Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal.

Volume 10 Number 10, August 2005 ISSN 1531-7714

Technical Documentation for Alternate Assessments

William D. Schafer
University of Maryland

The three usual criteria for assessments from the psychometric perspective - reliability, validity, and utility - are discussed in this paper in the context of alternate assessments that are individualized for students with severe cognitive disabilities. Possible sources of data for documentation of the technical quality of alternate assessments are discussed. Some suggestions for developing alternate assessments are presented.

Alternate assessments can arise from a need to represent, in a broad (e.g., statewide) assessment and accountability system, students with the most significant cognitive disabilities, whose levels of cognitive functioning are below that needed for instruction and assessment using the content and achievement standards and/or formats expected of students in the general (regular) instructional program. Students who participate in alternate assessments are expected to be rare (i.e., a maximum of 1% by 2014, as implied by current federal guidelines) and expectations are often tailored to their individual needs based on discussions of committees (e.g., an IEP, or Individualized Education Program, or similar team) or other means (e.g., teacher judgment). What implications may exist upon psychometric criteria of individualization of assessment activities, content standards, or achievement standards have not yet been explored except in specific examples (e.g., Almond & Bechard, 2005; Almond, Filbin, & Bechard, 2005), which have attempted to apply traditional methods for technical documentation.

A state's alternate assessments may be (but are not required to be) intended to allow judgments about student achievement based on alternate achievement standards. These standards must exhibit three characteristics, according to the United States Department of Education Peer Review Guidance (USED, 2004). They "must be aligned with the state's academic content standards (i.e., include knowledge and skills that link to grade-level expectations" (p. 15). They "must promote access to the general curriculum" (p. 15). They "must reflect professional judgment of the highest learning standards possible for the group of students with the most significant cognitive disabilities" (p. 15).

Using the results of alternate assessments, students are assigned to achievement levels. Placements are made for a variety of reasons, among them to allow evaluation of the instructional programs that the students receive and to report student progress to parents. According to current federal guidelines, the achievement levels that result from alternate assessments are to parallel those used in the regular assessment program (e.g., basic, proficient, advanced) for the content areas tested in the regular program (e.g., reading and math).

Each student who is tested using alternate assessments must nevertheless be allowed access to the content of the regular educational program. This criterion is commonly met either through extending the regular content standards, perhaps to the level of access skills demonstrated in the contexts of the content areas, or through reducing the complexity of the grade-level objectives that appear in the regular educational program to a similarly foundational level (Bechard, 2005). Another approach is to modify the assessment format but to present the student with the same tasks as the regular assessment (but since the modifications may alter the construct itself, it is not clear that this approach in fact differs from the other two).

Some assessments consist of modified standardized assessments that are intended to cover the same content and achievement standards as the regular assessments. Others are developed to cover an alternate set of content and achievement standards with standardized instruments. There may be no need to consider the problem of psychometric implications for either of these types of assessments since they do not individualize, and thus the same psychometric standards that apply to the regular assessments (and cut scores) apply also to them. In the material that follows, it is assumed that there exist some element(s) of individualization of content standards, achievement standards, or assessment activities (e.g., domain, prompts, administration, scoring, cut scores) in order to accomplish an alternate assessment.

Alternate assessments present some unique challenges for traditional psychometric practices. Assessments in the regular program are given to students under standardized conditions, cover identical content, and result in scores that have similar meaning throughout the achievement range. But some alternate assessments that are individualized no longer have these three characteristics. Students taking those alternate assessments may not have received instruction toward the same domain expectations (i.e., content-cognition combinations) as each other and/or their performance expectations may not be the same from student to student. Students may receive assessments that differ from each other in either content or complexity (or both), as well as in their achievement expectations, and that are presented in ways that match the individualized instructional conditions that students are supposed to have received. Unlike traditional assessments, therefore, alternate assessments do not necessarily have consistent meaning throughout the achievement scale and different students may be held to different standards for achievement level assignments. Some implications of these dimensions of individualization of assessments for psychometric evaluation in terms of reliability, validity, and utility are explored below.

RELIABILITY

Variance-Ratio Approach

One may conceptualize reliability as the degree to which variation in scores are caused by differences in those measured, which leads to a variance-ratio definition. This approach is most appropriate for assessments that are interpreted when scores earned by examinees are compared with scores earned by others (i.e., norm referencing) and applies well also to assessments that are interpreted in comparison with fixed cut-points. But when student results have different meaning across the learning domain, such as alternate assessments, conceptualizing reliability as a ratio of true and observed variances seems inappropriate. Not only does this approach treat scores as comparable with each other, it also ignores alternate assessments' changing referents (e.g., individualized cut scores). Internal consistency results or indeed any means of estimating reliability that evaluates relationships (e.g., correlations) across examinees is not likely to be very helpful since students are not each others' referents (i.e., identical scores may have different outcomes, or interpretations, for different examinees).

Consistency Approach

Another approach to characterizing reliability uses the concept of consistency of outcomes. Although it also applies to norm-referenced assessments, this latter approach seems most appropriate as it is applied to judgments of

achievement level assignments because students are being compared with fixed, consistent criteria rather than norms (i.e., rather than each other). The criterion of decision consistency is commonly applied to such judgments in the regular assessment program and might be thought to be appropriate for alternate assessments since achievement level assignments are also being made. However, the same achievement level that may be described as "proficient" for one student may be described as "not proficient" for another when made based on different individualized assessments. Therefore, the concept of decision consistency across students is not necessarily appropriate for alternate assessments.

Since alternate assessments are individualized, it seems most natural to study consistency for individual students. Parenthetically, this approach is compatible with a modern view of measurement error as conditional on degree of student achievement.

