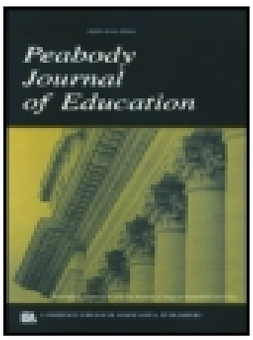

Peabody Journal of Education

ISSN: 0161-956X (Print) 1532-7930 (Online) Journal homepage: https://www.tandfonline.com/loi/hpje20

Psychometric Considerations for Alternate Assessments Based on Modified Academic Achievement Standards

Michael C. Rodriguez

To cite this article: Michael C. Rodriguez (2009) Psychometric Considerations for Alternate Assessments Based on Modified Academic Achievement Standards, Peabody Journal of Education, 84:4, 595-602, DOI: 10.1080/01619560903241143

To link to this article: https://doi.org/10.1080/01619560903241143

PEABODY JOURNAL OF EDUCATION, 84: 595–602, 2009

ISSN: 0161-956X print / 1532-7930 online
DOI: 10.1080/01619560903241143

Routledge
Taylor & Francis Group

Psychometric Considerations for Alternate Assessments Based on Modified Academic Achievement Standards

Michael C. Rodriguez
Educational Psychology, University of Minnesota

Because of the unique nature of the students eligible for alternate assessments based on modified academic achievement standards, their varied access to the general education curriculum, and their unique learning needs, innovative psychometric thinking and practice is needed to assure high technical quality of alternate assessments. Indeed, we at least must marshal state-of-the-art procedures to secure strong psychometric evidence to support appropriate and meaningful design and use of these important assessments. The authors contributing work to this special issue, Alternate Assessments Based on Modified Academic Achievement Standards, address important issues and provide guidance to policymakers, test developers, and educators. They also each raise important technical quality issues. This article offers a brief review of such psychometric considerations, in light of the work and comments of the special issue authors.

This special issue of the *Peabody Journal of Education* on Alternate Assessments Based on Modified Academic Achievement Standards: New Policy, New Practices, and Persistent Challenges provides researchers, educators, and policymakers with a set of important considerations and guidance in our current exploration of the world of alternate assessment design and implementation. Although not explicitly intended, the contributors to the special issue present issues that have implications for the technical quality of alternate assessments based on modified academic achievement standards (AA-MAS). To focus attention, I briefly review the technical issues raised. I then turn to psychometric considerations to provide guidance for establishing a high level of quality evidence to support some of the more ambitious claims about AA-MAS.

CONTRIBUTORS RAISE TECHNICAL QUALITY ISSUES

Authors of all six articles in this special issue present important challenges to the technical quality of AA-MAS. Some of this attention is due to the language in the federal guidance and peer review documentation. Guidance on AA-MAS development allows for the use of modified versions of the general education assessment or the development of new tests and the use of easier achievement standards and easier test questions. At the same time, the content covered

Correspondence should be sent to Michael C. Rodriguez, 250 Education Sciences, 56 East River Road, University of Minnesota, Minneapolis, MN 55455. E-mail: mcrdz@umn.edu

on the alternate assessment must be the same as the general assessment (requiring alignment with grade-level content standards). In their introduction to this special issue, Kettler, Elliott and Beddow (2009/this issue) raise three challenges in these requirements, including the identification of qualified students, the degree of appropriate modification to achievement standards, and the demands of adequately documenting technical quality. These challenges have psychometric implications and are reflected in the comments of the other authors.

Identification is the focus of the work by Zigmond and Kloo (2009/this issue). After a review of legislation leading up to the inclusion of students with disabilities in statewide assessment programs and the authorization for alternate assessments, Zigmond and Kloo examine decisions that states encounter to identify the appropriate population for the AA-MAS. They raise important questions regarding the low proficiency rates of students on the general assessment as a reflection of ability versus access or the presence of construct-irrelevant variance. With respect to the fit of the assessment, construct-irrelevant variance is a serious source of interference with test interpretation, misinforming subsequent decisions. Herein is a challenge to states: create a meaningful alternate assessment to meet the assessment needs of students with disabilities that considers their unique instructional and learning needs.

