Hindawi Publishing Corporation
Education Research International
Volume 2013, Article ID 952704, 12 pages
http://dx.doi.org/10.1155/2013/952704

Research Article

Test Accessibility: Item Reviews and Lessons Learned from Four State Assessments

Peter A. Beddow,[1] Stephen N. Elliott,[2] and Ryan J. Kettler[3]

[1] *Lipscomb University, Nashville, TN 37204, USA*
[2] *Arizona State University, Tempe, AZ 85287-2111, USA*
[3] *Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*

Correspondence should be addressed to Peter A. Beddow; peterbeddow@gmail.com

Received 17 February 2013; Revised 24 April 2013; Accepted 27 April 2013

Academic Editor: Huy P. Phan

Copyright © 2013 Peter A. Beddow et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The push toward universally designed assessments has influenced several states to modify items from their general achievement tests to improve their accessibility for all test takers. The current study involved the review of 159 items used by one state across four content areas including science, coupled with the review of 261 science items in three other states. The item reviews were conducted using the *Accessibility Rating Matrix* (Beddow et al. 2009), a tool for systematically identifying access barriers in test items, and for facilitating the subsequent modification process. The design allowed for within-state comparisons across several variables for one state and for within-content area (i.e., science) comparisons across states. Findings indicated that few items were optimally accessible and ratings were consistent across content areas, states, grade bands, and item types. Suggestions for modifying items are discussed and recommendations are offered to guide the development of optimally accessible test items.

1. Introduction

Access is a dominant concern in the pursuit of developing inclusive assessments for students with a broad range of abilities and needs. The push toward universally-designed assessments gained particular prominence following the passage of legislation that permitted a portion of students to participate in alternate assessments based on modified academic achievement standards (AA-MASs), prior to which access barriers were addressed primarily by the use of testing accommodations. Such accommodations are typically defined as changes in the administration procedures of a test to address the special needs of individual test takers [1]. With changes to the NCLB Act in 2007 [2, 3], test developers began to examine tests and items with the goal of modifying them to reduce the influence of intrinsic access barriers on subsequent test scores for a small group of students with disabilities, to increase test score validity for the population of students for whom standardized tests historically have posed difficulty.

This item modification process has been guided by accessibility theory [4]. To wit, accessibility—defined as the degree to which a test and its constituent item set permit the test taker to demonstrate his or her knowledge of the target construct—is conceptualized as the sum of interactions between features of the test and individual test taker characteristics (see Figure 1.) The validity of test score inferences is dependent on the accessibility of the test for the entirety of the target test taker population. To the extent a test contains access barriers for a portion of the tested population, inferences made from test scores of those individuals may be invalid; as well, the validity of subsequent norming procedures or comparisons across the population may be reduced. This paper represents the first comparison of the results of accessibility reviews of test items from several state achievement tests using accessibility theory.

Accessibility theory disaggregates test taker access skills or characteristics into five categories: physical, sensory/perceptive, receptive, emotive, and cognitive. In Figure 1, the left-hand column of the test event consists of these categories. Each of these categories is loosely paired with one or more of the test or item feature categories in the right-hand column, which indicate aspects of the test that can be adjusted

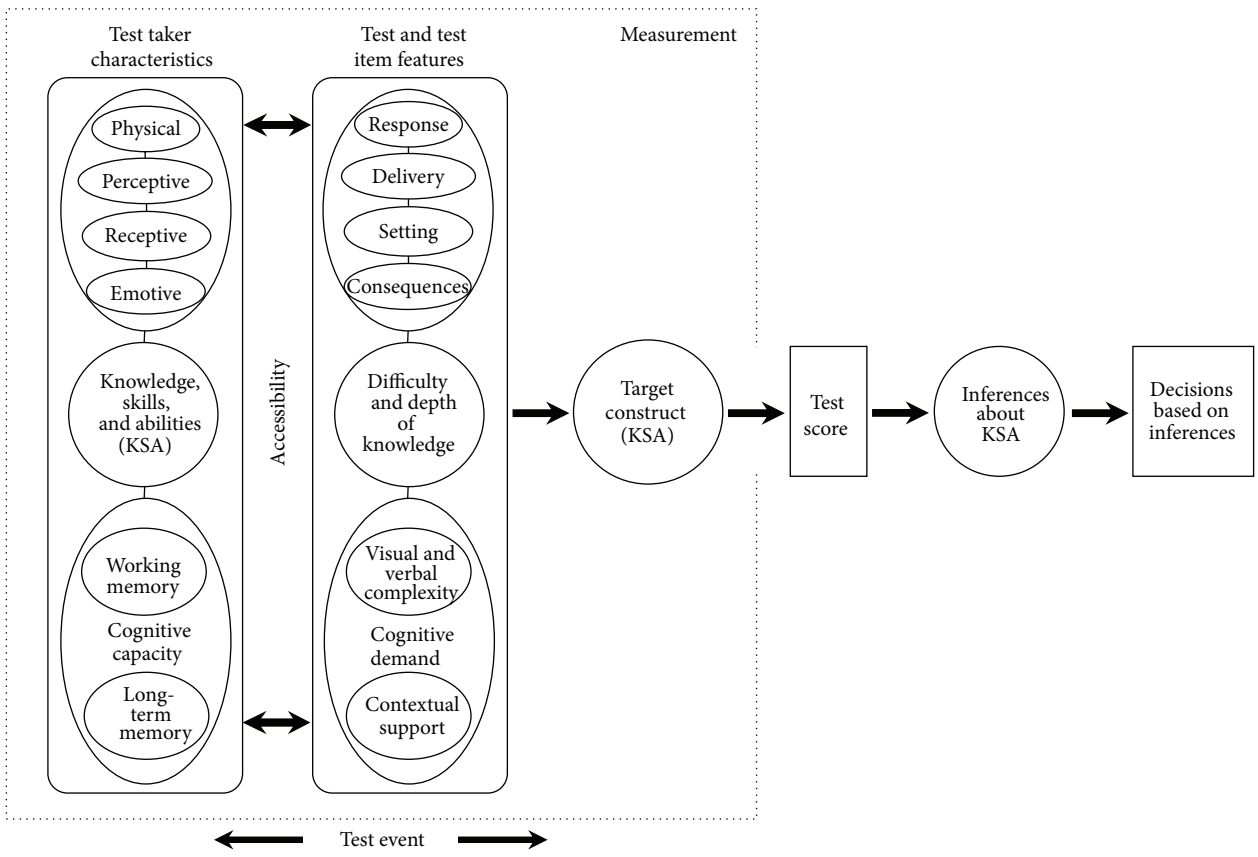


Figure 1: Accessibility theory ([4], used with permission).

to accommodate particular access needs and thus reduce the influence of ancillary interactions on subsequent test scores [4]. The concept is similar to the interactivity model of universally-designed assessments proposed by Ketterlin-Geller [5] and contains elements from the guidelines for universally-designed learning environments [6].

Physical access skills consist of the physical requirements to engage in the test event, such as entering the test room or using a pencil to fill bubbles on an answer document. Perceptive or sensory skills include seeing the test items or hearing spoken directions or prompts. Receptive skills involve the processing of perceived stimuli for comprehension (e.g., reading or listening). Emotions, while not necessarily involving abilities per se, include motivation and attitude (e.g., test self-efficacy or, alternatively, test anxiety). Indeed, it should be noted that research has indicated stress can negatively impact working memory, which arguably is chief among essential test-taking skills [7].

