Connecting Claims and Outcomes: Applying Accessibility Criteria to Alternate Assessments

Anne H. Davidson
EdMetric LLC

Kristine David
University of Kansas

Jill Christmus
South Carolina Department of Education

The study's purpose was to evaluate how accessibility is forwarded through technical test specification and how specifications are influenced by policy guidance. Although universal design (UD) is frequently identified as a guiding principle in test development, Johnson, Trantham, and Usher-Tate (2019) found many assessment programs neither realize these promises, nor ensure the necessary steps for optimal accessibility. We reviewed assessment development approaches and features in light of UD principles by conducting a qualitative review of relationships between UD elements and Peer Review Critical Elements (2018), and the relationships between UD elements and "Criteria for Procuring Evaluating High-Quality Assessments (CCSSO, 2014) using expert judgment (Patton, 2002). Results illustrated where raters identified UD elements within policy guidance and showed a concentration of references to UD in test development processes, consistent with findings from previous studies (Davidson, 2019). Results suggest the limited definition of fairness and a view that accessibility is only a consideration at the item level may contribute to the lack of connection to these UD elements in Peer Review guidance.

Keywords: accessibility, universal design, test specification, accountability criteria

INTRODUCTION

The broad purpose of the study was to evaluate how accessibility is forwarded through technical test specification and how specifications are influenced by policy guidance. Accessibility in test development has risen in priority for state decision makers, in large part because of its relation to validity arguments and contribution to a test's overall fairness (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In order to evaluate a claim that a given test is accessible to all students in a test's target population, criteria for accessibility must be defined. Therefore, policy makers and influencers have woven accessibility into those criteria that guide test development.

Despite the trend toward including accessibility concepts and practices in test specifications, the full realization of the goal for consequential, academic tests remains elusive. Therefore, we first examined technical test specification in light of the concept of Universal Design (UD) as one framework for accessibility considerations, and then we conducted an investigation into the relationship between UD and influential policy guidance that sets expectations for states' procurement and evaluation of consequential, large-scale assessments.

Role of Technical Specification

Test development practices have increasingly placed emphasis on *a priori* design, assuming that quality test design will translate to overall quality tests. For example, the Stanford Center for Assessment, Learning, & Equity (2016) called out item design principles critical to high-quality assessment. These design principles relate to two important, complementary ideas: evidence-centered design (ECD) and UD for learning (UDL; Rose, Meyer, & Hitchcock, 2005). Further, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) describe the components of consequential assessment programs that are required to support claims of fairness toward all examinees within a target population.

Universal Design

Universal Design is a concept increasingly applied in educational assessment as a guiding principle (Thompson, Johnstone, & Thurlow, 2002) and part of influential policy guidance. In contrast to specialized designs in which a technology is created for general use by people without disabilities and then modified or accommodated for use by people with disabilities, UD values the design of products and services that can be "usable by people with the widest range of functional capabilities" (Individuals with Disabilities in Education, 2004). Bowe (2000) defined UD in education as the "preparation of curricula, materials, and environments so that they may be used, appropriately and with ease, by a wide variety of people" (p. 45). Hehir (2002) argued that UD is a matter of simple justice that should be broadly applied to education.

Thompson, Johnstone, and Thurlow (2002) outlined seven key principles for use of UD in assessment. Table 1 summarizes these principles as elements.

TABLE 1
UNIVERSAL DESIGN ELEMENTS APPLIED TO ASSESSMENT DEVELOPMENT

UD Elements	Definition
1: Inclusive Assessment Population	Tests designed for state, district, or school accountability must include every student except those in the alternate assessment, and this is reflected in assessment design and field-testing procedures.
2: Precisely Define Constructs	The specific constructs tested must be clearly defined so all construct-irrelevant cognitive, sensory, emotional, and physical barriers can be removed.
3: Accessible, Non-Biased Items	Accessibility is built into items from the beginning, and bias review procedures ensure quality is retained in all items.
4: Amenable to Accommodations	The test design facilitates the use of needed accommodations (e.g., all items can be brailled).
5: Simple, Clear, and Intuitive Instructions and Procedures	All instructions and procedures are simple, clear, and presented in understandable language.
6: Maximum Readability and Comprehensibility	A variety of readability and plain-language guidelines are followed (e.g., sentence length and number of difficult words are kept to a minimum) to produce readable and comprehensible text.
7: Maximum Legibility	Characteristics that ensure easy decipherability are applied to text, to tables, figures, and illustrations, and to response formats.