Lazarus and Thurlow (2009/this issue) identify changes in states' AA-MAS between 2007 and 2008, based on information made available at state Web sites and follow-up with state personnel. They found online information for six states in 2007 and nine in 2008 with AA-MAS in place or under development. Regarding the characteristics of the assessments in 2008, they found all nine states included multiple-choice items, two had constructed-response items, five employed a writing prompt, and one included performance tasks. Lazarus and Thurlow also review methods of test design of the nine states: eight used fewer items, six removed a multiple-choice distractor, six employed simplified language throughout, five had fewer reading passages, and five had shorter reading passages. In addition, five states used fewer items per page, five used larger font size, three employed segmentation of passages, and two bolded or underlined key text. States are constructing their AA-MAS within the framework of their existing assessment programs.

A core concept in the development of AA-MAS and the intent of alternate assessments for students with disabilities overall is increasing access to grade-level general education curriculum. Roach et al. (2009/this issue) tackle this overriding goal. In all of the legislation and guidance documentation, the issue of access to the general curriculum is prominent, expected, and in many cases assumed. But as most of the authors acknowledge in regard to their specific topic, the students for whom AA-MAS would be most appropriate may not have full access to the general curriculum. Roach and colleagues aptly review the contexts, opportunities, and information needs to address curricular access for all students. They point to the need for a multidimensional measure of student engagement in general education programming, which can play an important part in interpreting the results of AA-MAS, improving our ability to make meaningful inferences about student achievement of grade-level content standards.

The task of taking an existing assessment system through an appropriate modification process (while maintaining the coherence between instruction, learning, and assessment for the students who would benefit the most) is a daunting challenge. Kettler, Elliott, and Beddow (2009/this issue) have undertaken this challenge in the development of their theory-guided data-based approach to design and modification. The Test Accessibility and Modification Inventory (TAMI) presents systematic guidance to developing, modifying, and evaluating items and tests to maximize accessibility and improve measurement. It is based on principles of cognitive load

theory, universal design, and item writing research and guidelines—essentially better measurement. The TAMI could easily be integrated into the development process of all assessments. The cognitive lab and experimental studies based on the use of the TAMI provide strong support for its use. Kettler, Elliott, and Beddow also review important sources of validity evidence based on the framework in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), including construct-related evidence, and evidence related to content, response processes, internal structure, relations to other variables, and the consequential bases of validity. They support their work with references to the *Standards* and empirical research based on the principles used to develop the TAMI, providing a strong basis for its use.

Continued use of an assessment in an accountability system requires the adoption of performance standards, including the development of performance level descriptors. Egan, Ferrara, Schneider, and Barton (2009/this issue) address the requirements to accomplish full implementation of AA-MAS. They argue that the technical quality of AA-MAS is critical, including evidence to support the validity of inferences about academic achievement of grade-level content standards, which support the decisions resulting from the application of performance standards. They also discuss the challenges in following standard procedures, as the unique characteristics of AA-MAS require innovative thinking. Some of the challenges—for example, that many state AA-MAS are shorter than general assessments—should be considered limitations of the methods we have developed for longer assessments; many methods, such as alignment and standard setting methodologies, rely on the presence of a great deal of information in long standardized tests. These limitations should not be construed as limitations of AA-MAS but limitations in existing methodologies to establish the technical quality of AA-MAS. Unfortunately, Egan and colleagues suggest that the process of gathering validity evidence "is conceptually complex and requires significant sustained effort" (p. 573). Certainly validation requires sustained effort and should be ongoing. The idea that it is conceptually complex is not helpful and points to the inability of validity theorists to communicate clearly to those who need strong guidance the most.

States continue to tackle these and many other challenges to provide access to high quality content and instruction and assess student performance. Palmer (2009/this issue) surveys assessment personnel from 22 states regarding their experiences with AA-MAS implementation. She notes that state personnel are looking for consistent guidance on the technical specifications necessary for acceptable AA-MAS. States appear uncertain about federal requirements and the effectiveness of AA-MAS. A strong example of these concerns is embodied in one respondent's comments: "We'd need to conduct research on comparability, validity, reliability, impact, use of accommodations, etc.—all of that costs quite a bit. We'd need to conduct a standard setting, and do an alignment study" (p. 581). States recognize the importance of technical quality and their limited resources to provide such evidence.

