statistics; generalized linear mixed modeling analysis.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Dembitzer, L. (2016). *Universal design and accommodations: Accessibility, reliability, and validity.*	**Purpose:** To compare the accessibility, reliability, and validity of a computer-based test with universally designed accommodations (i.e., audio presentation and extended time) on the performance of students with and without a functional reading impairment. **Research Questions:** 1. What is the perceived accessibility of a universally designed assessment with choices of accommodations? 2. To what degree will the accommodations of extra time and audio presentation affect the reliability of scores? 3. To what degree will the accommodations affect the validity of the inferences that can be drawn from tests scores?	Author describes computer-based tests with embedded accommodations, like the one in the study, as universally designed. **Frameworks:** UDA	**Research Method:** Mixed methods. ***Stage 1***—Developed computerized reading test with two forms and two versions each (accommodated, non-accommodated). ***Stage 2***—Examined accessibility using the Test Accessibility and Modification Inventory (TAMI; Beddow et al., 2009b). ***Stage 3***—Field testing. Researchers identified students with functional [reading] impairment (SWFI) and without (SWOFI) using median Oral Reading Fluency CBM scores. Students were randomly assigned by group to take field test assessment in counterbalanced design (e.g., form and accommodations condition). **Data Analysis:** ***RQ 1***—Survey data and TAMI results analyzed qualitatively and via t-tests; ***RQ 2***—Analysis of item level scores by group and condition. Coefficient Alpha tests for equality. ***RQ 3***—(a) Test scores by group and condition: T-tests of main effects and interaction; (b) Number of accommodations accessed and test score increases; (c) Pearson correlation; (d) Words read correctly per minute and test scores for students with disabilities by condition; Pearson correlation.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Dolan, R., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). *Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities.*	To compare the performance of students with learning disabilities on a universally-designed multiple-choice U.S. history and civics test under two conditions: one using a computer-based testing system with text-to-speech (CBT-TTS) and one using a traditional paper and pencil version of the test. **Research Questions:** 1. Is CBT-TTS effective? 2. Is CBT-TTS an effective means for assessing high school students with LD compared to traditional paper-and-pencil tests with no read-aloud accommodation? 3. What aspects of CBT-TTS make it effective? 4. If effective, which components of the system may be responsible? 5. Would students use CBT-TTS in the real-world? 6. Would students choose to use CBT-TTS during testing if it was available?	Authors describe a more flexible, individualized assessment as universally designed. **Frameworks:** UDL or CAST, UDA or NCEO	**Research Method:** Mixed Method. Counterbalanced administration of paper/pencil assessment and computer-based assessment with text to speech. Additional data collected from surveys, interviews, field observations, and usage tracking. **Data Analysis:** ***Test data***—Quantitative. Matched sample comparison of means (t-test) conducted across three different sets of test stimuli: all reading passages, long passages (100+ words), and short passages (100 words or less). A correlation statistic r based on a transformation of the Cohen's d index was used to measure effect size. ***Other data***—Qualitative analysis.
Fleming, J., Kearns, J., Dethloff, A., Lewis, P., & Dolan, R. (2006). *Technology skills checklist for online assessment.*	**Purpose:** To identify prerequisite skills students with disabilities need to successfully participate in an accessible computer-based assessment (KY CATS Online). **Research Questions:** What technology prerequisite skills do students need to use accessible computer-based accountability assessments?	Authors state that a computerized assessment designed to incorporate text or screen readers is universally designed. **Frameworks:** UDL, UDA, General use	**Research Method:** Qualitative. Interviews, focus group, document analysis, and descriptive analysis. **Data Analysis:** Descriptive statistics.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Guerreiro, M. A., Barker, E., & Johnson, J. L. (2020). *Exploring student reading comprehension performance, equity, and engagement by embedding comprehension items within reading passages.*	**Purpose:** To explore embedding items within reading passages to improve student experience and performance. **Research Questions:** 1. What item design best incorporates embedding items within text for common stimulus passages? 2. Does embedding items within text have an effect on outcome score for reading comprehension common stimulus passages?	Embedding items in the reading passages provides a Universal Design option for reading assessments. **Frameworks:** UDL, General use .	**Research Method:** Mixed Method. Researchers created a reading test with items embedded within reading passages. Students could answer comprehension questions immediately after reading a section of text. The approach potentially reduces memory load and anxiety, provides support for English learners, offers student choice, and allows students to interact in different ways with tests. They compared performance on the adapted test and a traditional test. ***Phases 1 and 2*** (2017–2018)—Review of item style by samples of students in grades 2-6 to ensure intuitiveness, ease of use, student perceptions of the approach and overall user experience. ***Phase 3***—Within-subjects counterbalanced quasi-experimental design measured reading comprehension performance on both assessment types. Researchers implemented input from phases 1-2, administered the assessment, and looked at student performance across item type. **Data Analysis:** Paired samples t-test.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2008). *Evidence-Centered Assessment Design for Reasoning about Accommodations for Individuals with Disabilities in NAEP Reading and Mathematics.*	**Purpose:** To show how evidence-centered assessment design can be applied in the area of accommodations for students with disabilities on large-scale assessments. NAEP reading and math is used as an example but the approach can be extended to other types of assessments. **Research Questions:** None given.	Assessments designed with accommodations and accessibility features are described as universally designed. **Frameworks:** General use	**Research Method:** Mainly theoretical with reference to some quantitative research findings from NAEP reading and math assessments. The paper presents how a validity argument could be constructed for an accommodated test score. A Toulmin diagram (Toulmin, 1958) is used to present the basic structure of the argument. The authors then describe considerations in creating a Bayes net (Belief network) model that attempts to address issues regarding some accommodations for NAEP reading and math. **Data Analysis:** Authors ran Bayes net model for several example students with disabilities using test accommodations (e.g., blind student using a read aloud accommodation), highlighting issues for further investigation. Reading and math models were run for seven exemplar students (no disability, blindness, low vision), on either the reading or math assessment, with three accommodations (large font, read aloud, braille).