The fifth test taker access category involves cognitive capacity, which consists of long-term memory and working memory (sometimes referred to as short-term memory.) It should be noted that we deliberately refrain from referring to cognition as a skill, since there is some debate whether cognitive capacity can be taught or learned; indeed, while longitudinal data indicate working memory span of high-performing learners increases to a greater degree over time than that of low-performing learners, it is unclear how these differences are causally related [8].

While the legislative push to develop AA-MASs for students identified with disabilities has driven the examination of accessibility as critical to test score validity, the resulting theory and evidence supporting its importance has indicated accessibility affects student test performance across the range of the test taker population. Indeed, the differential boost observed across the testing accommodations literature as noted in 2005 by Sireci et al. [9] is similarly evident in the results of research on item modifications [10]. This suggests accessibility affects test performance for students identified with disabilities to a greater degree than for their general education peers, but that some general education students may benefit from accessibility improvements as well. This phenomenon is likely a function of test taker characteristics that may be independent of disability status, such as reading fluency, test anxiety, and adaptability to high cognitive demand (see [11]). Research by Laitusis [12] suggests several factors that influence comprehension (e.g., reading fluency) and likely contribute to poor test performance on a range of assessment types, particularly for students identified

with disabilities. Recent guidelines released by the National Accessibility Reading Assessment Projects (NARAP; [13]) underscore the need for the development of assessments that isolate target constructs and reduce the need for accommodations without precluding their use for some test takers.

Results of experimental research suggest a relation exists between accessibility and student achievement. Elliott et al. [14] conducted a large-scale study with 755 eighth-grade students in four states and found significantly higher scores for students on test forms with modified items compared to forms that consisted of unmodified items. Modifications included reducing item word count, increasing space within items, and adding visuals. These differences were observed on both reading and mathematics tests for all the three groups who participated in the study: (a) students identified with disabilities who would be eligible for an AA-MAS ($n = 250$), students identified with disabilities who would not be eligible ($n = 236$), and (c) students not identified with disabilities ($n = 269$). The modification effects (original versus modified) for reading and mathematics were .38 standard deviations and .21 standard deviations, respectively. Additionally, preliminary results of a small-scale pilot study of mathematics and reading items in grades 7 and 10 indicated moderate positive correlations for ARM accessibility ratings with both item discrimination and item difficulty [15]. Further, the ARM has been used by several professional organizations to guide teams of assessment developers and to train individual item writers. The current evaluation study was conducted to provide systematic evidence of the accessibility of state achievement tests across four content areas and across four states. It provides a framework for addressing and improving the accessibility of items for AA-MASs and other forms of achievement tests. If data from the current study indicate the accessibility of the sampled achievement tests is less than optimal—as we anticipate based on the assumptions underlying the legislation permitting the design of alternate assessments with improved accessibility—we would argue that the process of accessibility review should be undertaken as a standard part of the development process of achievement tests.

2. Method

2.1. Sample. A representative sample of test items from the achievement tests of four states was reviewed: State A is in the midwest, States B and C are both northern plains states, and State D is a coastal southern state. The participating states were part of two federal grant projects: the Consortium for Modified Alternate Assessment Development and Implementation (CMAADI (CMAADI was a U.S. Department of Education General Supervision Enhancement grant codirected by Stephen N. Elliott, Michael C. Rodriguez, Andrew T. Roach, and Ryan J. Kettler; several studies on item modification were conducted within this multistate project during 2007–2010)), and Operationalizing Alternate Assessment of Science Inquiry Standards (OAASIS(OAASIS was a U.S. Department of Education Enhanced Assessment grant directed initially by Courtney Foster and ultimately by John Payne; Several studies on item modification were conducted within this multi-state project during 2008–2011)). The first of these reviews was conducted on a set of multiple-choice and constructed-response items from State A across grades 3–8 from the English language arts, mathematics, science, and Social Studies content domains ($N = 159$ items). The second set of reviews was conducted on science inquiry items from States B, C, and D across grades 4, 5, 8, and 11 ($N = 261$ items). The four states were the only states approached to participate in the current study, based on their participation in the aforementioned large-scale assessment research projects.

2.2. Materials. The *Accessibility Rating Matrix* (ARM; [16, 17]; http://peabody.vanderbilt.edu/tami.xml) is a noncommercial research tool for evaluating and modifying tests and items with a focus on reducing the influence of ancillary interactions during the test event due to unnecessary complexity in text and visuals, poor organization and/or item layout, and other item and test features that may limit access for some test takers. The ARM consists of a set of rating rubrics to guide the analysis of test items to yield an accessibility rating that reflects the degree to which the item is likely to be accessible for the entirety of the test taker population, on a 4-point scale (4 = *maximally accessible for nearly all test takers*; 3 = *maximally accessible for most test takers*; 2 = *maximally accessible for some test takers*; and 1 = *inaccessible for many test takers*). Table 1 contains heuristics for each of these accessibility levels based on the approximate percentage of the target population for whom a test or test item is optimally accessible (i.e., the test or test item permits the test taker to demonstrate his or her knowledge of the target construct). It should be noted that since the construct of accessibility heretofore has been unmeasured, these percentages were established to guide the process of item evaluation by a team of experts in assessment and are based on raters' assumptions of the effect of access barriers on test taker scores, not on research data. The ARM is organized into the five basic elements common to most multiple-choice test items: (a) item passage and/or stimulus; (b) item stem; (c) visual(s), including graphs, charts, tables, and figures; (d) answer choices; and (e) page and/or item layout. Figure 2 contains a visual representation of these item elements.

2.3. Procedures. Raters consisted of five researchers with test development experience, all of whom had been trained extensively in the use of the ARM. Specifically, 3 of the 5 researchers had directed or co-directed multi-state federal grants on assessment development and evaluation, with specific focus on universal design and accessibility. The other two were coauthors of the ARM with extensive experience in item rating and evaluation. The senior author of the instrument assigned one-fourth of each state's item sample to each of the four researchers and himself. It should be noted that for the three-state review, the passage/item stimulus element was disaggregated to generate two distinct ratings.

2.4. Reliability of the Process. Several steps were taken to ensure the reliability of the review process and the validity of the results. First, states were instructed to provide items and data with the following criteria in mind: (a) to support the external validity of results to the universal population of state test items, states were asked to provide a representative sample of their test items, as opposed to selecting their "best" or "worst" items in terms of accessibility; (b) the authors assumed the greater the amount of information available about each item, the greater the likelihood raters would accurately isolate the target construct of the item and identify features that may present access barriers for some test takers; therefore, states were asked to include descriptive and psychometric data for each item, including the target construct, performance indicator, or strand the item was designed to measure, depth of knowledge, difficulty, discrimination, distractor functioning, and response selection frequency. When possible, states were asked to disaggregate psychometric data by disability status or performance group. To ensure reliability of the process, item review procedures were designed to mirror the collaborative approach used by item modification teams across several states. To wit, agreement for an item was operationalized by agreement within one accessibility level on the overall analysis rating (e.g., adjacent overall agreement). When this criterion was not met, the team of raters collaborated until consensus was reached. The final ratings for items on which raters did not have exact agreement were determined by the lower of the two ratings for each of the item elements and the overall accessibility rating for the item. According to this criterion, agreement between raters pairs ranged from 87% to 100% for the four accessibility reviews, with a mean of 94%.