Thompson, Johnstone, & Thurlow, 2002

Test developers, in line with current policy guidance, have focused attention to UD at the item level. The assumption is, if tests are comprised of universally designed and unbiased items, the tests are fair and accessible. Identifying test formats available to students of all abilities is a challenge and requires rethinking all stages of the development and administration process. This has therefore preoccupied test developers in recent years.

It is worth noting that ECD (Mislevy, Steinberg, & Almond, 1999) has provided a general theoretical framework for designing, developing, and administering assessments to result in reliable and valid information about student performance on tested constructs. The approach starts with an end result in mind – what scores are intended to mean. Claims set out through the ECD process link the intended uses of test scores to all layers of design, development, and administration processes. The tested content domain is analyzed and modeled, a conceptual framework is developed, and assessment implementation including delivery is envisioned and developed (Mislevy & Risconscente, 2005). The ECD framework threads the validity argument into each step of the assessment development process, connecting intention to outcome and serving as a vehicle for weaving in UD principles.

Application of UD to Alternate Assessments

Alternate assessments of alternate achievement standards (alternates), including the work of states and consortia (Dynamic Learning Maps [DLM], National Center and State Collaborative [NCSC] and its successor the Multi-State Alternate Assessment [MSAA]), have become increasingly sophisticated in recent decades. This is noteworthy given the relatively recent inclusion of the target population (i.e., students with significant cognitive disabilities) in consequential testing programs and state accountability systems. Alternates have gained in technical rigor, driving toward greater standardization and psychometric quality, garnering sophistication through technology solutions, and allowing for accessibility features and supports previously thought unavailable to students and test administrators. Alternates are now an established part of states' assessment and accountability programs. However, as a relatively new type of assessment driven by the specific needs of the target student population, alternates continue to challenge test developers to refine how programs meet students' academic and accessibility needs, while ensuring appropriate inferences can be made regarding what students know and can do.

Given efforts by consortia and state programs to incorporate accessibility concepts through UD, an earlier study examined alternate assessments to better understand how UD principles were being carried through development to implementation (Davidson, 2019). In this earlier study, findings showed that, since items developed for alternates have trended toward greater technical rigor under NCLB (2001) and Every Student Succeeds Act (2015), new item and task types and administration procedures were being considered for incorporation into alternate assessments. So, in addition to looking at research studies, we turned to the current field of alternates for the breadth of test designs and item/task types in play.

Publicly available test content (e.g., practice tests, sample items, released tests) for the 2018-2019 school year was sampled from two state consortia representing 23 states and four other educational agencies. Test content was also sampled from five additional states with their own alternate programs. The five individual state alternates were selected to represent different regions of the country as well as the range of students with disabilities and number of students with disabilities (National Center for Educational Outcomes, 2019). In sum, the review represented alternate assessments in 28 states and four other jurisdictions.

The study reviewed items and tasks selected from available practice tests, released items, and sample items in mathematics and English language arts or reading. Up to 20 items or tasks were reviewed per program and content area. Given the various assessment designs, additional grade levels were included either because of common content across grades or to increase the number of items or tasks reviewed.

A review protocol was developed to collect comparable evaluation data across different programs with various item types and design elements. The protocol focused on common characteristics of items around the key design principles, characteristics of the tests, and characteristics of how students interact with test content enabled by accessibility features to illustrate the range of alternates currently used (Appendix A).

Key findings from this review of research and released assessment content support the premise that design characteristics and delivery approaches play a significant and influential role in the relative accessibility of test content and underlying constructs in at least three ways.

First, different item types offer different advantages to the assessment overall. For example, items that required a single action to select a response were suggested by the literature to be more accessible to students. Another finding was that technology-enhancement provided opportunities to present more complex items more efficiently than multiple-choice format. Items that provided feedback to students may improve accessibility in some cases but not all (e.g., low-performing students may be frustrated; Johnstone, et al., 2013), and alternates that required that students and test administrators to work together during the testing experience improved accessibility.

Second, hybrid approaches to administration allowed for more responsiveness to student needs for communication, engagement, and response. The consortia programs allowed for hybrid approaches to delivery (both online and paper-and-pencil), whereas individual state programs were either online or paper-and-pencil.

Finally, the study found embedded accessibility features varied across programs, suggesting a wealth of possibilities for meeting the UD principles as applied to test content. The most common embedded accessibility features were formatting (color, font size, image size), text-to-speech, and masking/guides. It was unclear whether specific embedded accessibility features supported student needs. Embedded accessibility features should be evaluated carefully as little is yet known about their overall relative importance in student access for students with significant cognitive disabilities. Diversity in the population's needs must be better understood and accessibility features must be flexible enough to allow for the full range of necessary features and tools. Most important in the alternate design is that the assessment can be adjusted for students with varying needs of accommodation. Embedded features should serve that end and be studied to ensure they do not confound or add complexity to the testing experience that could create challenges or barriers for students.