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Housh, K., Rehmat, A.P., Hmelo-Silver, C., Cisterna-Alburquerque, D., & Liu, L. (2020). *Evaluating computer-based science assessments for universal design learning principles.*	**Purpose:** To examine the extent to which computer-based science assessments with simulations cater to UDL principles. **Research Questions:** 1. How did the design elements of science assessment tasks cater to UDL? 2. What attributes of the science simulations contributed to students' engagement across the three tasks?	Universally designed assessments should have alternative formats, such as embedding assessments in immersive interfaces like the simulations used in this study. During qualitative data analysis identified three themes (relevance, interest, variety) that were mapped onto the UDL principles of representation, engagement, and expression and their sub-principles. **Frameworks:** UDL	**Research Method:** Qualitative. Students completed 3 simulation-based science assessments with 14 items. Each assessment contained a variety of item types including matching and ranking items, model building tasks, videos, simulations, multiple choice questions, and constructed response items. Each student then participated in a 20-minute semi-structured interview to discuss their experience. Interview questions not provided. **Data Analysis:** Qualitative analysis of interview themes.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Jamgochian, E. M. (2010). *Designing and validating a measure of teacher knowledge of Universal Design for Assessment (UDA).*	**Purpose:** To design and validate a measure of teacher knowledge of Universal Design for Assessment (TK-UDA). **Research Questions:** 1. Is the content of the measure representative of the seven UDA principles? 2. Does the measure yield scores that reflect a continuum of teacher knowledge? a. Is performance on background knowledge items correlated with performance on declarative and applied knowledge items? b. Are declarative and applied knowledge scores correlated, forming a single UDA knowledge measurement dimension? c. Are teachers' declarative and applied knowledge of UDA scores structured from high (declarative) to low (applied)? 3. Does the measure effectively differentiate levels of expertise, in relation to: a. Teacher knowledge of UDA (overall)? b. Types of knowledge (background, declarative, applied)? 4. Are UDA element domain scores (sub-scores) from applied knowledge (scenario) items useful for identifying professional development needs? a. Are domain scores correlated, forming a single UDA skill measurement dimension? b. Are domain scores differentially difficult? c. Do domain scores differentiate experts from non-experts?	Developed instrument to describe teachers' knowledge of universal design for assessment and to identify professional development needs at the UDA element level. **Frameworks:** UDA, General use	**Research Method:** Quantitative. Iterative design of an online assessment of teacher UDA knowledge. **Data Analysis:** Different methods of analysis were used for each research question, including correlations between performance on background knowledge items and declarative and applied knowledge items (% correct scores), correlation between the declarative and applied knowledge sections (% correct scores, IRT scale scores, correlation coefficients), Cronbach's alpha, test-retest correlations, correlations between domain scores using item difficulties for each UD element in the applied knowledge section, rank ordering of items by item difficulty (mean rank per element), MANOVA.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal Design and student performance.*	**Purpose:** To determine whether eliminating construct irrelevant materials in tests gives a more accurate measure of student performance. **Research Questions:** 1. Is there an overall difference in test performance between traditionally and universally designed tests? 2. Does the relationship in performance on traditionally and universally designed tests differ between students with and without disabilities? 3. Do students perceive differences in the ease in which they completed the traditionally designed and universally designed tests?	Test was universally designed (e.g., removed construct irrelevant content and language, sensitivity review, removed unnecessary diagrams, simplified language, shorter sentences, untimed, large font, high contrast etc.). **Frameworks:** UDA	**Research Method:** Mixed methods. ***Study 1***—Creation of a traditional math test from released state assessment items. Creation of the universally designed items by researcher and advisory panel. Half the participants took the traditional assessment with standard procedures first. Students with disabilities used their IEP accommodations. This was followed by the universally designed test. The other half of the group took the tests in reverse order. ***Study 2***—Interviews of a subset of Study 1 students with and without disabilities. **Data Analysis:** ***Study 1***—Calculated group means for both assessments. Matched sample t-test to determine whether there was a statistically significant difference between group means. Cohen's d effect size calculated. ***Study 2***—Interviews transcribed and qualitatively analyzed to identify themes and representative quotes.
Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners.*	**Purpose:** To focus on the Think Aloud Method (Cognitive Laboratory) research methodology to detect design issues in large-scale tests, based on a framework of universal design. **Research Questions:** None given.	Used Think Aloud Method (Cognitive Laboratory) to detect design issues in large-scale tests, based on a framework of universal design. **Frameworks:** UDA	**Research Method:** Mixed methods. Student think alouds with 12 released state test items (6 in 4th grade and 6 in 8th grade) and a standardized think aloud protocol. One researcher conducted the think alouds, which were also videotaped and audiotaped, while the other observed them. **Data Analysis:** Product analysis used to determine whether students answered test items correctly or incorrectly. Video tapes were watched and coded for the major processes of test completion.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Johnstone, C. J., Thompson, S. J., Moen, R. E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design.*	**Purpose:** To provide states and test designers with methods of analysis for discovering which test items are valid for all students and which items may have issues related to universal design. **Research Questions:** None given.	Large-scale statistical analyses of item results to look at differential item functioning are a strategy for ensuring universal design of assessments. **Frameworks:** UDA	**Research Method:** Quantitative. Analysis of extant norm-referenced, multiple choice assessment data for 2000 from a large, Midwestern state's database. Students took the test under various accommodation conditions and under standard conditions. All data were included in the data set and correct or incorrect responses were entered. **Data Analysis:** Examined differential item functioning in four different ways: (a) Item ranking; (b) Item total correlation; (c) Differential Item Functioning: Contingency Table Methods; (d) Differential Item Functioning: Item Response Theory Approaches.
Kavanaugh, M. (2017). *Examining the impact of accommodations and universal design on test accessibility and validity.*	**Purpose:** To explore whether a technology-based approach results in scores that have similar psychometric and underlying structural qualities as scores collected under non-accommodated, paper-based administration. **Research Questions:** 1. Is the underlying factor structure consistent for scores gathered under accommodated and non-accommodated conditions? 2. Do items function similarly under accommodated and non-accommodated conditions? Specifically, holding ability constant, are item difficulty and discrimination equivalent for accommodated students and non-accommodated students? 3. If differential item functioning is exhibited, do patterns of DIF and item characteristics suggest that accommodations or use of accessibility supports may be related to DIF?	Examines two strategies to addressing accessibility: the use of technology to implement principles of universal design to assessment and the provision of accommodations. **Frameworks:** UDL, UDA	**Research Method:** Quantitative. Secondary data analysis of 2009 11th grade NECAP science assessment data from three states. Data for this study were collected for a larger study on the feasibility, effect, and capacity to deliver state achievement tests using NimbleTools, a computer-based test delivery system with embedded accommodations and accessibility features. Schools were given the option of using a paper assessment with traditionally offered accommodations or a computer-based test (on the NimbleTools platform) that offered embedded accommodations and accessibility supports. In general, schools chose to use Nimble Tools. **Data Analysis:** Differential item functioning analysis and confirmatory factor analysis (CFA) to examine the psychometric and structural qualities of student scores collected under different testing conditions.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Ketterlin-Geller, L., Alonzo, J., & Tindal, G. (2004). *Use of focus groups to inform the construction of a universally designed mathematics test.*	**Purpose:** 1. To identify the features of a universally designed 3rd grade math test that could be improved. 2. To explore the universal design needs of the student population. **Research Questions:** None given.	Study explored whether Universally-designed assessment (i.e., with accommodations built in) would meet the needs of students. Defines UD as applicability for all. **Frameworks:** General use	**Research Method:** Qualitative. Focus groups to examine the accessibility needs of students taking a previously developed, universally-designed assessment. Participants interacted with the computer-based assessment to ensure familiarity with the test design. Focus groups were recorded and transcribed plus field notes were taken. A follow up survey asked about universally designed assessment features they tried out, and what was perceived to be beneficial and difficult for students to use. UD features incorporated include: (a) use of basic skills test to determine linguistic complexity of items to present to individuals (e.g., read aloud, simplified text or standard administration), and (b) variations in language format (e.g., English only, bilingual Spanish and English). **Data Analysis:** Qualitative analysis by student and adult groups separately using the constant comparative method with induction. Researchers noted common themes as well as individual participant opinions.
Liu, K. K., & Anderson, M. (2008). *Universal design considerations for improving student achievement on English language proficiency tests.*	**Purpose:** To determine the usefulness and applicability of existing UDA considerations for 3 states' English language proficiency assessments. **Research Questions:** None given.	Applied UDA to ELP assessments. **Frameworks:** UDA	**Research Method:** Qualitative. Modified Delphi approach. **Data Analysis:** Descriptive analysis.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
McMahon, D., Wright, R., Cihak, D. F., Moore, T. C., & Lamb, R. (2016). *Podcasts on mobile devices as a read-aloud testing accommodation in middle school science assessment.*	**Purpose:** To examine the effect of a digitized podcast to deliver read aloud testing accommodations on mobile devices to students with disabilities and students with reading difficulties. **Research Questions:** (1) Do statistically significant differences exist between student performance on science content tests when compared across standard administration, teacher read aloud accommodation, and podcast read aloud accommodation? (2) Do significant differences exist across the testing conditions between students in special education and general education?	Authors state test accommodations provide greater test accessibility which relates to UD. Results are tied to UDL components 1.3 (multiple means of representation), 6.2 (multiple means of action and expression), and 7.1 and 8.2 (multiple means of engagement). **Frameworks:** UDL	**Research Method:** Quantitative. Generalized Latin Square Comparative design used to examine differences within and between student groups across three experimental conditions. Ninety multiple choice items were selected from a 6th grade formative science assessment series were randomly split into three versions of a 30-item science test. Researchers collaborated with 6th grade science and special education teachers to create an assessment protocol. Participants were randomly assigned to one of three testing conditions: (a) standard paper test with no accommodations, (b) group administered assessment with teacher-delivered read aloud, and (c) researcher-create podcast read aloud delivered on a mobile device. **Data Analysis:** Factorial Analysis of Variance (fANOVA) with test conditions and student status as the fixed factors.
Rieke, R. L. Lazarus, S. S., Thurlow, M. L., & Dominguez, L. M. (2013). *2012 survey of states: Successes and challenges during a time of change.*	**Purpose:** To provide a snapshot of new initiatives, trends, accomplishments, and emerging issues in assessment of students with disabilities in 2012. **Research Questions:** None given.	To find out how states used universal design. **Frameworks:** General use	**Research Method:** Mixed Methods. Survey. **Data Analysis:** Descriptive analysis.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Shelton, A. (2012). *Comparing the performance and preference of students experiencing a reading aloud accommodation to those who do not on a virtual science assessment.*	**Purpose:** To examine how the use of a read aloud accommodation (RAA) in immersive virtual environment science assessments affects the performance of middle school students. **Research Questions:** 1. While engaged in an immersive virtual environment assessment, to what extent and how does the expression of science content knowledge and in-world actions differ between students who experience a RAA when compared to students who experience text only as indicated by their sci-tools usage, interactions with characters and artifacts, and their answers provided at the end of the module? 2. How do significant student actions in question 1 vary by gender, ethnicity, English language learner status, and disability status, if at all? 3. To what extent and how do students perceive the RAA as helpful or distracting?	Technology is one way to achieve more UD of assessments because it is the most effective and efficient way to present information in a variety of modes. This study used a computer-based virtual assessment but also explored the combination of UD and technology (i.e., offering aural information instead of requiring students to read text). **Frameworks:** UDL, UDA, General use	**Research Method:** Quantitative. Random assignment of classes to treatment or control group. ***Pre-module***—Survey collected demographic information, and information on science anxiety and self-efficacy in science inquiry. Introductory assessment to acclimate students to the environment. ***During module***—Students completed contextualized science assessment by solving the embedded problem. Control group read text in the module to answer assessment questions. Treatment group was asked a series of questions orally by a module character. Data on students' actions and answers from the virtual environment-based module were recorded. Each module had different sci-tools. ***Post-module***—Survey similar to pre-module survey minus demographic information. Treatment group received an additional survey on their feelings about the treatment. **Data Analysis:** ***Q1***—Recorded student tool usage and interactions with characters and artifacts (dependent variables), along with an overall score percentage per module (via researcher scoring of open-ended response, short answer and multiple choice questions). One-way MANOVA of three dependent variables (overall score percentage, number of interactions, number of measurements) for each module with treatment as the independent variable. ***Q2***—Used significant dependent variables from RQ 1. Multiple regression analysis to look for significance on overall score percentage as the criterion variable and gender, ethnicity, EL, and disability status, treatment and their respective interactions with the treatment as predictor variables. ***Q3***—Descriptive statistics of post-module survey questions and to four randomly selected responses that students supplied to illustrate in-depth feelings about the treatment.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Shobe, R. E. (2020). *Striving for the ideal: Technology's role in creating accessible tests for students with disabilities.*	**Purpose:** To examine educators' perspectives on accessibility and the decision-making process educators use to select and implement test accommodations for students with disabilities when using Technology Enabled Assessments designed with the principles of Universal Design. **Research Questions:** How do educators perceive test accessibility for SWDs taking the TEA-UD (SBAC)? a. How do educators perceive the influence of the use of UD principles on SBAC as they relate to test accessibility for SWDs? b. How have embedded digital test accommodations on the SBAC influenced the educator decision-making process used to select test accommodations for SWDs? c. How has the use of technology influenced test accessibility for SWDs?	Elicits educators' perspectives about providing accommodations on a universally designed technology enhanced assessment (SBAC). **Frameworks:** UDA, General use	**Research Method:** Qualitative. Exploratory, qualitative, multiple case study design with semi-structured interviews with educators responsible for selecting 2018-19 SBAC accommodations for students with disabilities. Interview questions were piloted with two educators first. The pilot informed a revised interview protocol which was then administered to three participants. **Data Analysis:** Qualitative data analysis techniques including the constant comparative method. Data were analyzed individually and then across cases.
Shyyan, V., Lazarus, S. S., & Thurlow, M. L. (2015). *2014 survey of states: Initiatives, trends, and accomplishments.*	**Purpose:** To provide a snapshot of the new initiatives, trends, accomplishments, and emerging issues in assessment of students with disabilities in 2014. **Research Questions**: None given.	Section of the survey addressed state use of universal design in developing assessments. **Frameworks:** General use	**Research Method:** Mixed methods. Survey of state education agencies. **Data Analysis:** Descriptive analysis of results.
Thompson, S. J., Johnstone, C. J., Anderson, M., Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments.*	**Purpose:** This report summarizes the results of a validation process for the list of considerations for universally designed assessments. **Research Questions:** None given.	Considerations (questions for test designers) were developed from the elements in the UDA framework. **Frameworks:** UDA	**Research Method:** Qualitative. E-mail Delphi review with experts. Rounds focused on feedback on draft considerations and additional considerations that might be needed. **Data Analysis:** Not described.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Thompson, S., Johnstone, C., Thurlow, M., & Altman, J. (2005). *2005 State special education outcomes: Steps forward in a decade of change.*	**Purpose:** To provide a snapshot of the new initiatives, trends, accomplishments, and emerging issues in assessment of students with disabilities in 2005. **Research Questions:** None given.	Survey asked states, in section on emerging practices, to indicate the way in which they were addressing universal design of assessments. **Frameworks:** UDA, General use	**Research Method:** Mixed methods. Survey of state directors of special education about the participation and achievement of students with disabilities during standards-based reform. **Data Analysis:** Descriptive analysis.
Thompson, S., & Thurlow, M. (2003). *2003 State special education outcomes: Marching on.*	To provide a snapshot of new initiatives, trends, accomplishments, and emerging issues in assessment of students with disabilities in 2003. **Research Questions:** None given	Section of questions on the state survey asked about state use of universal design in developing assessments. **Frameworks:** UDA	**Research Method:** Mixed methods. Survey of state education agencies. **Data Analysis:** Descriptive analysis.

