

Test Accessibility and Modification Inventory

TAMITM

Accessibility Rating Matrix

Technical Supplement

Peter A. Beddow, Stephen N. Elliott, & Ryan J. Kettler

VANDERBILT  UNIVERSITY

November 2009

Purpose and Rationale

The Test Accessibility and Modification Inventory (TAMI) is a multi-purpose training and evaluation tool with two components. First, the inventory component of the TAMI provides practical guidance for item writers and test developers to ensure tests and test items are optimally accessible. Second, the Accessibility Rating Matrix (ARM) is a measurement tool for evaluating the accessibility of tests and test items accessibility and suggesting modifications that may improve accessibility for more test-takers. The authors of the TAMI argue that the by applying the principles of accessibility theory and systematically using the TAMI and TAMI ARM during the test development process will result in more accessible tests and improved measurement test-takers with a broad range of abilities and needs. Students identified with disabilities comprise approximately 10-12% of the student populations of most states. Common characteristics of these students include reading problems, attention difficulties, and feelings of anxiety and frustration with schoolwork and testing. As a result, many of these students experience persistent academic difficulties and learn at a rate slower than their peers. Additionally, there are an estimated 5-10% of students across states who have not been identified with disabilities but for whom the demands of school and tests are similarly challenging.

Recent federal legislation permits states to report proficiency for calculating AYP for up to 2% of students using results from an alternate assessment based on modified academic achievement standards (AA-MAS). To be eligible for participation in an AA-MAS, a student must be identified with a disability and must have scored below-proficient on the statewide assessment in prior years with the data-based prediction that he or she is unlikely to achieve proficiency on the general state assessment during the current year. When an IEP team agrees to permit the student to participate in the AA-MAS, the implication is that because of this student's special needs, the general state assessment is unlikely to yield scores from which valid inferences can be made about the student's knowledge of the academic content standards.

The Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, & Elliott, 2008) and the TAMI Accessibility Rating Matrix (Beddow, Elliott, & Kettler, 2009) represent the combined effort of a team of educators, assessment leaders, researchers and test developers across several states to evaluate the accessibility of test items with the goal of improving testing for all students.

Accessibility Theory

Accessibility Theory provides a framework for improving tests for all individuals by offering a perspective on the measurement of this target construct in terms of three sets of variables: the test-taker, the test, and the test event. The theory is based on three primary areas of research and theory: universal design (UD) principles, cognitive load theory (CLT), and research on test and item design.

Current federal legislation requires the application of universal design principles to the development of all state and district-wide achievement tests. Universal design, as defined in the Assistive Technology Act (P.L. 105-394, 1998), is "a concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly usable (without requiring assistive technologies) and products and services that are made usable with assistive technologies" (§3(17)). While the term *accessibility* is not used in this definition, universal

design principles as applied to assessment technology clearly are intended to address issues of access while responding to the concern raised in the Testing Standards that the use of individualized accommodations may increase measurement error. Guidance from researchers at the National Center on Educational Outcomes (NCEO) provides recommendations for applying UD to testing (Thompson, Johnstone, & Thurlow, 2002; Johnstone, Thurlow, Moore, & Altman, 2006).