This review of alternate item-level content yielded evidence of how program specifications influence accessibility of individual items. Findings also suggested that test administration protocols (e.g., incorporation of different types of items; hybrid approaches to delivery) can make assessments more accessible through high-quality items.

Design Versus Implementation

Most assessment programs claim to use UD, in tandem with ECD, to create integrated development plans as these assessments are built from the onset. For example, DLM, developer of a large-scale alternate assessment used by multiple states, used concepts from ECD (DeBarger, et al., 2011) in developing design patterns, development specifications, and task templates (Bechard, Romine, Karvonen, & Kingston, 2019). Including UD principles in the test development process, developers incorporated "allowances for the wide variety of supports students with significant cognitive disabilities need" (p. 8).

The application of UD, together with ECD, has made a difference. For example, DLM's technical documentation presented how vocabulary can be treated using a specified vocabulary list that "reflects the research in core vocabulary in augmentative and alternative communication and words that students must express to demonstrate mastery" (Bechard, et al., 2019, p. 8). The National Center and State Collaborative (2015) considered components of ECD to develop design patterns that "incorporated a variety of approaches to obtaining evidence of targeted knowledge or skills and supported development of task templates" (p. 12). Others have used related processes to construct innovative item types and explore human-computer interactions (Parshall & Harmes, 2009).

Despite these efforts and claims, though, Johnson, Trantham, and Usher-Tate (2019) found that not all assessment programs actually realize these promises to ensure the necessary steps are taken for optimal accessibility for all students. For example, they found in their sample of technical documents that some programs did not meet the expectation, such as using UD in test design and development, review of tests by a panel with subgroup expertise, or full disclosure of development or scoring protocols. Overall conclusions about actualization of accessibility promises are not always substantiated. "[C]autions about

using the test with specific subgroups due to lack of evidence, suggestions for making test instructions or format more accessible to all test takers, and recommendations for strengthening studies or methodologies used (p. 13)" must be considered.

Role of Policy Guidance

Therefore, while inspecting technical specification is critical for evaluating whether tests are maximally accessible, a limited focus on test items and protocols is insufficient. Policies drive decisions related to large-scale, consequential assessment programs. Largely influenced by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014), federal Peer Review has been a major vehicle for providing guidance for the validity arguments states use to build assessment programs. Since the early 2000s with No Child Left Behind (NCLB, 2001), the Peer Review process has evolved, adjusting to clarify technical advisors' expectations and the transparency of panel membership. Federal demands of states to build "valid and reliable" assessments have been coupled with funding levers to drive states toward adopting tests that produce scores that can answer substantive questions about school system performance. Most states have engaged in reviews of their assessments.

The Council of Chief State School Officers (CCSSO) has played a major role in the dissemination and application of best practices to states through influence. Their white paper, "Criteria for Procuring and Evaluating High-Quality Assessments" (2014), provides guidance to state and local leaders as they navigate the challenges of balancing costs and high-quality educational experiences for their students. These CCSSO criteria are cited widely to justify decision-making.

Findings from annual, consequential testing regimens over time show gaps in student performance persisting (or even growing) with varying degrees of intractability since NCLB began. Embedded within the discussion of fairness is the claim (or often assumption) that the test is accessible to all students in the test-taking population. These findings contributed to the impetus for a new chapter in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) that addresses fairness, touching on issues related to accessibility. However, to date, the concepts of fairness, including accessibility, have treated the claim of accessibility apart from or in limited relation to validity arguments in a limited national discussion focused on item-level bias evaluations.

Study Purpose

This led us to ask broadly how the policy guidance drives test development related to UD. Where are the opportunities to integrate UD elements into the work of building, administering, and monitoring state testing programs? Are there missed opportunities? Do definitions of accessibility need to change or stretch to meet the goal of full access for every child?

Therefore, based on themes in research literature related to accessibility claims during test design as well as the rapid developments in the most recent generation of alternate assessments, this study focused on evidence of accessibility features and processes in alternates. The study examined how alternate programs integrate both theoretical principles and empirical evidence related to accessibility for all students within a target population (Thurlow, Lazarus, Christensen, & Shyyan, 2016).