Citation	Purpose and Research Questions	UD Connection	Research Method and Analysis
Wilson, C.L. (2015). *Students' perspectives on the impact universally designed assessments have on mathematics achievement.*	**Purpose:** To examine perspectives of 8th graders with mild disabilities to determine the impact a universally designed computer-based assessment had on their ability to demonstrate mastery of a math concept. **Research Questions:** What impact do universally designed computer-based assessments have on the math achievement of eighth grade students with mild disabilities? (a) How do the students' perceptions of a universally designed computer-based assessment and a paper-based assessment affect their ability to demonstrate mastery of a math concept? (b) What are students' perceptions concerning the ease of use of a universally designed computer-based math assessment and a paper-based math assessment? (c) Would students perceive the universally designed computer-based assessment as a better option for evaluating math achievement when compared to the paper-based assessment?	Author describes the computerized test with built in accommodations as universally designed. **Frameworks:** UDL and UDA	**Research Method:** Mixed method. Administration of a paper-pencil math test (not described) and a universally-designed computer-based assessment taken on the Kurzweil 300 UD-integrated literacy program with 5 accommodations available: (a) Text to speech, (b) Talking electronic calculator, (c) Word predictor, (d) Dictionary, and (e) Highlighters. **Data Analysis:** Qualitative analysis (e.g., document analysis, classroom observations, student interviews, teacher interviews, student focus groups) determined how students scored and what their preferences were for the UD CBT vs. the paper pencil test. Thematic analysis with the constant comparative method. Also calculated basic descriptive statistics for the percentage of students answering interview questions in a particular way.