Table 1: Test item accessibility levels.

Level	Description	Heuristic
4	Maximally accessible for nearly all test takers	Optimally accessible for 95–99% of the population
3	Maximally accessible for most test takers	Optimally accessible for 90–95% of the population
2	Maximally accessible for some test takers	Optimally accessible for 85–90% of the population
1	Inaccessible for many test takers	Optimally accessible for less than 85% of test takers

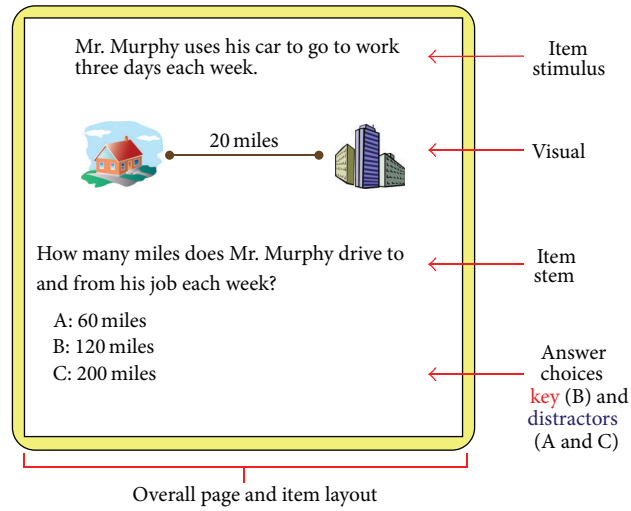


Figure 2: Anatomy of a multiple-choice item ([16, 17], used with permission).

3. Results

Table 2 contains the results of the State A item review for the CMAADI project, organized into the item elements as defined in the ARM [16, 17]. Across the reviewed items, the means of various item elements ranged from 2.8 (visuals) to 3.4 (answer choices), with mean accessibility rating for overall items of 2.8. Of the 159 items, 13% received a rating of 4 (optimally accessible; $n = 20$). An additional 58% of the items were rated 3 (*maximally accessible for most test takers*). The evaluation team identified several positive attributes of the reviewed set of items, including embedding items in related passages to reduce cognitive demand and memory load, the use of positively worded item stems, sufficient use of white space, and the use of three answer choices. The evaluation team also suggested a number of modifications to improve the accessibility of the items, including simplifying and decreasing the length of item stimuli and stems, eliminating unnecessary visuals, and reducing the spread of information required for responding to the items across pages. Table 3 contains the percentages of each suggested modification by content area.

Tables 4 and 5 contain the combined results of the four state item reviews in the domain of science. Across states, the means of various item elements ranged from 2.5 to 3.7. Overall accessibility ratings ranged from 2.5 to 2.8. For the total sample, only 2% of items were rated as optimally accessible ($n = 6$), while 62% received a rating of 3 (*maximally accessible for most test takers*). Summary data for each state indicated there were a variety of strengths and weaknesses in the items sampled across states. For State A ($n = 26$ science items in grades 4 and 6), strengths included clearly written item stems that used the active voice, and few items with negative stems. The majority of item visuals clearly depicted the intended content and were embedded in the item layout to minimize the demand for representational holding. Nearly all items contained all elements on a single page, and most items used large font sizes. Common suggestions for improving access for all students included simplifying the text of item stimuli and stems, eliminating unnecessary or implausible distractors, attending to balance issues in the answer choices, and reducing the length of answer choices. Of the 26 reviewed items, only 21 received overall ratings due to the items being in unfinished form. Of these items, 77% received an overall rating of 3 or higher.

For State B ($n = 60$ end-of-course science items), the evaluation team identified several positive attributes of the items, including clearly written item stems and answer choices, while suggesting a number of modifications to improve the accessibility of the items, including simplifying

Table 2: Accessibility ratings by grade, content area, and item type for state A.

Content area/grade	# (% of sample)	Item analysis rubric ratings					Overall analysis rubric ratings
		Passage/item stimulus	Item stem	Visuals	Answer choices	Page/item layout	
		M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Across	159 (100%)	3.1 (0.7)	3.3 (0.7)	2.8 (0.9)	3.4 (0.8)	3.3 (0.8)	2.8 (0.7)
Elementary	88 (%)	3.2 (0.7)	3.3 (0.7)	2.8 (0.8)	3.4 (0.7)	3.3 (0.8)	2.9 (0.7)
Middle	71 (%)	3.0 (0.6)	3.2 (0.7)	2.9 (0.9)	3.3 (0.9)	3.4 (0.8)	2.8 (0.7)
Language Arts	67 (42%)	3.0 (0.7)	3.2 (0.7)	2.4 (0.6)	3.5 (0.7)	3.1 (0.8)	2.8 (0.6)
Elementary	44 (66%)	3.0 (0.7)	3.2 (0.7)	2.3 (0.6)	3.4 (0.7)	3.1 (0.8)	2.8 (0.7)
Middle	23 (34%)	3.1 (0.8)	3.3 (0.8)	2.6 (0.7)	3.7 (0.6)	3.1 (0.8)	2.9 (0.5)
Mathematics	50 (31%)	3.2 (0.6)	3.4 (0.7)	3.4 (0.8)	3.2 (0.9)	3.4 (0.7)	2.8 (0.7)
Elementary	19 (38%)	3.6 (0.5)	3.6 (0.6)	3.6 (0.7)	3.4 (0.6)	3.4 (0.7)	3.1 (0.3)
Middle	31 (62%)	3.0 (0.5)	3.3 (0.7)	3.2 (0.9)	3.0 (1.0)	3.4 (0.8)	2.6 (0.8)
Science	21 (13%)	2.9 (0.8)	3.0 (0.8)	3.7 (0.5)	3.1 (0.8)	3.6 (0.7)	2.8 (0.6)
Elementary	11 (52%)	3.0 (0.9)	2.9 (1.0)	3.7 (0.5)	3.2 (0.8)	3.5 (0.7)	2.7 (0.6)
Middle	10 (48%)	2.9 (0.7)	3.0 (0.7)	—	3.0 (0.7)	3.7 (0.7)	2.9 (0.6)
Social studies	21 (13%)	3.4 (0.7)	3.2 (0.6)	2.5 (0.0)	3.8 (0.6)	3.6 (0.5)	2.9 (0.8)
Elementary	14 (67%)	3.6 (0.5)	3.4 (0.5)	4.0 (0.0)	3.8 (0.6)	3.6 (0.5)	2.9 (0.8)
Middle	7 (33%)	2.0 (0.0)	3.0 (0.8)	1.0 (0.0)	3.8 (0.4)	3.6 (0.5)	2.7 (0.8)
Item type							
Constructed response	39 (25%)	2.9 (0.7)	3.0 (0.7)	2.5 (0.9)	—	3.1 (0.7)	2.7 (0.6)
Multiple choice	120 (75%)	3.2 (0.7)	3.3 (0.7)	3.0 (0.9)	3.4 (0.8)	3.4 (0.8)	2.9 (0.7)

Accessibility ratings should be interpreted as follows:
4 = maximally accessible for nearly all test takers; 3 = maximally accessible for most test takers;
2 = maximally accessible for some test takers; 1 = inaccessible for many test takers.

the item layout, changing the formatting of item text to distinguish item stimuli from item stems, simplifying visuals, attending to items with multiple plausible correct responses, and eliminating unnecessary answer choices. Of the 60 reviewed items, 68% of the items received overall accessibility ratings of 3 or higher.