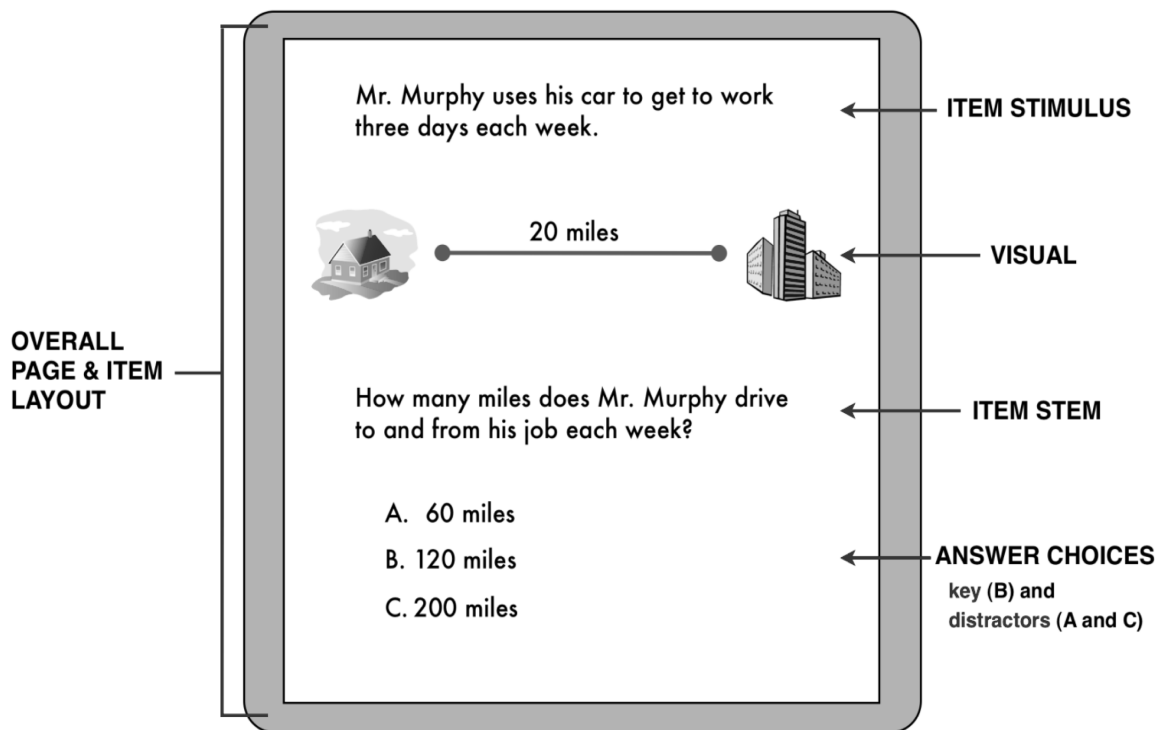
Originally conceived and applied as a model for understanding the demands of instructional activities, cognitive load theory has clear implications for the development of accessible tests. The model disaggregates the cognitive demands of learning tasks into three load types: intrinsic load, germane load, and extraneous load. The triune model of cognitive load was encapsulated by Paas, Renkl, and Sweller (2003) as follows: “Intrinsic, extraneous, and germane cognitive loads are additive in that, together, the total load cannot exceed the working memory resources available if learning is to occur”(p.2). With regard to assessment, application to CLT focuses on certain aspects of test items that may require the test-taker to use cognitive resources in excess of those needed to demonstrate knowledge of the tested content. Excessive cognitive load may be reduced by eliminating extraneous material from items, reducing reading load, and using visuals efficiently.

Further, evidence based on the collective expertise of test and item development scholars has informed the development of the TAMI and has advanced our understanding of test and test item accessibility. Among these is Haladyna, Downing, and Rodriguez’s (2002) research-based taxonomy consisting of a number of recommendations for writing multiple-choice items. Another is Rodriguez’s (2005) meta-analysis of over 80 years of research on item development which led to his conclusion that three answer choices are optimal for multiple-choice items. Rodriguez’ results indicate that reducing items from 4 or 5 answer choices to 3 tends to result in small, nonsignificant changes in item difficulty, discrimination, and reliability, and such modifications are sensible in that the reduced reading load results in a test that is more manageable for a group of students who typically take much longer to finish. Likewise, the decreased reading load may permit test developers to add more items and increase the reliability of the test. Accordingly, the authors of the TAMI support the practice of reducing the number of response options from 4 to 3, when doing so eliminates an implausible distractor and results in a set of answer choices that are balanced and free of cueing.

Description of the Instruments

The TAMI and TAMI ARM are user-friendly tools that have been used by over 100 educators and assessment experts to improve the accessibility of test items. These tools provide item developers and evaluators a systematic structure for creating new items or enhancing existing items. The two parts of this pragmatic approach to improving tests for all students are described next.

Figure 1. Anatomy of a multiple-choice item.