Table D2. UD Research-based Resources—Participants, Results, Conclusions/Implications/Recommendations

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Abell, M., & Lewis, P. (2005). *Universal design for learning: A statewide improvement model for academic success.*	Kentucky students with disabilities and their teachers are primary participants. Some English learners also use the tool. Number is unclear.	TextHelp software was most used in language arts and reading instruction. Use was sporadic but grew over time and writing grew the most. Teacher and student feedback was generally positive. Students were satisfied with the tool overall. Some teachers lacked consistent computer access and up-to-date technology to run the software. They wanted more support integrating the software tool into learning activities. For the online assessment with a built in text reader, students indicated they felt they performed better and that the reader helped them focus. The software allowed them to work at their own speed and they were less embarrassed at having a computer read to them.	Text reader technology was empowering to individual students. They were less self-conscious and could manage their own learning needs without depending on the teacher. With core content available in accessible digital formats paired with computerized reading and writing supports, students have more tools to assist and manage their own learning needs. ***Implications:*** None given.
Andersen, L., & Nash, B. (2016). Making science *accessible to students with significant cognitive disabilities.*	1,606 students with significant cognitive disabilities from four states (36% = elementary; 35% = middle school; 29% = high school; 54% = at target linkage level; 20% = precursor linkage level; 26% = initial linkage level).	*RQ 3:* 15% of pilot test items across grade levels were flagged for low p-values. The majority of items (85%) met the minimum threshold for the percentage of students answering correctly (i.e., 35% of students). Six of 81 testlets were rejected for being too difficult. Of these, five were at the precursor linkage level and one was at the initial linkage level. No data patterns indicated access differed by grade level. Data provided evidence of text accessibility. Item and test review data supported claims that items were high quality and not too difficult. Precursor level linkage descriptors asked students to use complex skills (e.g., developing models and making evidence-based claims). Success on these items relied on application of a science skill and a student's memory of facts.	Providing more context and activating prior knowledge by adding science stories could help compensate for memory limitations experienced by many students in this population. ***Implications:*** Special education teachers involved in developing test items/testlets struggled to understand how to make content accessible for students with significant cognitive disabilities. This indicates a need for professional development to ensure students are getting access to the content measured on assessments.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Baker, H. E. (2008). *Effects of plain language revision on item difficulty, discrimination, and DIF.*	Approximately 450,000 3rd-10th graders taking the CSAP in 2005 and 460,000 in 2007. English Learners: 69,872 in 2005 and 88,060 in 2007. Study compared English learners at all levels from non-English proficient to fluent English proficient (recently exited from EL programs) and non-English learners. Number of non-English learners unclear.	Results indicate little, if any, difference in item performance at individual grade levels. Overall, EL performance on revised items was higher than on non-revised items; the increase did not occur for non-ELs. *RQ 1:* DOK change was difficult to examine due to change of raters and calculation methods across years. DOK of both control items and plain language revised items was higher in 2007. *RQ 2:* The only significant change was for 5th grade ELs. EL control group item difficulty did not change across years. Item difficulty for 2007 5th grade revised items was significantly different from 2005 item difficulty. ELs performed better on revised items. *RQ 3:* Differences in item discrimination between 2005 and 2007 were not statistically significant. *RQ 4:* Overall, differences in DIF were not impacted by plain language revision. *RQ 5:* Specific types of revisions did not provide greater access for ELs than others.	Plain language revision holds promise as a best practice for item development. As an accommodation, it may not be sufficient to ensure access to assessment items for ELs. ***Recommendations:*** Conduct more research with/on: 1. Plain language items used with specific linguistic accommodations. 2. Matching a student's English proficiency score and content scores to better examine item difficulty. 3. A within-subjects design to better isolate changes in DOK ratings. 4. An in-depth examination of rater perceptions of items to understand rating variations. 5. Plain language and accommodations for non-English proficient students. 6. Distractor choices by subgroups to look at item bias. 7. DIF by linguistic subgroup rather than performance grouping. Future research should also continually examine the accuracy of DOK of assessment items, ensure there equal numbers of items in different revision categories, and gather information on students' educational background to shed light on the impact of linguistic barriers versus lack of content knowledge.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Beddow, P. A. (2011). *Effects of testing accommodations and item modifications on students' performance: An experimental investigation of test accessibility strategies.*	*Students:* 432 7th graders from CA and AZ. 103 students had an IEP; most in the area of specific learning disability. 329 7th graders with no IEP. One documented EL. Extensive missing demographic data. Two years of test data were available for only 71% of total sample.	*Q1:* No detectable effect of accommodations on student test performance for either group or for total sample. There was a moderate effect for modification (i.e., Form). *Q2:* Correlations between ARM rating and difficulty of the two forms were very small and negative for Form A, and small to moderate for Form B. Correlations between ARM rating and item discrimination were moderate. *Q3:* Ratings of comprehension, cognitive ease, and perceived performance for students with IEPs were significantly lower than for students with no IEPs. Mean helpfulness of accommodations, however, was not significantly higher for students with IEPs, and desirability was low for both groups. Most students (93%) did not use available accommodations.	If a student did not know the test content, an accessible test did not improve their score. Some students without IEPs also had difficulty accessing test items. The word count of an item appeared to have a stronger relationship to the student's score than the modification or accommodation. There are several relations between accessibility and validity-related variables including item difficulty, item discrimination, readability, student test self-efficacy, and ultimately student proficiency. The study may have benefited from delivering a specific package of accommodations to each student regardless of need or desire to use it. ***Recommendations:*** 1. Use the computer-based test delivery system to obtain student feedback on items and record the amount of time spent on each item. This would allow for an examination of the relationship between accessibility, cognitive load, and performance. 2. Replicate the study with a more robust sample including multiple grade-levels with and without disabilities, and ELs.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Bernstein, E. E. (2021). *Using the think aloud lab method to gather student input on state-wide science assessments.*	33 students (13 5th graders, 12 8th graders, and 8 high school students). Higher performing students were selected to increase likelihood of completing a think aloud.	No statistically significant relationships existed between the median Maze scores and the number of correct responses to science items. The average number of words per item was similar across grades. The Flesch-Kincaid grade reading levels of testing items were significantly different between fifth and eighth grade, and between eighth and high school, indicating that the reading levels of the tests were in keeping with students' grade levels. Students in lower grades had greater difficulty with think alouds and performed worse on the Maze assessments than students in upper grades. Older students used more of the embedded features. Raw scores were higher in younger grades. Most students felt the tests were not too hard and were good measures of their science knowledge.	Results provided enough validity evidence to support the conclusions that the TNReady science test items evoked the intended cognitive processes, participants interacted with the tests as intended, and the assessments were technically sound measures of students' science knowledge. Reading load was not a barrier. ***Recommendations:*** Test developers should ensure that item wording is clear, and the graphics are helpful and necessary. Schools should consider offering students more practice with computer-based tests and flexible options for test delivery. Future studies could include a more diverse sample (e.g., race and ethnicity, socioeconomic status, ELs, and students with disabilities).
Cohen, D. J., Zhang, J., & Wothke, W. (2019). *Effects of item modifications on test accessibility for persistently low-performing students with disabilities.*	*Target:* 5th graders with an IEP the previous year who were low performers on Ohio general state assessment for previous two years. *General Education:* 5th graders without IEPs. Numbers not specified.	Cognitive deficits were significantly related to low performance for students in general. Persistently low-performing students with disabilities showed higher possibilities of cognitive deficits than the general education students. Item scaffolding was an effective modification for both mathematics and reading. Other modifications, such as bolding/underlining of keywords, hindered test performance for low-performing students.	Item modification or test accommodation methods should be chosen to improve test accessibility for students with cognitive deficits without making the test less difficult for all. Modifications and accommodations make the test more valid for the general education students with cognitive deficits as well.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Dembitzer, L. (2016). *Universal design and accommodations: Accessibility, reliability, and validity.*	131 12th graders from three New Jersey high schools. 117 general education,5 special education, some not identified. Students with and without functional reading impairment chosen. Majority female 98 (75%) and ethnically diverse. African American = 43 (33%); Asian-Pacific Islander = 2 (2%); European American = 60 (46%); Latino/a = 10 (8%); Other = 16 (12%).	Students liked having accommodations available and rated the testing platform as highly accessible, but did not use many accommodations. There were complaints about the audio accommodation and the test was untimed. SWFIs spent more time and used more accommodations than SWOFIs. Reliability change using Feldt's test yielded mixed results for different forms of the test. Validity results were mixed as well. There was no increase in scores for either group with accommodations, no differential boost, and no relationship between boost and accommodations. SWFIs, who were the students who could benefit most from the accommodations offered, spent more time using them and used them more often. The relationship between reading fluency and reading comprehension lessened for SWFIs with accommodations.	TAMI results indicated an adequately accessible test. Consider individual need for accommodations, such as the need for reading assistance to overcome the barrier of reading fluency to support valid test inferences. Measurement of access skills can predict student need and use of accommodations
Dolan, R., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). *Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities*.	Nine grade 11 and 12 students with learning disabilities who were partially or fully included in general education classes.	Overall, students answered 65.3% of the items correctly when performing the test with the CBT-TTS versus 58.7% with the paper-pencil version. This difference was not statistically significant. However, when responding to items associated with long reading passages, students scored approximately 22 percentage points higher on the CBT-TTS administration (mean percentage score 76.7%) than the paper-pencil administration.	Results provide preliminary support for the potential benefits and usability of digital technologies in creating universally designed assessments.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Fleming, J., Kearns, J., Dethloff, A., Lewis, P., & Dolan, R. (2006). *Technology skills checklist for online assessment.*	*Focus groups:* 40 elementary, middle, and high school students with an IEP or 504 plan who took assessment and used screen readers. *Interviews:* 7 teacher test administrators. *Validation:* 10 teachers who administered 2003 assessment.	The research resulted in a Technology Skills Checklist that lists prerequisite skills in five areas: (a) *Basic Computer Skills*, (b) *Keyboarding Skills,* (c) *Word Processing Skills,* (d) Text-Reader or Screen-Reader Skills, (e) *Interaction with Online Assessment Skills (CATS Online).* The authors also identified issues needing attention prior to administering online assessments (e.g., Hardware/ Software Operation Check, Screen Reader/ Text-Reader Feature Check, and Ergonomic Items Check).	The technology checklist will help educators in identifying computer and software skills that must be taught prior to a student taking the KY assessment. It will also help school staff become aware of issues in need of attention to promote an effective testing environment and optimal systems operations.