For State C (n = 101 science items across grades 5, 8, and 11), positive attributes of the sampled items included high mean ratings for all of the item elements, large and readable font sizes, clear visuals, and the lack of demand for turning the page to access information necessary for responding. The team recommended a number of modifications to improve the accessibility of the State C items, including simplifying the item layout, changing the formatting and organization of item text to distinguish item stimuli from item stems, simplifying or eliminating visuals, and attending to items with multiple plausible correct responses. Of the 101 items, 73% were rated 3 or higher.

For State D (n = 100 items across grades 4, 8, and 11), the team identified several positive attributes of the reviewed set of items, including clearly written and positively worded item stimuli and item stems, and the use of sufficient white space and large font sizes. The team also suggested simplifying common item stimuli and visuals, attending to items with multiple plausible correct responses, eliminating unnecessary answer choices, and reducing the spread of item elements across pages. Of the 100 items in the review sample, no items were rated as optimally accessible, while half of the sample received a rating of 3.

Table 6 contains a tabulation of the percentages of items at each accessibility level for each state, with State A's data disaggregated by content area. For States A, B, and C, approximately two-thirds of the items received ratings of 3, or *maximally accessible for most test takers*, and approximately one-quarter of the items received ratings of 2, or *maximally accessible for some test takers*. Ratings for State D were equally divided between levels 2 and 3. Across states, comparatively few items received the highest and lowest accessibility ratings.

4. Discussion

Access to the construct being measured with a test item is fundamental to a claim of a valid score. The present study reported research on test item accessibility for a sample of achievement test items from four states involved in federally funded projects focusing on the development of modified alternate assessments. These assessments are for a rather small group of students identified with disabilities who are receiving grade level instruction, but are unlikely to progress at a rate that will allow them to be academically proficient by the end of the school year.

The analysis of a representative sample of items from existing state achievement tests was guided by the ARM [16, 17]. Item ratings and modification suggestions were

Table 3: Suggested modifications for items from State A, percentages by content area.

Element	Language arts ($N = 67$)	Mathematics ($N = 50$)	Social studies ($N = 21$)	Science ($N = 21$)	Across Content areas ($N = 159$)
Passage/stimulus					
Add a passage or stimulus	0%	0%	13%	0%	1%
Simplify/shorten text	75%	62%	25%	42%	63%
Change text formatting (bold, etc.)	30%	16%	0%	11%	21%
Item stem					
Simplify/shorten stem	45%	45%	36%	48%	44%
Clarify question or directive	15%	24%	41%	38%	24%
Use active voice	1%	0%	0%	5%	1%
Eliminate negative stem	0%	2%	0%	14%	2%
Change text formatting (bold, etc.)	10%	29%	14%	24%	19%
Visuals					
Add a visual(s)	0%	10%	18%	8%	7%
Eliminate visual(s)	94%	0%	0%	0%	49%
Simplify visual(s)	4%	25%	50%	10%	13%
Move visual(s)	0%	0%	33%	11%	2%
Answer choices					
Simplify/shorten text	14%	9%	6%	15%	9%
Revise answer choices	10%	12%	26%	19%	26%
Eliminate distractor(s)	24%	54%	6%	60%	36%
Change the order of choices	16%	12%	6%	0%	10%
Balance issues	27%	20%	24%	20%	23%
Rationale can be made for more than one correct response	0%	6%	0%	5%	3%
Page/item layout					
Embed item in passage	45%	0%	0%	0%	22%
Increase white space	1%	35%	32%	24%	19%
Change the size of item elements	0%	4%	5%	0%	2%
Change the font size	6%	0%	5%	16%	6%
Reduce spread of information across pages/screens	70%	14%	18%	4%	36%

Table 4: Accessibility ratings for science items by state.

Grade	Number of items	Item analysis rubric ratings						
		Passage M (SD)	Item stimulus M (SD)	Item stem M (SD)	Visuals M (SD)	Answer choices M (SD)	Page/Item layout M (SD)	Overall analysis rubric ratings M (SD)
State A	21	—	2.9 (0.8)	3.0 (0.8)	3.7 (0.5)	3.1 (0.8)	3.6 (0.7)	2.8 (0.6)
State B	60	3.4 (0.5)	3.0 (0.7)	3.2 (0.7)	3.0 (0.9)	2.8 (0.8)	3.5 (0.6)	2.7 (0.7)
State C	101	3.4 (0.6)	3.0 (0.6)	3.1 (0.7)	3.0 (1.1)	3.0 (0.7)	3.3 (0.6)	2.7 (0.5)
State D	100	2.5 (0.6)	3.6 (0.6)	3.0 (0.7)	2.8 (1.0)	2.7 (0.7)	3.0 (0.7)	2.5 (0.5)

Accessibility ratings should be interpreted as follows: 4 = maximally accessible for nearly all test takers; 3 = maximally accessible for most test takers; 2 = maximally accessible for some test takers; 1 = inaccessible for many test takers.

remarkably similar across the major categorical variables. In State A, average item accessibility ratings were within .2 on a 4-point scale across content area (2.9 for social studies and 2.8 for the other three content areas), grade band (2.9 for middle school versus 2.8 for elementary school), and item type (2.9 for multiple choice and 2.7 for constructed response). Across states, mean overall item accessibility ratings were within .30 on the same 4-point scale in science (State D = 2.50; State B = 2.80).

Across states and content areas, the most recommended modifications were also quite similar. In all states, the most common recommendation to improve a passage or stimulus

Table 5: Suggested modifications for science items, percentages by state.

Element/grades	State A ($N = 21$) 4, 6	State B ($N = 60$) 11	State C ($N = 101$) 5, 8, 11	State D ($N = 100$) 4, 8, 11
Passage*				
Add a passage or stimulus	—	0%	0%	0%
Eliminate passage or stimulus*	—	0%	7%	2%
Simplify/shorten text	—	57%	43%	91%
Reorganize information*	—	0%	7%	22%
Modify the directions*	—	0%	0%	0%
Change text formatting (bold, etc.)	—	0%	0%	10%
Stimulus*				
Add a passage or stimulus	0%	4%	0%	0%
Eliminate passage or stimulus*	—	18%	18%	8%
Simplify/shorten text	42%	68%	67%	33%
Reorganize information*	—	0%	5%	0%
Modify the directions*	—	0%	0%	0%
Change text formatting (bold, etc.)	11%	0%	0%	3%
Item stem				
Simplify/shorten stem	48%	42%	54%	69%
Clarify question or directive	38%	23%	23%	20%
Change stem to a question*	—	8%	6%	0%
Use active voice	5%	0%	3%	0%
Eliminate negative stem	14%	0%	0%	2%
Change text formatting (bold, etc.)	24%	13%	16%	6%
Visuals				
Add a visual(s)	8%	3%	6%	0%
Eliminate visual(s)	0%	0%	12%	23%
Simplify visual(s)	10%	46%	35%	36%
Move visual(s)	11%	8%	6%	18%
Answer choices				
Simplify/shorten text	15%	17%	15%	24%
Revise answer choices	19%	12%	19%	26%
Eliminate distractor(s)	60%	58%	64%	74%
Change the order of choices	0%	0%	0%	0%
Balance issues	20%	22%	12%	23%
Rationale can be made for more than one correct response	5%	10%	5%	14%
Page/item layout				
Embed item in passage	0%	0%	7%	16%
Increase white space	24%	20%	53%	8%
Change the size of item elements	0%	3%	0%	2%
Change the font size	16%	0%	0%	27%
Move item/change item order*	—	0%	0%	5%
Reduce spread of information across pages/screens	4%	0%	2%	61%

*In the version of the ARM used to rate the IN items, the item passage and stimulus elements were considered a single element. Additionally, the version did not contain a code for every suggested modification.

was to simplify or shorten the text, and the most common recommendations to improve an item stem were to simplify and/or shorten it and to clarify the question or directive. The most common recommendation across states in science to improve answer choices was to eliminate a distractor, and this was also the most common recommendation in mathematics. In language arts, to address balance issues and to eliminate answer choices were both likely recommendations, and in social studies to address balance issues and to revise answer choices were both likely recommendations. Across states and content areas simplifying a visual was a common recommendation, although in language arts, eliminating the visual was recommended almost universally (94%). Finally, the most common recommendation for improving item

Table 6: Percentages of items at each overall accessibility level, by content area and state.