Test Accessibility and Modification Inventory. The inventory component of the TAMI consists of two sections: Item Analysis and Computer-Based Test Analysis. The TAMI Item Analysis section contains 46 descriptors organized into five primary categories based on the key elements of most test items: the item passage and/or stimulus, item stem, visuals, answer choices, and layout (see Figure 1.) An additional Fairness category consists of a distillation of the Fairness Review Guidelines (Educational Testing Service, 2009). The Computer-Based Test Analysis section follows the same format as the Item Analysis section and contains 35 descriptors across four categories: Test Delivery System, Test Layout, Training, and Audio. Guidance on web accessibility standards and current research and practice regarding computer-delivered tests (e.g., NimbleTools; see <http://www.nimbletools.org>) were primary influences for this section of the TAMI.

TAMI Accessibility Rating Matrix. The TAMI ARM consists of two scoring rubrics: the Item Analysis and the Overall Analysis. After writing the item number on the ARM Record Form, the rater begins by using the Item Analysis rubric to evaluate the accessibility of the item according to the key elements of a multiple-choice test item. While individual test items may or may not include each of these elements, the ARM was designed with the assumption that this anatomic structure can be used to evaluate most current assessment item formats.

Using the Item Analysis rubric, the rater determines the accessibility level of each item element on a 4-point scale (see Figure 2). After assigning a rating to each item element, the rater uses the Modification Guide to select modifications that are likely to improve the accessibility of

the item. Additionally, there is a space at the bottom of the ARM Record Form for raters to code suggested modifications that are not listed on the Modification Guide. After rating the individual item elements, the rater reviews the item element ratings and selects an Overall Accessibility rating using the Overall Analysis rubric.

Figure 2. TAMI accessibility levels.

Level	Description	Heuristic
4	Maximally Accessible for Nearly All Test-Takers	Optimal accessibility for between 95-99% of the population
3	Maximally Accessible for Most Test-Takers	Optimal accessibility for between 90-95% of the population
2	Maximally Accessible for Some Test-Takers	Optimal accessibility for between 85-90% of the population
1	Inaccessible for Many Test-Takers	Optimal accessibility for fewer than 85% of the population

Evidence for the Reliability of Accessibility Ratings

Accessibility involves an interaction between item features and test-taker characteristics; raters, therefore, should have both expertise in the test content and an intimate understanding of the range of abilities and needs across the intended test-taker population. Levels of accessibility on the ARM are based on the extent to which the item is *maximally accessible* for a given portion of the intended test-taker population. A maximally accessible item is an item that contains no barriers that would limit or hinder the test-taker from demonstrating his or her knowledge of the target construct. An item that is maximally accessible for nearly all test-takers, therefore, requires few, if any, cognitive resources in excess of those needed to show what the test-taker knows. If the extraneous cognitive load demand of an item differentially impacts performance on the item across the test-taker population, the item’s accessibility is less than optimal.

The ARM rubrics are intended to guide professional judgments about test accessibility rather by providing a flexible algorithm for assigning accessibility ratings to test items. It is essential, therefore, that raters are trained in evaluating accessibility prior to using the TAMI to ensure sufficient reliability of scores and validity of subsequent diagnostic inferences and recommendations. Based on the current data, the reliability of accessibility ratings by evaluators with extensive training in accessibility principles is high. Specifically, across three item reviews totaling over 350 test items, perfect or adjacent agreement for ratings by an evaluation team of assessment researchers and professors trained to use the ARM was over 94%.

Evidence for the Validity of Accessibility Ratings

To date, several studies have yielded empirical data supporting the use of TAMI to facilitate the development of accessible tests and contributing to an increased understanding of test accessibility. These studies have emerged from three federal grant projects aimed at supporting states’ efforts to develop an AA-MAS for students with persistent academic

difficulties: (a) the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES; Elliott & Compton, 2006-2008), (b) the Consortium for Modified Alternate Assessment Development and Implementation (CMAADI; Elliott, Roach, & Rodriguez, 2008-2010) and (c) Operationalizing Alternate Assessment of Science Inquiry Skills (OAASIS, 2009-2011). Results have indicated students perform comparatively better on items modified with a focus on accessibility than on unmodified sibling items (Elliott, Kettler, Beddow, Kurz, Compton, et al., In press). Evidence also indicated the magnitudes of differences in item difficulty across modified items for students identified with disabilities who would be eligible for an AA-MAS are greater than those observed for their general education peers who would not be eligible for a modified assessment. Further, students who completed surveys and cognitive interviews following testing reacted positively to the item changes (Roach, Beddow, Kurz, Kettler, & Elliott, In press).