Guerreiro, M. A., Barker, E., & Johnson, J. L. (2020). *Exploring student reading comprehension performance, equity, and engagement by embedding comprehension items within reading passages.*	130 3rd graders from Midwest elementary school. Female = 59 (45.4%), Male = 49 (37.7%); Unidentified = 22 (16.9%); White = 50 (38.5%); Non-White = 80 (61.5%); EL = 24 (18.5%); Gifted and Talented = 18 (13.8%); Special Education = 15 (11.5%); 504 = 5 (3.8%); Free and Reduced Lunch = 17 (13.1%)	*Phases 1 and 2:* Based on feedback, changes included designation of location and type of symbol used to denote an item embedded within the text, layout of passage and items, intuitiveness of item experience, ease of directions, and overall engagement with this new experience. Results were used to create the final item prototype used in Phase 3. *Phase 3:* On average, students scored higher on embedded items assessment (M = 227.91, SE = 3.28), than the traditional items assessment (M = 189.43, SE = 0.76). This difference, 38.48, 95% CI [32.30, 44.66], was significant t(114) = 12.33, p = .000, and represented a large-sized effect, r = 0.57.	Embedding items in text may be more valid approach to measuring reading comprehension and result in both improved comprehension scores and increased student engagement. Approach has potential to become a more equitable measurement of reading comprehension, particularly for marginalized groups.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2008). *Evidence-centered assessment design for reasoning about accommodations for individuals with disabilities in NAEP reading and mathematics.*	N/A—No human participants. Report describes issues for exemplar students with disabilities but says process is also applicable to ELs who receive accommodations.	ECD-based techniques were useful in analyzing the effects of accommodations and other accessibility features on validity. Generally NAEP accommodations practice is confirmed. For individuals with disabilities, NAEP allows accommodations that are valid according to the model and disallows ones that are not valid. There appears to be a general consistency between NAEP accommodation policy and the model based estimate of validity (generated via the Bayes net).	Application of ECD can help NAEP reconcile the need for accessible assessments for students with disabilities with the need to maintain or strengthen the validity argument for assessment score interpretations. ECD can help assessment planners expanding the list of available accommodations that yield valid assessment results and thus include students with a wider variety of disabilities. A robust validity framework can help determine the nature and scope of features that are made available under the heading of universal designed assessments for a given audience and purpose.
Housh, K., Rehmat, A. P., Hmelo-Silver, C., Cisterna-Alburquerque, D., & Liu, L. (2020). *Evaluating computer-based science assessments for universal design learning principles.*	20 6th and 9th grade students from rural Midwestern intermediate and jr. high school. Equal representation of grades and genders. No information on student demographics.	Students appreciated the variety of assessment tasks that helped engage them and helped them better understand the science constructs on the assessment. Analysis of interview transcriptions identified three themes (relevance, interest, variety) that were mapped onto the UDL principles of representation, engagement, and expression and their attributing tasks. Multiple forms of representation such as videos, simulations, and model building mapped on to all three themes. Multiple forms of engagement, such as model building, simulation, and videos also mapped onto all three themes. Multiple forms of Action or Expression (e.g., model building, constructed response, multiple choice questions, and matching and ranking items mapped on to interest and variety).	Assessments in this study are an example of how UDL considerations may be included in science assessments. Assessment simulations were designed around the interests of students to be inviting and appealing. UDL concepts must be prevalent in the curriculum and assessment.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Jamgochian, E. M. (2010). *Designing and validating a measure of teacher knowledge of Universal Design for Assessment (UDA).*	*Q1:* 9 participants total. Internal review = 3 researchers; External review = 3 experts + 3 teacher reviewers. *Q2:* 4 experts, 66 in-service teachers, 16 preservice teachers, *Retest participants:* 1 expert, 11 in-service teachers, 3 preservice teachers.	Results provided evidence indicating the need for measure revisions before claiming that the TK-UDA accurately described levels of teacher knowledge of Universal Design for Assessment. Based on results from *t*-tests and MANOVAs, no significant differences between groups (based on level of expertise) were found. Item Response Theory (IRT) scaling of items along a continuum indicated that declarative knowledge items were generally less difficult than applied knowledge items. IRT scaling of person scores represented a narrow range of knowledge within the sample. Reliability estimates from IRT scaling and test-retest indicated strong item reliability, relatively weak person reliability, and satisfactory test-retest reliability. A Kruskal-Wallis rank-order test conducted to evaluate the differential difficulty of UDA elements within the applied knowledge section provided initial evidence for identifying professional development needs at the element level.	*Recommendations:* The number of UDA elements could be reduced and the language of the elements simplified to be applied at the classroom level. Some of the elements may be less applicable to classroom assessments than they are to large-scale assessments (e.g., inclusive assessment population) and could be eliminated.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal Design and student performance.*	*Study 1:* 231 students from underperforming schools and populations in the South West. 165 Native American, 25 Latino/a, 23 Anglo, 18 missing demographic information. Also, 31 with a learning disability, 109 ELs, and 132 reading below grade level. *Study 2:* 23 students chosen from Study 1 because they showed 1.5 S.D. change between the control and experimental test scores.	Overall Universal Design principles applied as a group had a positive effect on student performance. 155 of 231 students had higher scores on a universally design test compared to a traditional test, 51 had lower scores, and 25 had the same scores. All improved test scores were significant but only 17 had a statistically significant negative score. Every subgroup showed improvement with the universally designed test. The difference in the means was statistically significant at the .05 level. Cohens d of .39 indicated a small-moderate effect of the universally designed assessment.	If constructs are held constant, the design of an assessment overwhelmingly influences student performance. Educators can use the findings to develop a better understanding of diverse students' capabilities. Item writers need training in universal design. More research is needed to verify the results and expand the definition of universal design. Universal Design has the potential to be the guiding philosophy for all test design. Its specific features could be continually researched and revised.
Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners.*	231 students from four schools in the Southwest. Demographic data were missing for 18 subjects. 165 Native American, 25 Latino/a, and 23 white. 31 students had specific learning disabilities, 109 were English learners, and 132 were reading below grade level.	Students who typically under-performed scored .39 standard deviations higher on tests that were designed using Universal Design principles. Descriptive statistics demonstrated that all subgroups achieved at a higher level on universally designed tests than traditionally designed tests. According to students, timing, readability, and recognizable materials were most important for high achievement on tests.	Overall, data demonstrated that UDA principles applied as a group to test items had a positive effect on student performance. If constructs are held constant, the design of a test can influence how a student performs. Findings provided educators with a better understanding of the capabilities of students who have diverse learning needs. They also reinforced the importance of ensuring that all item designers have been trained in the principles of Universal Design.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Johnstone, C. J., Thompson, S. J., Moen, R. E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design.*	4th and 8th graders from a variety of disability categories who took a large-scale math test in 2000. Numbers in each disability group ranged from 5 to 5,464 (4th grade) and from 4 to 5,498 (8th grade).	Results varied by analysis. All test items had universal design issues for some students. Each statistical test identified different problematic items. Pragmatic decision rules (e.g., finding patterns across disability groups and across analysis techniques) helped to reduce complexity of items with universal design issues.	Methods described in this paper may produce a manageable number of flagged items to investigate further. Secondary approaches, such as finding patterns in items across analysis methods, may help to better identify potentially problematic items (i.e., not universally designed). Large-scale analyses are relatively effective at finding problematic items, but they cannot say why items are problematic. Additional analyses (e.g., expert judgment, qualitative measurements) will support holistic decisions about items that are not universally designed and how they can be improved or eliminated from assessments. Improved items, in theory, will increase access to assessments for all students.
Kavanaugh, M. (2017). *Examining the impact of accommodations and universal design on test accessibility and validity.*	*Group 1:* 656 students in New Hampshire, Vermont, and Rhode Island completing the 11th grade science assessment using NimbleTools and assigned at least one embedded support. *Group 2:* 2,343 students completing paper-based assessment and assigned 1+ accommodations. *Group 3:* Random sample of 2,000 students selected from 28,464 students completing unaccommodated paper-based assessment. All students were eligible for accommodations. Overall, sample included 42.8% students with IEPs and 57.2% without IEPs. Fifty-five percent were male and 44.9% female.	Overall item functioning and underlying factor structure was consistent across accommodated and unaccommodated conditions, regardless of whether accommodations were provided with a paper form or a universally designed computer-based test delivery system.	Results support the viability of using technology-based assessments as a valid means of assessing students and offering embedded, standardized supports to address access needs.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Ketterlin-Geller, L., Alonzo, J., & Tindal, G. (2004). *Use of focus groups to inform the construction of a universally designed mathematics test.*	14 students (7 = general education; 2 = with disabilities; 5 = Spanish-speaking ELs). 16 adults (8 = parents of general education students, 3 = parents of students with disabilities; 5 = parents or relatives of Spanish speaking ELs; 7 = teachers and administrators; 5 = members of child advocacy groups).	Participant suggestions for improvement fell into 3 categories: (a) Make math test easier for people who have difficulty with using a mouse—some confused the right and left click buttons; (b) Reduce areas of uncertainty or confusion about test directions—clarify language and provide extra information about resources like read aloud (i.e., accommodations); (c) improve the test format so it is less distracting, clearer and has built-in help. In addition, survey results indicated that ELs potentially had differential access to computers compared to English-only speaker. This finding had implications for interpretations of scores from a computerized assessment.	Focus groups allowed researchers to better understand the characteristics of the test users and to identify features of the universally designed math test that inhibited student performance.
Liu, K. K., & Anderson, M. (2008). *Universal design considerations for improving student achievement on English language proficiency tests.*	Round 1: 21 experts representing assessment, ESL or bilingual education, and special education. Round 2: 17 of the 21 experts from Round 1. 4 ESL or Bilingual Education experts dropped out.	Considerations related to the creation of concise and readable text in assessment items received consistently high ratings overall. In contrast, considerations relating to the creation of test items that allow for format changes (e.g., Braille, oral presentation, sign language) showed the most variability in ratings.	Several changes were recommended to the original list of universal design considerations, including adding sections on types of tests and computer-based testing. Other changes addressed the format (e.g., adding notes on issues, developing a short form, adjusting some wording, and adding references, for example).