State	Content area	1 (inaccessible for many test takers)	2 (maximally accessible for some test takers)	3 (maximally accessible for most test takers)	4 (maximally accessible for nearly all test takers)
State A	Across ($n = 159$)	2% ($n = 3$)	27% ($n = 43$)	58% ($n = 93$)	13% ($n = 20$)
	Language Arts ($n = 67$)	0% ($n = 0$)	31% ($n = 21$)	55% ($n = 37$)	13% ($n = 9$)
	Mathematics ($n = 50$)	4% ($n = 2$)	22% ($n = 11$)	64% ($n = 32$)	10% ($n = 5$)
	Social Studies ($n = 21$)	5% ($n = 1$)	24% ($n = 5$)	52% ($n = 11$)	19% ($n = 4$)
	Science ($n = 21$)	0% ($n = 0$)	29% ($n = 6$)	62% ($n = 13$)	10% ($n = 2$)
State B	Science ($n = 60$)	8% ($n = 5$)	23% ($n = 14$)	63% ($n = 38$)	5% ($n = 3$)
State C	Science ($n = 101$)	3% ($n = 3$)	24% ($n = 24$)	73% ($n = 73$)	1% ($n = 1$)
State D	Science ($n = 100$)	0% ($n = 0$)	50% ($n = 50$)	50% ($n = 50$)	0% ($n = 0$)

layout in most states and content areas was to increase white space. In language arts reducing the spread of information and embedding the items in passages were much more common recommendations. In state D, the most common recommendations for improving item layout were to reduce the spread of information across pages/screens and to change the font size.

The ARM [16, 17] is a tool being used by test developers and companies to operationalize the principles of accessibility theory, train item writers, and improve test items (http://edmeasurement.net/itemwriting; http://www.nwea.org). Like item fairness reviews, we believe item accessibility reviews conducted systematically and with the ARM will lead to improved items for all students.

4.1. Example Items. The practical results of this research are refined test items. Figure 3 consists of a biology end-of-course item from State D's item review sample. To the right of the item are accessibility ratings by item element and overall, using the ARM. The item stimulus reads as follows: "In animals, the lack of skin color is recessive; the presence of skin color is dominant. A male which is homozygous for skin color is crossed with a female who is heterozygous for skin color." The element earned a rating of 4 (*maximally accessible for nearly all test takers*) because it contains no extraneous verbiage and is clearly worded, and all information is necessary for responding to the item. The item stem reads as follows: "What is the probability that their offspring will lack skin color?" The item stem also received a rating of 4; it is presented separately from the stimulus, the directive is clearly written and uses the active voice, and it facilitates responding by using the same target word (i.e., "lack") as the noun in the first sentence of the stimulus. The item does not contain a visual, and the addition of one would not enhance the accessibility of the item. The answer choices are optimally accessible as well, presenting the four possibilities (one correct, the others representing common errors), in numerical order to maintain balance and prevent cueing of any option. Overall, the item received a rating of 4, as no rater observed any indicators of possible accessibility problems nor made suggestions for modifying the item to improve its accessibility for more test takers.

We created the item presented in Figure 4 to mirror one of State B's items that could not be released for publication. The item yielded a different result due mostly to the magnitude of cognitive demand, specifically what Sweller [18] refers to as *element interactivity*. Based on the item stem, the target construct appears to be integrating information from a graph using knowledge of biological concepts (i.e., environmental adaptation). The item requires the test taker to be simultaneously mindful of several element sets (dichotomies) to respond. First, the item stimulus provides background information about the amount of rainfall that reaches two different regions of Madagascar ("central-eastern Madagascar receives abundantly more rainfall than the southwestern region of the country.") Second, the test taker must recall, and store in short-term memory, the knowledge that the southwestern and central-eastern regions are equated with low and high amounts of monthly rainfall, respectively. A third dichotomy involves the two animals. Both animals' names consist of challenging phonologies (i.e., Verreaux's Sifaka and the Lac Alaotra Gentle Lemur). The item provides the genus-species names of each animal in parentheses and provides a local name for the second animal (i.e., Bandro). The latter is required to interpret graph and respond to the item stem. Given the high element interactivity with respect to the target construct of the item, the item stimulus received a rating of 2, or *maximally accessible for some test takers*.

According to Haladyna et al. [19] item stems should be written in the active voice. The item stem in Figure 4 is in the passive voice ("What conclusion *can best be drawn*...?"). Otherwise, the stem is written plainly. The stem received a rating of 3, *maximally accessible for most test takers*.

In the current item, the apparent purpose of the visual is to elicit demonstration of part of the target skill (i.e., integrating information from a graph). The visual requires the test taker to refer to the stimulus to understand that the notations B and V represent the two animals introduced in the stimulus. The diminutive arrow on the ordinate is the only indication that survival rate is the highest at the top of the graph (although this arguably is self-evident). The visual received a rating of 2, or *maximally accessible for some test takers*.

In rating the answer choices, we refer again to the notion of element interactivity, which presents a fourth set

00. In animals, the lack of skin color is recessive; the presence of skin color is dominant. A male which is homozygous for skin color is crossed with a female who is heterozygous for skin color.

What is the probability that their offspring will lack skin color?

√ A. 0%
B. 25%
C. 75%
D. 100%

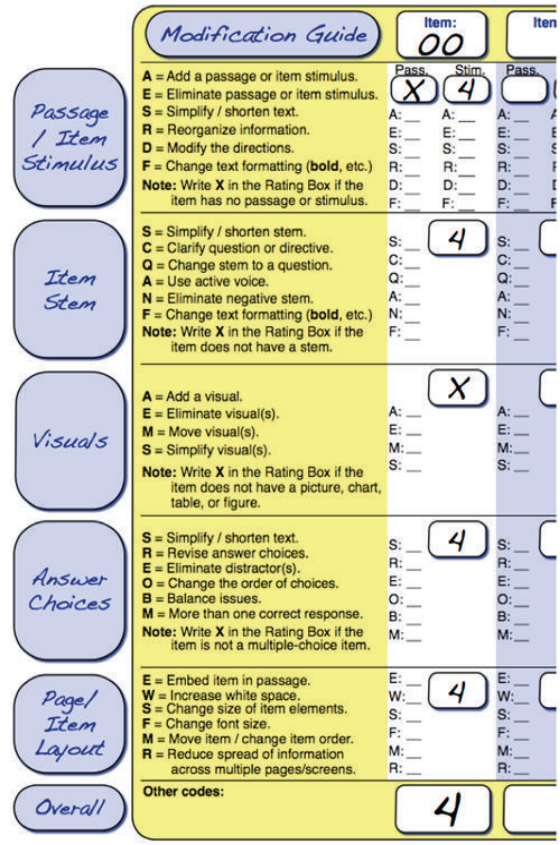


Figure 3: Grade 10 science item with ARM ratings.