This evaluation study further examined the types of evidence needed to support accessibility claims between development approaches and resulting assessment features and processes in light of UD principles. We asked, *How do UD elements relate to influential policy guidance (i.e., Peer Review Critical Elements) intended to make assessments accessible?*

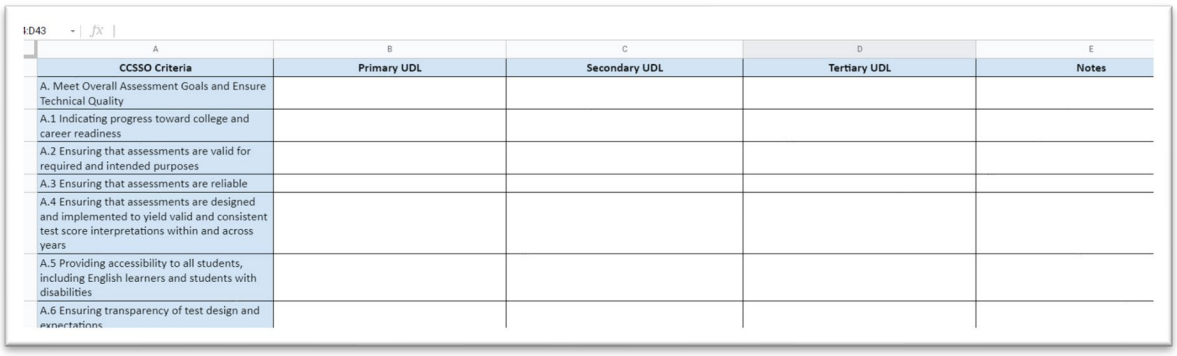
METHODS

To investigate our research question, we conducted a qualitative review of relationship between Peer Review Critical Elements and the UD elements using evaluation strategies (Patton, 2002). To triangulate

the review, we also looked at the relationship between the CCSSO Guidelines and the UD elements. The specific elements reviewed are included in Appendix B.

To illustrate the coding process, Figure 1 is an outtake of the rating data collection sheet for the Peer Review guidance.

FIGURE 1
ILLUSTRATION OF RATING SHEET



CCSSO Criteria	Primary UDL	Secondary UDL	Tertiary UDL	Notes
A. Meet Overall Assessment Goals and Ensure Technical Quality				
A.1 Indicating progress toward college and career readiness				
A.2 Ensuring that assessments are valid for required and intended purposes				
A.3 Ensuring that assessments are reliable				
A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years				
A.5 Providing accessibility to all students, including English learners and students with disabilities				
A.6 Ensuring transparency of test design and expectations				

Two veteran educators with extensive test development experience and expertise reviewed the policy guidance and aligned UD principles. Raters conducted the following steps:

1. Aligned each Peer Review Critical Element to UD elements (in prioritized order); and
2. Aligned each CCSSO Quality Element in light of UD elements (in prioritized order).

Raters independently reviewed each element, and then discussed differences in their ratings. They came to consensus about how to characterize each element's relationship to UD and recorded a consensus round. They also took notes about what they observed. Agreement rates ranged from 28% to 100% across the ratings. Raters reached consensus and calibrated their approach during the second round of ratings.

RESULTS

Results of the review of the policy guidance illustrated where expert raters identified UD elements within the articulated Peer Review and CCSSO elements. These results are summarized as "heat maps" using conditional shading to illustrate the relative focus of UD elements within the policy guidance (Tables 2-5).

Tables 2 and 4 report the unweighted counts, meaning the results do not reflect the raters' prioritization of the alignment ratings, for CCSSO results and Peer Review results, respectively. Tables 3 and 5 report the results after a weight was applied to reflect the prioritization, or degree, of alignment between UD elements and the policy element.

TABLE 2
CCSSO FINAL RATINGS BY CATEGORY (UNWEIGHTED)*

UDL	A	B	C	D	E	F	Total
1: Inclusive Assessment Population	2	0	1	1	0	1	5
2: Precisely Define Constructs	4	7	4	1	0	0	16
3: Accessible, Non-Biased Items	3	5	2	0	0	1	11
4: Amenable to Accommodations	1	1	1	0	0	1	4
5: Simple, Clear, and Intuitive Instructions and Procedures	1	1	1	0	0	0	3
6: Maximum Readability and Comprehensibility	0	2	0	0	0	0	2
7: Maximum Legibility	1	1	1	0	0	0	3
Total	12	17	10	2	0	3	44

* Counts are total, regardless of the hierarchy (primary, secondary, tertiary).