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
McMahon, D., Wright, R., Cihak, D. F., Moore, T. C., & Lamb, R. (2016). *Podcasts on mobile devices as a read-aloud testing accommodation in middle school science assessment.*	47 urban 6th graders with reading difficulties who were from the southeastern U.S. 16 were students with disabilities who typically received teacher read aloud on tests, 31 were students without disabilities who had reading difficulties. All students participated in reading intervention program and had reading levels between 1st and 3rd grade. 36 students (76.6%) were African American, 8 white (17.0%), 2 Hispanic (4.3%), 1 Native American (2.1%). More than half of students were male.	Both students with disabilities and general education students with reading difficulties demonstrated statistically significant gains based on their testing conditions. The data indicated podcast delivery of read-aloud testing accommodations increased scores of both students with disabilities and general education students with reading difficulties compared to the standard assessment condition and the accommodation did not alter student achievement significantly compared to the teacher read-aloud condition. There was a greater increase in the percentage of items correct for students with disabilities than for students with reading difficulties using the podcast read aloud format.	Results suggest podcast read aloud produced an increase in scores compared to the no accommodation condition and did not alter student achievement significantly compared to teacher read-aloud condition. It did so for both groups of students which may be due to the reading demands of science tests and particularly multiple choice test items. The authors state that results support the use of read aloud via podcast delivery as a way to make science assessments more accessible for students who struggle with reading fluency. The podcast read aloud accommodation may have also increased scores because it helped students focus their attention and have more time to complete items as a result. However, the podcast speaker may have unintentionally given cues to item answers through their delivery. Authors stated that the content knowledge of students with disabilities in particular may have been better measured when they were provided with both written and audio versions of science items. Authors directly linked the use of a read aloud accommodation to UDL principles. Both the teacher-delivered read aloud and the podcast-delivered read aloud provided multiple means of representation (*Principle 1.3*—provide an alternate to visual information). In addition, the podcast-delivered read aloud provided multiple means of engagement (*Principles 7.1*—optimize individual choice and autonomy and *8.2*—vary demands and resources to optimize challenge). It allowed students to be more independent in accessing the read aloud accommodation and they could choose which words and multiple means of action and expression (*Principle 6.2*—options that support planning and strategy development).