of elements. Specifically, the answer choices are constructed such that options A and B use "better adapted," option C uses "poorly adapted," and option D uses "equally well adapted," with options A and B referring to the central-eastern region, option C referring to the southwestern region, and option D referring only to the amount of rainfall. The explanation for this can be difficult, and the reader is cautioned to be mindful of the subset of test takers who may choose to "overthink" an item such as this. If a particular test taker, observing that the graph's ordinate contains no scale, decides that one *could* draw the conclusion that the two animals both are equally adapted for life where there is zero rainfall, thus, he or she could overlook option C and select D as the correct response. The large majority of successful test takers (such as, we presume, those reading this paper) immediately see the flaw in this test taker's reasoning, seeing that there is a much better answer in option C. Indeed, distractor statistics for the item on which this example was based reveal only 7% of test takers selected option D. Nevertheless, we contend the test item is not intended to measure the test taker's ability to find a best answer among others that may be logically correct, so it may behoove the item writer to remove this option. According to these suggested changes, the answer choices received a rating of 2, or *maximally accessible for some test takers*.

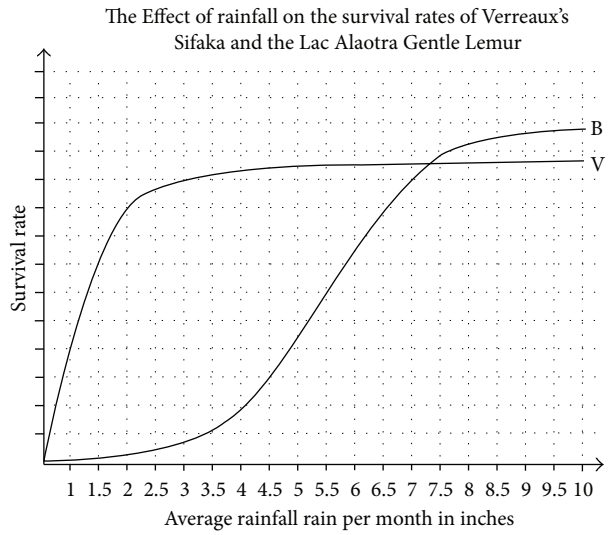
Finally, the item is presented on one page, is clearly organized, and contains sufficient white space. However, the font sizes could be increased, and a sans serif font could be used. Accordingly, the page and item layout received a rating of 3, *maximally accessible for most test takers*. Overall, for the reasons indicated previously, the item received a rating of 2, *maximally accessible for some test takers*.

Harking back to the target construct as defined above, we contend the degree of element interactivity in the item in Figure 4 is higher than necessary for measuring the test taker's ability to integrate information from a graph using his or her knowledge of environmental adaptation. We demonstrated how the item received a considerably lower overall accessibility rating than the previous item primarily because of the low ratings of the stimulus, visual, and answer choices.

Figure 5 contains the same item after making suggested changes to exemplify how ARM ratings may be used to guide item modification. In its modified form, the item measures the same construct as the item in Figure 4, but the element interactivity has been reduced. Specifically, the item stimulus has been reorganized and simplified. It begins with a statement introducing the primary elements of the item, Animals A and B. It proceeds to present two informational statements about the rainfall in the two regions of Madagascar, using bullet points to set the statements apart for later reference. The next part of the stimulus introduces the graph. The visual has been modified slightly, reducing the number of words in the title, adding labels to the ordinate, simplifying the scale

17. Central-eastern Madagascar receives abundantly more rainfall than the southwestern region of the country.

Verreaux's Sifaka (*Propithecus verreauxi*) lives in the dry, deciduous forests of southwestern Madagascar. The Lac Alaotra Gentle Lemur (*Hapalemur alaotrensis*), known locally as the Bandro, lives further north, in central-eastern Madagascar. The graph below shows the effect of monthly rainfall on the survival rates of Verreaux's Sifaka (V) and the Bandro (B).



What conclusion can best be drawn about the graph above?

A. Verreaux's Sifaka is better adapted to the central-eastern region than the Bandro.
B. The Bandro is better adapted for life in the central-eastern region than Verreaux's Sifaka.
√ C. The Bandro is poorly adapted for life in southwestern Madagascar compared to Verreaux's Sifaka.
D. Verreaux's Sifaka and the Bandro are equally well adapted for life with very little rainfall.

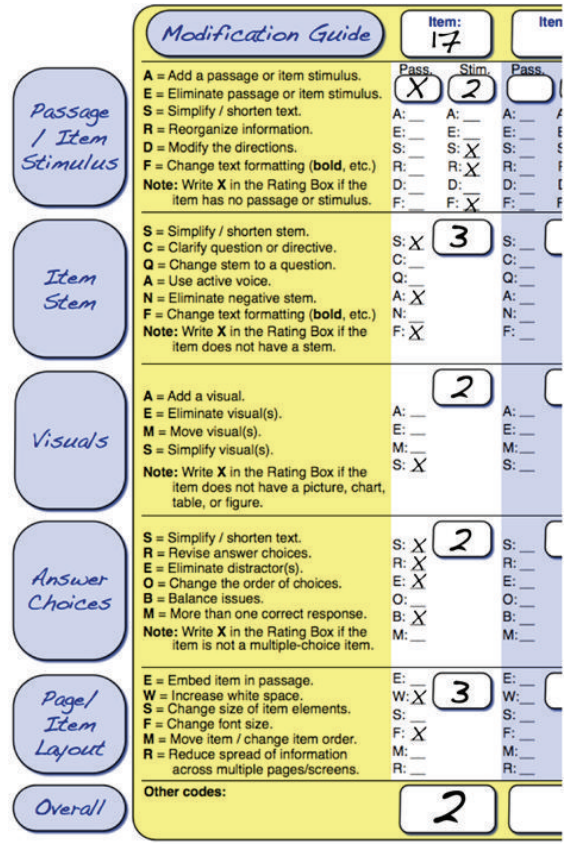


Figure 4: Grade 10 science item with ARM ratings, original.

on the abscissa, and clarifying the labels for the two animals. The stem has been shortened and reworded in the active voice. The answer choices all contain comparative statements about the two plants. As per the recommendation of [20], an implausible answer choice has been removed, leaving one key and two distractors. All of the answer choices refer to southwestern Madagascar. The second answer choice mirrors the first, changing only the comparative order of the two animals. The third choice changes the comparative statement to an equality.

In its current form, the item stimulus is improved from the original. However, the stimulus could be revised to reduce the element interacivity even further, by eliminating the reference to Madagascar. Indeed, the construct arguably would be preserved even if the test taker were required only to identify the adaptability by the amount of rainfall. Thus, the item stimulus received a rating of 3, *or maximally accessible for most test takers*. The revised item stem, containing simplified language and written in the active voice, received the highest accessibility rating. The visual, while relatively clear by comparison to its original "parent," could be further simplified, or perhaps expanded to increase white space, and thus received a rating of 3, *maximally accessible for most test takers*. The answer choices received the highest accessibility rating. In its modified form, the item received a rating of 3, *maximally accessible for most test takers*, indicating that the item is highly accessible but may be improved further with another iteration of modification.