TABLE 3
CCSSO FINAL RATINGS BY CATEGORY (WEIGHTED)*

UDL	A	B	C	D	E	F	Total
1: Inclusive Assessment Population	2.00	0.00	0.75	1.00	0.00	1.00	4.75
2: Precisely Define Constructs	3.75	6.75	3.50	0.75	0.00	0.00	14.75
3: Accessible, Non-Biased Items	2.25	3.75	1.75	0.00	0.00	1.00	8.75
4: Amenable to Accommodations	0.50	0.50	0.50	0.00	0.00	0.75	2.25
5: Simple, Clear, and Intuitive Instructions and Procedures	1.00	0.75	0.75	0.00	0.00	0.00	2.50
6: Maximum Readability and Comprehensibility	0.00	2.00	0.00	0.00	0.00	0.00	2.00
7: Maximum Legibility	0.50	0.75	1.00	0.00	0.00	0.00	2.25
Total	10.00	14.50	8.25	1.75	0.00	2.75	37.25

* Counts are weighted: primary=1, secondary=0.75, tertiary=0.5.

TABLE 4
PEER REVIEW FINAL RATINGS BY CRITICAL ELEMENT (UNWEIGHTED)*

UDL	1	2	3	4	5	6	Total
1: Inclusive Assessment Population	5	1	1	1	3	2	13
2: Precisely Define Constructs	3	2	4	3	0	1	13
3: Accessible, Non-Biased Items	1	2	0	3	3	0	9
4: Amenable to Accommodations	0	3	0	3	4	0	10
5: Simple, Clear, and Intuitive Instructions and Procedures	0	2	0	0	0	0	2
6: Maximum Readability and Comprehensibility	0	0	0	0	0	0	0
7: Maximum Legibility	0	0	0	0	0	0	0
Total	9	10	5	10	10	3	47

* Counts are total, regardless of the hierarchy (primary, secondary, tertiary).

TABLE 5
PEER REVIEW FINAL RATINGS BY CRITICAL ELEMENT (WEIGHTED)*

UDL	1	2	3	4	5	6	Total
1: Inclusive Assessment Population	5.00	1.00	0.75	1.00	3.00	2.00	12.75
2: Precisely Define Constructs	2.25	1.75	4.00	2.75	0.00	0.75	11.50
3: Accessible, Non-Biased Items	0.50	1.25	0.00	2.75	2.25	0.00	6.75
4: Amenable to Accommodations	0.00	2.00	0.00	2.00	2.75	0.00	6.75
5: Simple, Clear, and Intuitive Instructions and Procedures	0.00	2.00	0.00	0.00	0.00	0.00	2.00
6: Maximum Readability and Comprehensibility	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7: Maximum Legibility	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	7.75	8.00	4.75	8.50	8.00	2.75	39.75

* Counts are weighted: primary=1, secondary=0.75, tertiary=0.5.

In addition to the summaries of ratings, raters noted questions and areas for further exploration. These notes were compiled. Four of the CCSSO Guidance elements and four of the Peer Review Critical Elements were identified as having the lowest rater agreement and the highest number of notes and questions from the raters. These elements were called out (Appendix B) for further exploration.

DISCUSSION AND CONCLUSION

We anticipated confirmation of the influence of UD principles applied in the test development process and the importance of stakeholder/user representation in all stages of development. Results raised questions about where accessibility can be considered in more ways during test development. Future

studies should explore how accessibility concepts and frameworks can inform test blueprints more directly.

Results showed a concentration of references to UD in the test development processes, consistent with findings from the content review of alternate assessments and their emphasis on item-level accessibility. However, raters saw connections with other aspects of assessment development (Peer Review Critical Elements 1-5; all CCSSO elements).

Of note is that two UD principles were not incorporated into the Peer Review Critical Elements, specifically #6 *Maximum Readability and Comprehensibility* and #7 *Maximum Legibility*. The lack of apparent expectation for these two UD elements within the Peer Review guidance could point to the lack of requirement in the *Standards for Educational and Psychological Testing* as defined by American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). The limited definition of fairness and a view that accessibility is only an issue at the item level may contribute to the lack of connection across the Peer Review requirements.

Through ratings and discussion, the raters identified where UD could be further applied, prioritized which UD elements could be integrated, identified which UD elements may need more attention, and identified areas where guidelines appear to not relate. In cases where the UD element was not aligned, they asked, *Is that reasonable? What other aspects of accessibility are necessary, where UD doesn't cover?*

For example, policy guidance could set expectations for representation in terms of who participates in each stage of the development process, including construct definition itself. To consider accessibility at the construct definition stage would not only demand changes to policy guidance, but also push at the UD definitions described by Thompson, Johnstone, and Thurlow (2002). Not only must constructs be precisely defined and communicated, constructs must reflect representative stakeholder engagement and input.

In addition, the results raised the question of how and where attention to opportunity to learn appears in the policy guidance and even in the UD principles. Without opportunity to learn, test content cannot be accessible to students. This fundamental issue stands at the heart of accessibility and deserves further attention.

REFERENCES

American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington DC: AERA.