A PSYCHOMETRIC REVIEW

There are multiple psychometric considerations raised in this special issue, and each one deserves another manuscript. For example, Egan and colleagues make a psychometric suggestion to support standard setting with AA-MAS, by augmenting the ordered item booklet used in some standard setting procedures. If the use of fewer items (typical in AA-MAS) results in limited coverage of the

ability continuum, augmenting the item booklet could improve the standard setting task—because the task of the standard setting panel is not to identify the number of items necessary to reach a certain performance level, but the ability that is associated with that level (based on performance-level descriptors). So if augmenting the item booklet improves the ability continuum coverage, particularly in the regions associated with the thresholds between the performance levels, the resulting cut scores are better identified and supported.

This is related to a point made by Kettler, Elliott, and Beddow regarding validity evidence based on internal structure. They suggest that item response theory (IRT) is useful to improve the match between ability assessment and item difficulty. IRT provides a model to estimate the amount of information provided by test items across the ability continuum (the item characteristic curve). This can be used to construct a test with specified information loads across the ability scale by selecting items to achieve those specifications.

To review all of the psychometric considerations made throughout the issue would be beyond the charge of this review. Instead, I see two areas as being most helpful as we move forward in the design and use of strong AA-MAS. First, there is a fair amount of discussion in the field of educational measurement regarding validity frameworks, what those frameworks should include, and how they should be used. Second, perhaps there is not enough attention paid to test score quality, as it is the test score on which inferences are based.

CONSIDERATIONS FOR A PSYCHOMETRIC FRAMEWORK FOR AA-MAS

In the work on AA-MAS in this issue and elsewhere, psychometrics have focused on the importance of validity evidence. Validity evidence in an assessment system is perhaps local, at the point at which inferences and uses of assessment results are meaningful and appropriate. As such, a framework for AA-MAS validation and technical quality assurance must be local and focus on the immediate and future needs of the assessment program for students, families, schools, and states.

Consistent with the measurement field's current thinking about validity and technical quality more generally, every aspect of an assessment, from conceptualization and design to administration, reporting, and use, requires quality evidence. Federal guidance requires AA-MAS to meet the same technical quality standards as general assessments, as they should. Alternate assessments are important tools for accountability, program evaluation, and decision making regarding student achievement. Because of the unique nature of alternate assessments, the use of the same psychometric criteria may not be appropriate; however, there is no reason why alternate assessments need to result in lower technical quality. It is important to support the use of alternate assessments with the highest level of technical quality and psychometric standards, perhaps even beyond that achieved by general assessments, particularly for eligible students. The goal is to provide eligible students with test scores that are of higher quality than they would be from the general assessment. This may require small-scale experimental designs, sampling eligible and noneligible students to take both versions of the assessment.

Because of the high-stakes nature of alternate assessments, their unique characteristics, their relative infancy in the field of educational assessment, and the unique learning needs of eligible students, much of the effort to establish technical quality has focused on content and consequences.

Content is clearly an important characteristic, because for most achievement assessments, the primary inference is about content knowledge and skills—content-related validity evidence is a core concern. And to some extent, the consequential-basis of validity has been a focus in evidence gathering activities—perhaps because of the unique learning needs and curriculum of eligible students. Although these are important inferences and provide appropriate evidence for the interpretation and use of test results, the technical quality of the test scores themselves must not be overlooked.

Psychometric evidence regarding the quality of scores can be gathered in many ways, including distractor quality (if multiple-choice), item quality, and test score quality. Distractor quality is typically identified through two metrics, proportion selecting the distractor and the distractor-total point-biserial correlation (a discrimination index for distractors). To be an effective distractor, each distractor should be selected approximately at the same rate (as each distractor should be plausible and potentially contain useful information regarding misconceptions or common errors), but probably at least by 5% of the respondents (given the overall difficulty of the item). Also, selection of a distractor should be negatively correlated with the total score (associated with lower total scores), where the point-biserial correlation should be less than –.20. In addition, IRT can provide even stronger evidence of distractor functioning, through a differential item functioning model for distractors, addressing differential distractor functioning across important subgroups (Green, Crone, & Folk, 1989; Middleton & Cahalan Laitusis, 2007). Qualitative evidence is equally useful when examining distractors, including evidence of cognitive aspects of distractors through think-aloud methodologies: Are the distractors eliciting the kinds of cognition we anticipated and value?