Additionally, results of two field tests of the TAMI indicate accessibility ratings are sensitive to items across a range of accessibility levels. In both studies, raters were randomly assigned a set of items to evaluate and results were analyzed to determine whether ratings were higher for items that had undergone modification to improve their accessibility. In the second of these studies, members of a team of raters who had recently received comprehensive training in the use of the ARM were randomly assigned to rate items from two item pools: (a) publicly-released items from the Texas Assessment of Knowledge and Skills (TAKS) and (b) corresponding items in modified form for the modified version of the TAKS (called the TAKS-M). Results indicated accessibility ratings of original (unmodified) versions of the items had significantly lower accessibility ratings than those on modified items. Further, ratings by the evaluation and training team corresponded highly with ratings by the novice raters, which is another indicator that when used by trained raters, the TAMI yields reliable ratings of accessibility.

Accessibility Reviews

Most states developing an AA-MAS have made modifications to the pool of items used for the general assessment to enhance their accessibility for the new test. The procedures used for modifying items vary across states and until recently, no state has attempted to quantify the accessibility of their assessment items. Without a thorough examination of the accessibility across a pool of test items, predictions about how the items may perform when used for the group of test-takers for whom the items are intended likely will be unreliable or invalid.

To date, the TAMI evaluation team has conducted several reviews of large item pools across a number of content domains and grade levels for several state education departments. For each review, the team examined each item in a representative item pool based on the five elements that are common to the majority of large-scale test items, including the item passage or stimulus, item stem, visuals, answer choices, and layout (Figure 1), noting patterns across the reviewed items and suggesting modifications that may improve the accessibility of the items for more test-takers. The team provided a report following each review to summarize features of the item pool that may enhance or reduce the accessibility of the items for portions of the test-taker population and to provide recommendations for improvement. Evidence suggests if test developers follow the team's suggested modifications in revising the items, the changes should result in optimally accessible items for nearly all test takers.

Contact Information

For copies of the TAMI or the TAMI Accessibility Rating Matrix, please visit the TAMI webpage at <http://peabody.vanderbilt.edu/tami.xml>. For further information about accessibility theory or the accessibility review process, please contact the senior author of the TAMI by mail, phone, or email:

Peter Beddow
Peabody Box #59
230 Appleton Place
Nashville, TN 37203-5721
(615) 403-8206
peter.beddow@vanderbilt.edu

References

- Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009). *TAMI Accessibility Rating Matrix*. Nashville, TN: Vanderbilt University.
- Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test Accessibility and Modification Inventory (TAMI)*. Nashville, TN: Vanderbilt University.
- Educational Testing Service. (2009). ETS Guidelines for Fairness Review of Assessments. Retrieved on November 13, 2009 from http://www.ets.org/Media/About_ETTS/pdf/overview.pdf
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M. C., Roach, A. T., & Bolt, D. (In press). Using modified items to test students with persistent academic difficulties. *Exceptional Children*.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-344.
- Johnstone, C., Thurlow, M., Moore, M., & Altman, J. (2006). *Using systematic item selection methods to improve universal design of assessments* (Policy Directions 18). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. Modified multiple-choice items for alternate assessments: Reliability, difficulty, and the interaction paradigm. Manuscript under review by *Applied Measurement in Education* for publication.
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38, 1-4.
- Roach, A. T., Beddow, P. A., Kurz, A., Kettler, R. J., & Elliott, S. N. (In press). Using student responses and perceptions to inform item development for an alternate assessment based on modified achievement standards. *Exceptional Children*.