Bechard, S., Clark, A., Swinburne Romine, R., Karvonen, M., Kingston, N., & Erickson, K. (2019). Use of Evidence-Centered Design to Develop Learning Maps-Based Assessments. *International Journal of Testing*, *19*(2), 188–205.

Bowe, F.G. (2000). *Universal design in education: Teaching nontraditional students*. Westport, CT: Bergin & Garvey.

Davidson, A.H. (2019). *Indiana's Alternate Measure (I AM): Item development and accessibility* [White Paper]. Marshall, MO: EdMetric LLC.

DeBarger, A.H., Seeratan, K., Cameto, R., Haertel, G., Knokey, A-M., & Morrison, K. (2011). *Alternate assessment design—mathematics. Implementing evidence-centered design to develop assessments for students with significant cognitive disabilities: Guidelines for creating design patterns and development specifications and exemplar task templates for mathematics* (Technical Report No. 9). Retrieved from http://alternateassessmentdesign.sri.com/techreports/AAD_M_TechRpt9_032911final.pdf

Dynamic Learning Maps Consortium. (2018). *Accessibility manual for the Dynamic Learning Maps alternate assessment, 2018-2019*. Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from https://dynamiclearningmaps.org/sites/default/files/documents/Manuals_Blueprints/Accessibility_Manual.pdf

Dynamic Learning Maps Consortium. (2016, June). *2014-2015 Technical Manual – Year-End*. Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from https://dynamiclearningmaps.org/sites/default/files/documents/publication/Technical_Manual_IM_2014-15.pdf

Dynamic Learning Maps Consortium. (2014a). *Summary of results from the fall 2013 pilot administration of the Dynamic Learning Maps™ Alternate Assessment System* (Technical Report #14-01). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from https://dynamiclearningmaps.org/sites/default/files/documents/publication/pilot_summary_of_findings.pdf

Eberhart, T. (2015). *A Comparison of multiple-choice and technology-enhanced item types administered on computer versus iPad*. (Unpublished dissertation.) Lawrence, KS: University of Kansas. Retrieved from https://kuscholarworks.ku.edu/bitstream/handle/1808/21674/Eberhart_ku_0099D_14325_DATA_1.pdf?sequence=1

Every Student Succeeds Act, Pub. 114-95, 114 U.S.C. (2015).

Hehir, T. (2002, Spring). Eliminating ableism in education. *Harvard Educational Review*, *72*(1), 1–32. Minneapolis, MN: University of Minnesota, National Center and State Collaborative. Retrieved from http://www.ncscpartners.org/Media/Default/PDFs/Resources/NCSCBrief6.pdf

Individuals with Disabilities Education Act, Pub. Law 94-142, 20 U.S.C. (2004).

Johnson, J.L., Trantham, P., & Usher-Tate, B.J. (2019). An evaluative framework for reviewing fairness standards and practices in educational tests. *Educational Measurement: Issues and Practices*, *38*(3), 6–19.

Johnstone, C., Figueroa, C., Yigal, A., Stone, E., & Laitusis, C. (2013). *Results of a cognitive interview study of immediate feedback and revision opportunities for students with disabilities in large scale assessments* (Synthesis Report 92). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from https://nceo.umn.edu/docs/OnlinePubs/Synthesis92/SynthesisReport92.pdf

Karvonen, M., Romine, R.S., & Clark, A.K. (2016). *Validity evidence to support alternate assessment score uses: Fidelity and response processes*. Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from: https://dynamiclearningmaps.org/sites/default/files/documents/publication/Validity_Evidence_AA_Score_Uses_NCME2016_Karvonen_Romine_Clark.pdf

Lazarus, S.S., Thurlow, M.L., R.R., Halpin, D., & Dillon, T. (2012). *Using cognitive labs to evaluate student experiences with the read aloud accommodation in math* (Technical Report). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Mislevy, R.J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (Technical Report). Retrieved from https://padi.sri.com/downloads/TR9_ECD.pdf

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (1999). *Evidence-centered assessment design.* Princeton, NJ: Educational Testing Service. Retrieved from https://pdfs.semanticscholar.org/ebef/dc95626cf8467473b333eda5526ecd570dd9.pdf

National Center and State Collaborative. (2019). *NCEO Data Analytics: Percent of Students with Disabilities by Disability Categories for 2005-06 to 2016-17 (#8)* (Online Datasource). Minneapolis, MN: National Center on Educational Outcomes at the University of Minnesota. Retrieved from https://nceo.info/

National Center and State Collaborative. (2016). *2015 Operational Assessment Technical Manual.* Minneapolis, MN: National Center on Educational Outcomes at the University of Minnesota. Retrieved from http://www.ncscpartners.org/Media/Default/PDFs/Resources/NCSC15_NCSC_TechnicalManualNarrative.pdf