Psychometric quality evidence at the item level includes information about item difficulty, item discrimination, and in an IRT model, item location and item fit. Pragmatically, we need items to cover a range of difficulty, including very easy items to motivate students and more difficult items to challenge students. But there are two concerns regarding this common guidance. First, classical item difficult (proportion correct) is relative, depending on the ability level of the responding students. In classical analyses, maximum efficiency is obtained when items are at about mid-difficulty, somewhere between chance level and 1.0. The current guidance that items may be "easier" is also relative, suggesting that for eligible students (and by implication, all students), items on the AA-MAS may be easier than items on the regular assessment, which will result in higher scores. Second, it is more important to be true to the intended test content and depth of knowledge (or cognitive task of items), because the most useful inferences from achievement tests are about subject matter knowledge and skill.

Item discrimination (point-biserial correlations) should be positive and greater than .30. Discrimination is naturally limited for the most difficult and the easiest items—as it is a correlation metric and depends on variability. Similarly, when students are guessing, there is no relation between item score and total score (discrimination goes to 0).

However, IRT presents a more practical model where item location (IRT difficulty) can be matched to the location of persons on the ability scale. Another common strategy for the location of items attempts to select enough items at the cut-point location so that greater precision is obtained at the critical points on the ability scale (the decision points). These are the principles that make computer adaptive testing so effective. IRT models provide a test of item fit to the model—that is, whether responses to an item fit the IRT model, such that individuals with lower ability tend to get the item incorrect more so than individuals with higher ability. As ability

increases, the probability of correctly responding increases; if not, item responses are likely influenced by construct-irrelevant factors, and the item fit indices indicate misfit. Again, methods such as think-alouds can contribute heavily here.

Downing and Haladyna (1997) provided guidance for gathering item validity evidence, noting the importance of supporting score meaning at the item level. Among their advice, they suggested gathering evidence about (a) item writer training (documentation of training materials and methods), (b) item-writing principles used (evidence of compliance with rules and item review processes), (c) cognitive behavior and depth of knowledge (the cognitive classification system and documentation of its use), (d) item editing procedures (including the credentials of editors and results of item review), (e) item tryout results (including item field tests, cognitive lab, and experimental test results), and (f) key validation (documentation of the process to verify the key). The TAMI can accomplish much of this (Kettler et al., 2009/this issue).

Finally, test score quality is important to establish because it is the test score that is used as the basis for inferences about levels of achievement. Test score quality is often characterized in terms of precision and accuracy. Precision is a matter of score consistency or score reliability and accuracy is a matter of score validity or hitting the target of measurement (what we intended to measure). Regarding precision, we typically employ indices of reliability and the standard error of measurement, from which we create score intervals acknowledging the imprecision of measurement.

Rethinking Reliability

The most commonly reported reliability coefficient is alpha. However, coefficient alpha assumes essentially tau-equivalent measurement, an assumption regarding the nature of item true scores and error scores that is rarely tested. When item standard deviations vary a great deal (suggesting that some items are measured on a different scale), when there are a relatively small (or smaller in comparison) number of items, and when there are multiple response formats, we likely violate the assumption of tau-equivalence, where alpha underestimates reliability. A more accurate estimate can be obtained through the use of a more appropriate measurement model, which matches the measurement properties of the test (e.g., a congeneric model that assumes each item measures the same construct, but possibly with different scales and different levels of precision). Graham (2006) reviewed procedures to obtain model-based estimates of reliability in a structural equation model approach that estimates the most appropriate measurement model for a given test.

Cronbach (2004) and others have argued that alpha is not a particularly useful or appropriate estimate of reliability, given its hypersensitivity to group variability and numbers of items (associated with the assumption of tau-equivalence). A more comprehensive and sensitive estimate of measurement error is obtained through Generalizability Theory. Generalizability Theory allows us to partition measurement error to multiple sources of variation introduced because of the nature of the measurement procedure, which results in the most flexible means for estimating a wide range of reliability coefficients (including alpha).