4.2. The Art and Science of Item Reviews: Iterative and Inclusive. The process of identifying item features that may present access barriers for some test takers, quantifying the accessibility of item elements to suggest changes to improve their accessibility and undertaking modification to enhance suboptimally accessible items, requires knowledge of the content and testing standards of the target test as well as knowledge of the intended test taker population. We therefore contend the item review and modification process should involve multiple trained raters, with the assumption that collaboration increases the likelihood that the process will yield information about potential access concerns that may be overlooked by an individual rater.

Additionally, we argue that item modification is best defined as an iterative process, whereby items undergo an initial review process by a team of item accessibility experts, followed by a round of modifications. Accessibility reviews

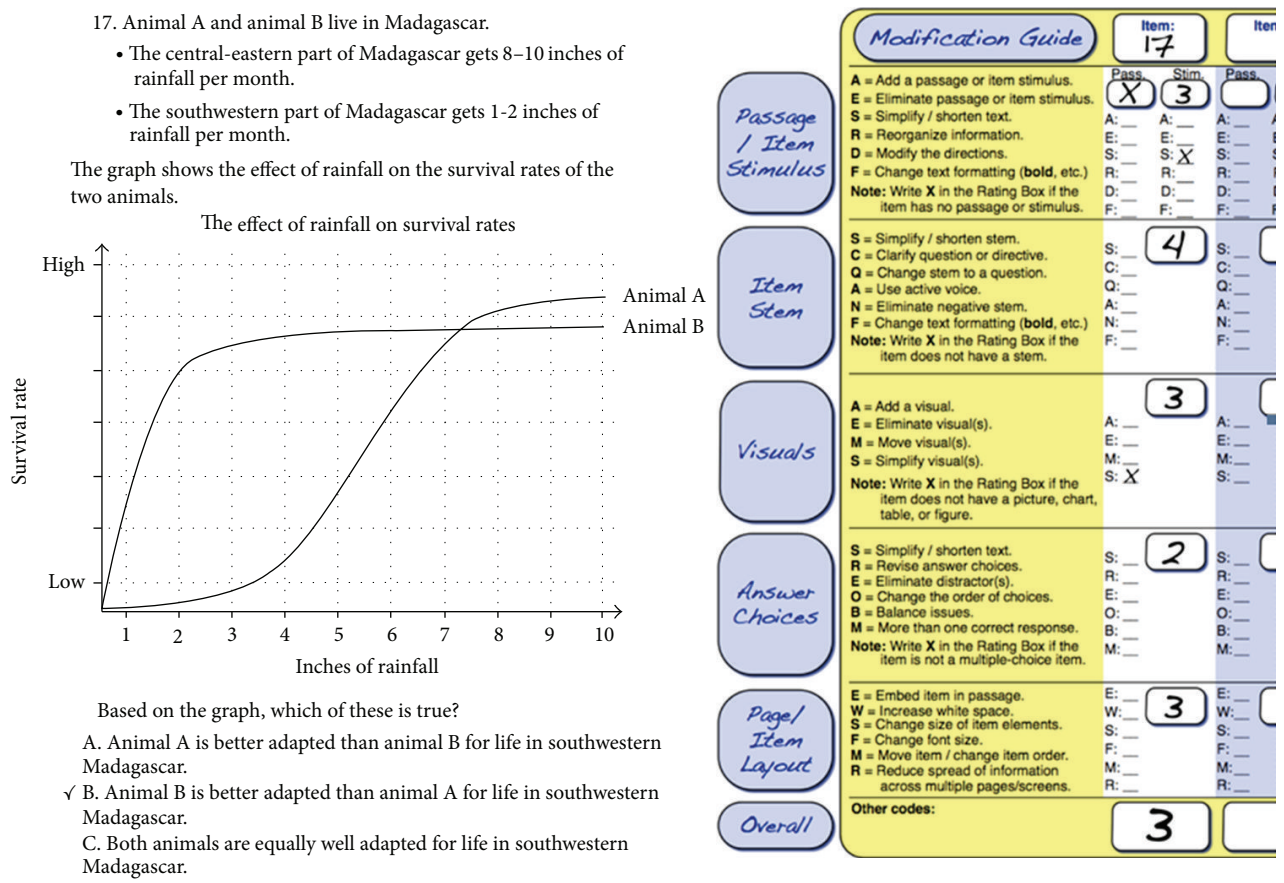


Figure 5: Grade 10 science item with ARM ratings, modified.

can be conducted prior to, or following, Rasch testing and other statistical tests of validity. We recommend the test developers run cognitive labs and field-test the items, followed by an analysis of resulting psychometric changes. Based on these data as well as on a subsequent item review, the test developer should identify items that may have been poorly modified or that may benefit from additional modification and should revise those items accordingly. Such a process is particularly important when tests are used for "high stakes" decision making, but should be considered for assessment for learning as well (i.e., formative assessment; [21]), as this research indicates item-level accessibility problems are common.

Additionally, we suggest that the quantitative accessibility review process described in this article does not preclude test developers from using other means of evaluating the accessibility of test items; indeed, cognitive labs and/or interviews, statistical item analysis, and field testing certainly may inform the review process.

Finally, the use of accessibility reviews should not be limited to students identified with disabilities. The comparatively greater ability of general education students to navigate tests and test items notwithstanding, the validity concerns that plague tests of students identified with disabilities likely exist for those used with the general population as well.

5. Conclusion

The results of four state accessibility reviews indicated the state tests, as defined by a representative sample of test items, consisted of approximately 30% items that were less than maximally accessible for most test takers. This suggests there are large numbers of test takers whose test scores may not accurately reflect the extent of their knowledge due to access barriers intrinsic to the items. In general, test developers of both formative and summative assessments are advised to systematically examine all test items with a focus on reducing extraneous complexity in text and visuals and organizing information to facilitate responding. As well, they should ensure, to the extent possible, items are free from nonessential material that may impose demands on the test taker that may siphon essential cognitive resources from the targeted interaction, thus introducing error into subsequent test score inferences. Advances in testing and the validity of resulting scores improve when items are highly accessible for all test takers.

References

[1] K. Hollenbeck, *Determining when Test Alterations are Valid Accommodations or Modifications for Large-Scale Assessment*,

Large-Scale Assessment Programs for All Students: Validity Technical Adequacy, and Implementation, 2002.

[2] U. S. Department of Education, *Modified Academic Achievement Standards: Non-Regulatory Guidance*, Washington, DC, USA, 2007.

[3] U. S. Department of Education, *Standards and Assessments Peer Review Guidance*, Washington, DC, USA, 2007.

[4] P. A. Beddow, A. Kurz, and J. R. Frey, "Accessibility theory: guiding the science and practice of test item design with the test taker in mind," in *Handbook of Accessible Achievement Tests*, S. N. Elliott, R. J. Kettler, P. A. Beddow, and A. Kurz, Eds., Springer, New York, NY, USA, 2011.

[5] L. R. Ketterlin-Geller, "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*, vol. 27, no. 3, pp. 3–16, 2008.