National Center and State Collaborative. (2015). *Student interaction study: Exploring student interactions with and teacher perceptions of mathematics and reading Items.* (Technical Report). Minneapolis, MN: University of Minnesota, National Center and State Collaborative.
National Conference for Student Assessment. (2014, June). *So…what makes an innovative item an innovative item?* Los Angeles, CA: University of California. Retrieved from file:///C:/Users/MCG/Downloads/Innovative%20Items%20-%20full%20presentation%20(2).pdf
Nebelsick-Gullett, L., Towles-Reeves, E., Perkins A., & Deters, L. (2015). *Evaluating the quality and impact of items, products, and procedures: NCSC writing alternate assessment based on alternate achievement standards*. Minneapolis, MN: University of Minnesota, National Center and State Collaborative. Retrieved from http://www.ncscpartners.org/Media/Default/PDFs/Resources/AERA-NCME-2015/Evaluating%20the%20quality%20and%20impact%20of%20items.pdf
No Child Left Behind Act of 2001, Pub. L. No. 107−110, 1 U.S.C. (2002).
Parshall, C.G., & Harmes, J.C. (2009). Improving the quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology*, *10*(1). Retrieved from http://jattjournal.com/index.php/atp/article/view/48349/39219
Patton, M.P. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.
Rose, D.H., Meyer, A., & Hitchcock, C. (2005). *The universally designed classroom: Accessible curriculum and digital technologies*. Cambridge, MA: Harvard Education Press.
Stanford Center for Assessment, Learning, & Equity. (2016). *Evaluating item quality in large- scale assessments, phase I report of the study of state assessment systems*. Stanford, CA: Author. Retrieved from https://scale.stanford.edu/sites/default/files/Evaluating%20Item%20Quality%20in%20Large-Scale%20Assessments_v19.pdf
Thompson, S.J., Johnstone, C.J., & Thurlow, M.L. (2002, June). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
Thurlow, M.L., Lazarus, S.S., Christensen, L.L., & Shyyan, V. (2016). *Principles and characteristics of inclusive assessment systems in a changing assessment landscape* (NCEO Report 400). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from https://nceo.umn.edu/docs/OnlinePubs/Report400/NCEOReport400.pdf

APPENDICES

Appendix A – Item-Level Review Protocol

TABLE A1
INITIAL ALTERNATE ASSESSMENT REVIEW PROTOCOL

1) **Stimulus**: Does the **stimulus** contain the following? (Mark "x" all that apply)
Data/Data representation
Text
Image/Picture
Sound/Recording
Other (specify in notes below)

Notes/Questions
2) **Answer Options**: What type of **answer options** does the item demand? (Mark "x" all that apply)
Right/wrong options only
Open-ended response
Single response required
Multiple responses required
Response allows for partial credit scoring
3) **Number of answer options (#)**
4) **Choice Reduction**: Does the item direct the TA to **reduce the # of answer options**? (Y, N)
5) **Item Demands**: What **verb(s)** describes what the item demands the student to do in response? (Mark all that apply)
indicate (e.g., mark, click, circle, tap, note, name, select, point, find)
manipulate (e.g., place, move, match, order, sort, reduce)
generate (e.g., write, draw, table, graph, calculate, represent, speak)
6) **Stimulus Presentation**: By what means is the stimulus presented?
Online
On paper
7) **Response Collection**: By what means is the item response collected?
Online
On paper
8) **Role**: Who inputs the student response?
Student
Test Administrator
9) **Embedded Accessibility Features**: Check accessibility functions available to the student without additional support from TA.
Calculator (Math only)
Glossary – Text
Glossary – Pictorial

Translation (including ASL)
Text-to-speech/read aloud
Speech-to-text
Formats to color
Formats to font size
Formats to image size
Masking
Line guides
Other (specify in notes below)
Notes/Questions

Appendix B – Policy Guidance

TABLE B1
PEER REVIEW CRITICAL ELEMENTS
(UNITED STATES DEPARTMENT OF EDUCATION, 2018)