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Rieke, R. L. Lazarus, S. S., Thurlow, M. L., & Dominguez, L. M. (2013). *2012 survey of states: Successes and challenges during a time of change.*	Representatives from 49 regular states and 6 unique states.	Most states indicated they considered UD during test conceptualization and construction. Many also considered UD during the final review, during the expert review process, and in the test development request for proposals (RFP). Fewer states considered it as part of statistical analysis processes or through think-aloud methods during field testing. Few unique states addressed elements of universal design in the test development process.	None provided.
Shelton, A. (2012). *Comparing the performance and preference of students experiencing a reading aloud accommodation to those who do not on a virtual science assessment.*	791 6th to 8th graders in urban (N=285) and near urban (N=506) districts across two years. Slightly more females than males. Race/Ethnicity: White (48.5%), African American (11.9%), Latino/a (9.9%), Asian (8.3%), mixed (2.3%), American Indian (1%), European (.5%), Other (.6%), Missing (16.9%). Also, English learners (N=58), and students with learning disabilities (N=32). Extensive missing demographic data.	*RQ1:* Read aloud accommodation (RAA) treatment influenced student performance on one module (sheep) but not the other two (basketball, weather). RAA did not affect student's in-world data gathering via character interactions or tool usage. *RQ2:* Learning Disability and EL were the only significant predictors for student assessment performance. The effect of RAA in the first research question disappeared when demographic variables were considered. Student perceptions of RAA varied. Overall, survey means indicated that students had positive feelings about the RAA, although consensus was not overwhelming. Within case studies, only two students indicated that RAA helped them perform better during the assessment.	For students who scored higher with the RAA, it may have lessened construct irrelevant variance and more validly assessed their science knowledge. Use of the RAA did not provide a differential boost for students with learning disabilities or ELs. When student demographics were considered, the significant differences between the treatment and control groups disappeared, although data loss potentially skewed results. EL and learning disability status accounted for more of the variance between groups. Either reading construct irrelevance may not have been a major issue or demographic variables interfered. RAA treatment does not appear to remove systemic barriers. However, in general, students had positive perceptions of the accommodation. Because structural inequities were still evident despite the treatment, virtual environment assessments may also be perpetuating the unfairness of assessments.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Shobe, R.E. (2020). *Striving for the ideal: Technology's role in creating accessible tests for students with disabilities.*	4 licensed Oregon educators from 4 districts who were responsible for selecting test accommodations for students with disabilities on the 2018-19 Smarter Balanced Consortium assessment.	Teachers perceived the TEA-UD as more accessible for students with disabilities compared to the previous large-scale assessment or agreed that it had crucial accessibility features. Participants were able to identify UD principles within the test design that they thought contributed to accessibility but the identification of the principles' contribution varied. They also identified principles that they believed did not contribute to accessibility (but the identification of principles varied). Many test decision-making factors were discussed but only 1/3 were consistent across case studies. Participants said about 1/3 of the decision-making factors were influenced by whether the test was computerized with embedded accommodations. Participants strongly agreed that using technology to deliver assessments allows for more accessibility features but paper-pencil tests were more straightforward and required less time to train students. They also thought computerized assessments required more computer literacy skills. Five key themes: (a) Perception of overall test accessibility for students with disabilities; (b) Factors influencing the educator decision-making process; (c) Influence of UD on educators' perceptions of test accessibility (i.e., reflection on how Mace & Thompson et al.'s UD principles relate to SBAC); (d) Influence of use of technology-based assessments on educators' perceptions of accessibility (e.g., computer test allows for greater accessibility).	1. SBAC represents a step forward in the provision of accessible tests for students with disabilities in Oregon. 2. TEA-UD requires a significant amount of digital fluency. 3. Educators created a quick guide to condense and summarize details about the TEA-UD and accessibility that they could use in training other educators. 4. The 1%-2% cap on AA-AAAS participation could be unrealistic (e.g., some students fall in a gap between the TEA-UD and the AA-AAAS; the state test should really include all of those students to be universally designed).