Rethinking Construct Validity as a Unifying Framework

It is important to target the collection of validity evidence to defend the most important claims we hope to make from the results of an AA-MAS to defend the accuracy of measurement. However, it is not clear that a focus on construct validity (in its unitary conceptualization) is helpful. Messick (1989), following the lead of Cronbach and Meehl (1955) and others, argued for a unified model of validity, under the umbrella of construct validation. Many measurement specialists have since argued against the daunting demands of construct validation as being beyond the capacity of the typical test developer as well as test takers and test users. Even Cronbach (1989), 34 years following his seminal article on construct validity, apologized for advancing such a limiting view of test score meaning: "It was pretentious to dress up our immature science in positivist language; and it was self-defeating to say that a construct not part of a nomological network is not scientifically admissible" (p. 159). Some have argued for the primacy of content-related evidence for achievement tests, because the target of measurement is subject matter knowledge and skill (e.g., Lissitz & Samuelsen, 2007). Others point to the important contributions of cognitive psychologists and suggest a stronger focus on score meaning (e.g., Gorin, 2007). Some argue for a more pragmatic approach, one that addresses the claims and assumptions made in score interpretation and use (e.g., Kane, 2002).

In a comprehensive modern perspective on validity, all evidence that can be gathered to bear on test score interpretation is useful, including reliability evidence, content evidence, response processes, the internal structure of the test, and relations to other variables (elements mostly included in the *Standards*). There remains significant debate regarding the role of consequences and the degree to which they bear on the accuracy of test scores. Scriven (2002) for one, argued that issues related to test score use are important but are more about the utility of test scores rather than the meaning of test scores. Nevertheless, evidence that supports inferences and uses of test scores (to the extent that uses are supported by the meaningfulness and appropriateness of a particular test score) is important to document and report, particularly in the context of high-stakes testing, regardless of whether it is gathered under the auspices of validity.

For the purposes of AA-MAS, there are federal guidelines regarding the basic validity-related evidence that must be gathered; the authors in this volume have contributed to our understanding of these guidelines. But for schools, educators, families, and students, perhaps there are more basic requests regarding the information value from AA-MAS, including a clearer understanding of the meaning of scores and descriptions of the cognitive performance producing those scores. Although the current demands on AA-MAS are federal, the real importance of these assessments in the lives of students is local.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Cronbach, L. J. (1989). Construct validity after thirty years. In R.L. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391–418.

Cronbach, L. J., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Downing, S. M., & Haladyna, T.M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*, 61–82.

Egan, K. L., Ferrara, S., Schneider, M. C., & Barton, K. E. (2009/this issue). Writing performance level descriptors and setting performance standards for assessments of modified achievement standards: The role of innovation and importance of following conventional practice. *Peabody Journal of Education, 84,* 552–577.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*, 456–462.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement, 66*, 930–944.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26*, 147–160.

Kane, M. T. (2002). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

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

Lazarus, S. S., & Thurlow, M. L. (2009/this issue). The changing landscape of alternate assessments based on modified academic achievement standards: An analysis of early adopters of AA-MASs. *Peabody Journal of Education, 84,* 496–510.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36,* 437–448.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Middleton, K., & Cahalan Laitusis, C. (2007). *Examining test items for differential distractor functioning among students with learning disabilities* (ETS Research Rep. No. 07-43). Princeton, NJ: Educational Testing Service. Retrieved June 10, 2009, from http://www.ets.org/Media/Research/pdf/RR-07-43.pdf

Palmer, P. W. (2009/this issue). State perspectives on implementing, or choosing not to implement, an alternate assessment based on modified academic achievement standards. *Peabody Journal of Education, 84,* 578–584.

Roach, A. T., Chilungu, E. N., LaSalle, T. P., Talapatra, D., Vignieri, M. J., & Kurz, A. (2009/this issue). Opportunities and options for facilitating and evaluating access to the general curriculum for students with disabilities. *Peabody Journal of Education, 84,* 511–528.

Scriven, M. (2002). Assessing six assumptions in assessment. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 255–275). Mahwah, NJ: Erlbaum.

Zigmond, N., & Kloo, A. (2009/this issue). The "two percent students": Considerations and consequences of eligibility decisions. *Peabody Journal of Education, 84,* 478–495.