Critical Element 1.1 – State Adoption of Academic Content Standards for All Students
Critical Element 1.2 – Challenging Academic Content Standards
Critical Element 1.3 – Required Assessments
Critical Element 1.4 – Policies for Including All Students in Assessments
Critical Element 1.5 – Meaningful Consultation in the Development of Challenging State Standards and Assessments
Critical Element 2.1 – Test Design and Development
Critical Element 2.2 – Item Development
Critical Element 2.3 – Test Administration
Critical Element 2.4 – Monitoring Test Administration
Critical Element 2.5 – Test Security
Critical Element 2.6 – Systems for Protecting Data Integrity and Privacy
Critical Element 3.1 – Overall Validity, Including Validity Based on Content
Critical Element 3.2 – Validity Based on Cognitive Processes
Critical Element 3.3 – Validity Based on Internal Structure
Critical Element 3.4 – Validity Based on Relations to Other Variables
Critical Element 4.1 – Reliability
Critical Element 4.2 – Fairness and Accessibility
Critical Element 4.3 – Full Performance Continuum
Critical Element 4.4 – Scoring

Critical Element 4.5 – Multiple Assessment Forms
Critical Element 4.6 – Multiple Versions of an Assessment
Critical Element 4.7 – Technical Analysis and Ongoing Maintenance
Critical Element 5.1 – Procedures for Including Students with Disabilities
Critical Element 5.2 – Procedures for Including English Learners in Academic Content Assessments
Critical Element 5.3 – Accommodations
Critical Element 5.4 – Monitoring Test Administration for Special Populations
Critical Element 6.1 – State Adoption of Academic Achievement Standards for All Students
Critical Element 6.2 – Achievement Standards Setting
Critical Element 6.3 – Challenging and Aligned Academic Achievement Standards
Critical Element 6.4 – Reporting

TABLE B2
CCSSO CRITERIA (CCSSO, 2014)

A. Meet Overall Assessment Goals and Ensure Technical Quality
A.1 Indicating progress toward college and career readiness
A.2 Ensuring that assessments are valid for required and intended purposes
A.3 Ensuring that assessments are reliable
A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years
A.5 Providing accessibility to all students, including English learners and students with disabilities
A.6 Ensuring transparency of test design and expectations
A.7 Meeting all requirements for data privacy and ownership
B. Align to Standards – English Language Arts/Literacy
B.1 Assessing student reading and writing achievement in both ELA and literacy
B.2 Focusing on complexity of texts
B.3 Requiring students to read closely and use evidence from texts
B.4 Requiring a range of cognitive demand
B.5 Assessing writing
B.6 Emphasizing vocabulary and language skills
B.7 Assessing research and inquiry
B.8 Assessing speaking and listening
B.9 Ensuring high-quality items and a variety of item types
C. Align to Standards – Mathematics
C.1 Focusing strongly on the content most needed for success in later mathematics
C.2 Assessing a balance of concepts, procedures, and applications
C.3 Connecting practice to content
C.4 Requiring a range of cognitive demand
C.5 Ensuring high-quality items and a variety of item types

D. Yield Valuable Reports on Student Progress and Performance
D.1 Focusing on student achievement and progress to readiness
D.2 Providing timely data that inform instruction
E. Adhere to Best Practices in Test Administration
E.1 Maintaining necessary standardization and ensuring test security
F. State Specific Criteria (as desired)
Sample criteria might include
• Requiring involvement of the state's K-12 educators and institutions of higher education
• Procuring a system of aligned assessments, including diagnostic and interim assessments
• Ensuring interoperability of computer-administered items

TABLE B3
PEER REVIEW CRITICAL ELEMENTS FOR FURTHER CONSIDERATION

PEER REVIEW	Primary	Secondary	Tertiary
Critical Element 2.4 – Monitoring Test Administration	5: Simple, Clear, and Intuitive Instructions and Procedures	4: Amenable to Accommodations	No Applicable Element
Critical Element 4.4 – Scoring	No Applicable Element	No Applicable Element	No Applicable Element
Critical Element 4.5 – Multiple Assessment Forms	3: Accessible, Non-Biased Items	4: Amenable to Accommodations	No Applicable Element
Critical Element 4.6 – Multiple Versions of an Assessment	3: Accessible, Non-Biased Items	4: Amenable to Accommodations	No Applicable Element

TABLE B4
CCSSO GUIDANCE ELEMENTS FOR FURTHER CONSIDERATION

CCSSO	Primary	Secondary	Tertiary
C.4 Requiring a range of cognitive demand	7: Maximum Legibility	3: Accessible, Non-Biased Items	2: Precisely Define Constructs
C.5 Ensuring high-quality items and a variety of item types	3: Accessible, Non-Biased Items	5: Simple, Clear, and Intuitive Instructions and Procedures	4: Amenable to Accommodations
F • Requiring involvement of the state's K-12 educators and institutions of higher education	1: Inclusive Assessment Population	No Applicable Element	No Applicable Element
F • Procuring a system of aligned assessments, including diagnostic and interim assessments	No Applicable Element	No Applicable Element	No Applicable Element