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Shyyan, V., Lazarus, S. S., & Thurlow, M. L. (2015). *2014 survey of states: Initiatives, trends, and accomplishments.*	State directors of special education and assessment in 50 states and eight unique states.	More than 3/4 of regular states and 1/2 of unique states addressed Universal Design, especially during test conceptualization and construction, requests for proposals for test development, final reviews conducted with test contractors, and expert reviews. Several states indicated that Universal Design was addressed in other ways including in development of custom items and via test specifications.	None given.
Thompson, S. J., Johnstone, C. J., Anderson, M., Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments.*	13 experts	Universal Design considerations originally sent to reviewers were rated as somewhat important to extremely important for designing and reviewing assessments. One consideration was deleted based on expert feedback, while others were added or revised. Primary additions were the expansion of considerations for computer-based testing.	*Recommendations*: Include UD in early stages of test development, involve experts in disability, technology, and language acquisition, and provide professional development. Specific ideas for item reviews and field testing were provided.
Thompson, S., Johnstone, C., Thurlow, M., & Altman, J. (2005). *2005 State special education outcomes: Steps forward in a decade of change.*	State directors of special education in 50 states and six unique states.	Most regular states and a couple of unique states addressed universal design for their state assessments. Most did so in RFPs and during item development and review. Most states included at least one disability representative in bias or content review teams. States identified several emerging issues, including funding, timelines, and staff resources.	Document was descriptive without conclusions or interpretations.
Thompson, S., & Thurlow, M. (2003). *2003 State special education outcomes: Marching on.*	State directors of special education in 50 states and 11 federal jurisdictions.	Most regular states used some approach to universally designed assessments, primarily having a disability representative on the assessment bias review committee. None of the special education personnel in the unique states were aware of the development of Universal Design approaches for their state assessments.	Document was descriptive without conclusions or interpretations.

Citation	Participants	Results	Author Conclusions and Implications or Recommendations
Wilson, C. L. (2015). *Students' perspectives on the impact universally designed assessments have on mathematics achievement.*	Purposeful sample of 7 African American 8th graders with either mild ID or LD who received math in self-contained special education classroom in a rural Louisiana school district and their teacher. Focus group with other students included for triangulation of findings.	*RQ 1a:* Students had varying preferences for paper-pencil vs. computer testing. Students were generally more comfortable with paper and lacked familiarity with a computer. Those who liked the computer test thought it was easier to understand and said the paper test required too much writing. Teacher preferred paper tests. Students were observed to ask for help more often with paper pencil test form. *RQ 1b:* Four of 7 preferred the UD CBT because it was easier and they finished faster. Three of seven thought paper was easier. Students tended to experience more stress and anxiety on their non-preferred type of test. Question 1c: Answers aligned to students' assessment format preference. They chose the type of test they saw as less complex. Teacher said students more likely to guess at answers on CBT and less likely to use accommodations because students lacked computer experience. Overall: Most participants (77%) demonstrated higher academic achievement on the paper-based math assessment. Students' perceptions about the tests did not impact their performance. Students' lack of experience with the universally designed computerized assessment may have contributed to the rate of error demonstrated on this examination.	Student perceptions of test do not necessarily align to or impact test outcomes. Most students got lower scores on UD CBT than paper test even though they said they liked computer test better. Move to computer testing may be costly and non-productive. Novice users of UD assessments may get scores representing a "false negative" of their academic ability. Implied that this false negative is because test is on computer. Drastic declines in school performance schools may result, with other negative consequences.

Appendix E

Table E1. Universal Design Frameworks Referred to by Research-based Resources and Links Between Findings and Frameworks

Citation	UDL	UDA	General Use of UD	Multiple Frameworks	Findings Tied to a UD Framework
Abell & Lewis (2005)	X				
Andersen & Nash (2016)	X				
Baker (2008)		X			
Beddow (2011)	X	X	X	X	
Bernstein (2021)		X	X	X	
Cohen et al. (2019)		X			
Dembitzer (2016)		X			
Dolan et al. (2005)	X	X		X	
Fleming et al. (2006)	X	X	X	X	
Guerreiro et al. (2020)	X		X	X	
Hansen et al. (2008)			X		
Housh et al. (2020)	X				X
Jamgochian (2010)		X	X	X	
Johnstone (2003)		X			
Johnstone et al. (2006)		X			X
Johnstone et al. (2005)		X			
Kavanaugh (2017)	X	X		X	
Ketterlin-Geller et al. (2004)			X		
Liu & Anderson (2008)		X			
McMahon et al. (2016)	X				X
Rieke et al. (2013)			X		
Shelton (2012)	X	X	X	X	
Shobe (2020)		X	X	X	
Shyyan et al. (2015)			X		
Thompson, S. J., Johnstone, C. J., Anderson, M., Miller, N. A. (2005)		X			
Thompson, S., Johnstone, C., Thurlow, M., & Altman, J. (2005)		X	X	X	
Thompson & Thurlow (2003)		X			
Wilson (2015)	X	X		X	

Citation	UDL	UDA	General Use of UD	Multiple Frameworks	Findings Tied to a UD Framework
Total (28)	11 (39%)	19 (68%)	12 (43%)	11 (39%)	3 (11%)

Notes. UDL = Universal Design for Learning; UDA = Universal Design of Assessment; UD = Universal Design

INSTITUTE on COMMUNITY INTEGRATION

UNIVERSITY OF MINNESOTA

NCEO is an affiliated center of the Institute on Community Integration