[6] CAST, *Universal Design For Learning Guidelines Version 2.0*, Wakefield, Ma, Boston, 2011.

[7] M. Luethi, B. Meier, and C. Sandi, "Stress effects on working memory, explicit memory, and implicit memory for neutral and emotional stimuli in healthy men," *Frontiers in Behavioral Neuroscience*, vol. 2, article 5, 2008.

[8] R. Bull, K. A. Espy, and S. A. Wiebe, "Short-term memory, working memory, and executive functioning in preschoolers: longitudinal predictors of mathematical achievement at age 7 years," *Developmental Neuropsychology*, vol. 33, no. 3, pp. 205–228, 2008.

[9] S. G. Sireci, S. E. Scarpati, and S. Li, "Test accommodations for students with disabilities: an analysis of the interaction hypothesis," *Review of Educational Research*, vol. 75, no. 4, pp. 457–490, 2005.

[10] R. J. Kettler, M. C. Rodriguez, D. M. Bolt, S. N. Elliott, P. A. Beddow, and A. Kurz, "Modified multiple-choice items for alternate assessments: reliability, difficulty, and differential boost," *Applied Measurement in Education*, vol. 24, no. 3, pp. 210–234, 2011.

[11] S. N. Elliott, R. J. Kettler, P. A. Beddow, and A. Kurz, "Research and strategies for adapting formative assessments for students with special needs," in *Handbook of Formative Assessment*, H. L. Andrade and G. J. Cizek, Eds., pp. 159–180, Routledge, New York, NY, USA, 2010.

[12] C. C. Laitusis, "Examining the impact of audio presentation on tests of reading comprehension," *Applied Measurement in Education*, vol. 23, no. 2, pp. 153–167, 2010.

[13] M. L. Thurlow, C. C. Laitusis, D. R. Dillon et al., *Accessibility Principles For Reading Assessments*, National Accessibility Reading Assessment Projects, Minneapolis, Minn, USA, 2009.

[14] S. N. Elliott, R. J. Kettler, P. A. Beddow et al., "Effects of using modified items to test students with persistent academic difficulties," *Exceptional Children*, vol. 76, pp. 475–495, 2011.

[15] S. N. Elliott, M. C. Rodriguez, A. T. Roach, R. J. Kettler, P. A. Beddow, and A. Kurz, *AIMS EA, 2009 Pilot Study*, Learning Sciences Institute, Vanderbilt University, Nashville, Tenn, USA, 2009.

[16] P. A. Beddow, S. N. Elliott, and R. J. Kettler, *Accessibility Rating Matrix (ARM)*, Vanderbilt University, Nashville, Tenn, USA, 2009.

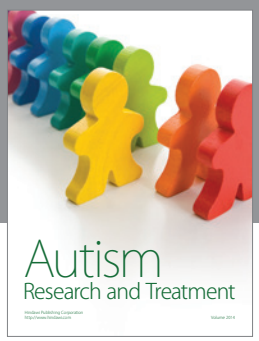
[17] P. A. Beddow, S. N. Elliott, and R. J. Kettler, *Test Accessibility and Modification Inventory (TAMI) Technical Supplement*, Vanderbilt University, Nashville, Tenn, USA, 2009.

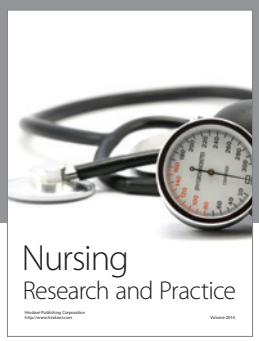
[18] J. Sweller, "Element interactivity and intrinsic, extraneous, and germane cognitive load," *Educational Psychology Review*, vol. 22, no. 2, pp. 123–138, 2010.

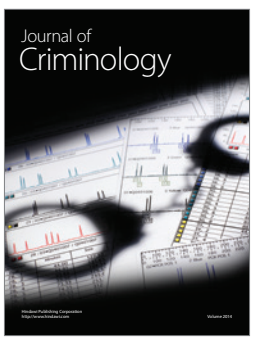
[19] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," *Applied Measurement in Education*, vol. 15, no. 3, pp. 309–334, 2002.

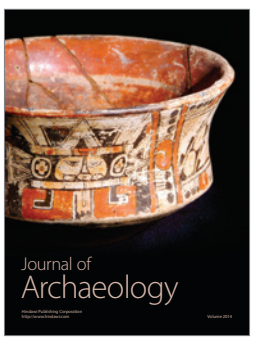
[20] M. C. Rodriguez, "Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research," *Educational Measurement*, vol. 24, no. 2, pp. 3–13, 2005.

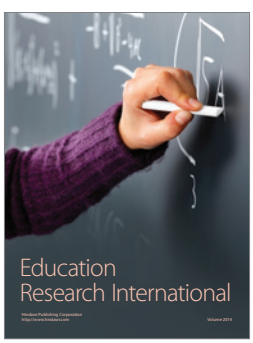
[21] H. L. Andrade and G. J. Cizek, *Handbook of Formative Assessment*, Routledge, New York, NY, USA, 2010.

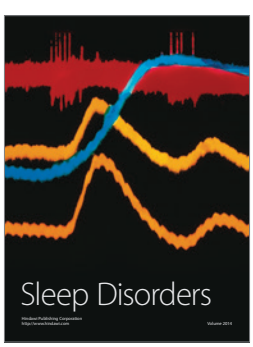


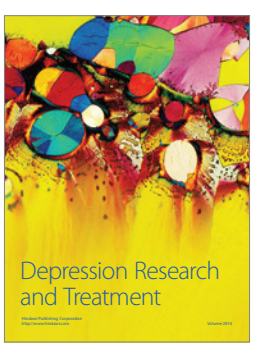


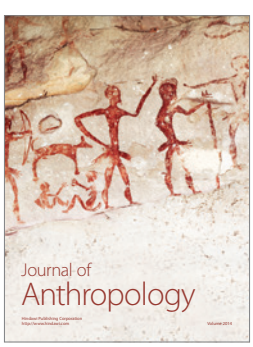


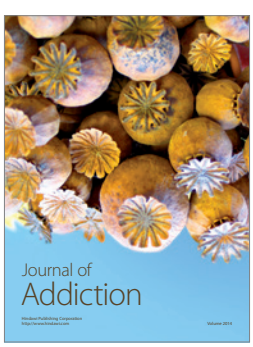


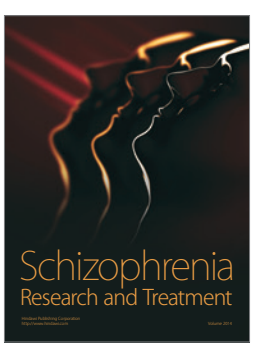


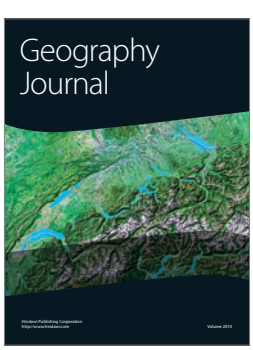


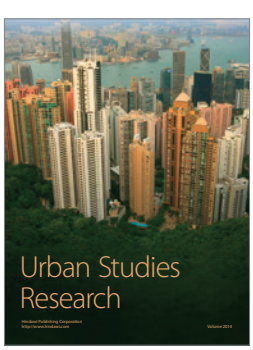












Hindawi

Submit your manuscripts at
http://www.hindawi.com