Normally, assessment error is only evaluated with respect to the result that is earned on an assessment. However, for alternate assessments there are two fundamental elements to be considered: the observations (e.g., score or scores) of the student and the referent (e.g., the cut score or scores) that the observations are compared with. Reliability of the latter is normally unimportant (or considered separately) in regular assessments since cut scores are set by panels and are used for everyone. But in alternate assessments there may be varied expectations for the broad range of student abilities that any assessment must be developed for. Students of the same age are not always studying toward the same objectives. Thus, the reliability of the reference (cut) score is also important to address. Some way must be found to incorporate both of these elements, the student score and its referent, into a reliability analysis when we focus on achievement level judgments for alternate assessments.

Evaluating Consistency

Consistency of alternate assessment outcomes and of performance expectations (referents for student results) across tasks, occasions, and scorers as appropriate should be reasonable assessment quality expectations. Consistency of individual student results across different tasks, for different occasions, and if applicable, among scorers all can and arguably should be studied empirically, either through ongoing studies that compare independent collections of evidence or through special studies. Consistency of referents can also be studied by focusing on how they are developed and then evaluating the extent to which independent developments of them differ from each other.

Consistency across evidence types is another issue. Common evidence types are videotapes of student performance, written performance products such as worksheets, and teacher or examiner descriptions of performance. These may not yield consistent results when the same performance is rated in these different ways. Again, differences in consistency might be studied within and across tasks, occasions, and scorers as possible moderator variables (i.e., variables that alter the consistency findings).

The consistency of the process for comparing student performance with achievement level expectations should also be documented. Consistent results and consistent referents are not adequate if the method of comparing them is capricious. This could be studied empirically through independent replications.

Finally, there must be an adequate quantity of evidence generated. It may be helpful to study consistency of achievement level judgments across subsets of the assessment tasks in order to generate a classification consistency statistic for individual examinees (this could be studied annually for samples of students, perhaps chosen to represent a range of assessment challenges). Another approach that may be helpful was suggested by Smith (2003), who recommended placing confidence intervals around proportion-correct scores for individual students using the usual standard error of a proportion when sampling from a binomial

(dichotomous) population. Finally, a re-sampling strategy may be useful (e.g., bootstrapping).

VALIDITY

Primacy of Interpretations

A modern view of validity holds that it exists to the extent that we can justify an interpretation derived from the test score. In order to study validity, then, the nature of the interpretation, itself, needs to be considered.

Because validity is defined only for making inferences, validity research is best begun by explicitly stating each inference (intended and perhaps unintended) that is being studied. This point can be applied to all assessments. Validity for each inference should be considered separately, though, since some may be more justifiable than others. Questions that need to be addressed in validity research may be organized within each inference by possible sources of threat to the validity hypothesis (or assumption) that supports drawing the inference.

Contextualizing Inferences

Interpretations need to be contextualized, and that is commonly done through referencing. Some interpretations reference the student, such as by describing what the student can do or how much the student has grown. These normally require a well-understood domain that has some universally agreed-upon understandings about success, as exist in such examples as landing an airplane or performing an appendectomy or running a marathon.

Other interpretations reference points on a score distribution, such as comparisons with norms or cut scores. Professional certification decisions are an example, as are achievement level judgments in regular assessment programs in schools.

Still other interpretations reference other earned scores. Selection decisions in competitive environments are an example.

On some alternate assessments, outcomes are related to the degree of support (e.g., cuing of, or assisting the student toward, the desired response; sometimes called scaffolding) necessary to achieve a correct response where the same task is presented to all students in a given grade and developmental level. In those assessments, degree of support could be combined with the task descriptions to define individualized achievement targets for students taking alternate assessments.

Some alternate assessment programs use expanded benchmarks, extending each of the skills and concepts in the regular curriculum through decreasing degrees of complexity to a basic, or foundational, level (Bechard, 2005). The assessment development process may move from state standards to expanded benchmarks to the assessments. Essentially, the expanded benchmarks play the role of curriculum standards for students taking alternate assessments. Students may receive assessments that demand a level of complexity similar to the regular assessment program at one extreme to a level at the other extreme that makes any connection between the assessment and the regular content standards seem quite tenuous.

Other alternate assessment programs focus on attaining prerequisites skills and concepts needed to achieve in the regular curriculum. In these programs, each student's instructional program, and thus his or her set of individual learning expectations, is based on his or her prior achievements and ability to progress toward the skills and concepts in the regular curriculum, using them as ultimate, but not necessarily yearly goals for attainment. Students may receive assessments that cover almost all of the concepts and skills in the regular assessment program (and some may even take some of the regular assessments) at one extreme to concepts and skills at the other extreme that are mastered by students in the regular program well before they enter school.

These and likely any other approach to alternate assessment present similar challenges in establishing validity. In any case, students have received individualized instructional programs that can require domain-level adjustments to the assessments. Presentation and response conditions

need to reflect instruction. Performance criteria for assignments to achievement levels may need to be developed to reflect individual learning goals. These aspects of individualization have implications for the questions that arise in validity research. They are explored here as possible threats to the validity of the assessment program, although it should be remembered that for individual assessments, validity will likely vary from student-to-student and like consistency, may need to be studied for representative samples of examinees. Each threat is identified according to the assumption of the alternate assessment that supports the inference implied by assignment of students to achievement levels.

Adequacy of the Alternate Assessment Learning Domains

The learning domains of the alternate assessments must allow access to the full range of the regular assessment content domains and achievement standards. Most professionals agree that placement into a program that implies participation in alternate assessments should not be final. It should be possible for a student to achieve to a level where success in the regular assessment program is a reasonable expectation. Whether through content (knowledge), process (cognitive activities) or both, evidence of pathways from the domain of any alternate assessment to that of the regular assessment should be developed in order to study this threat.

One unique threat to validity of alternate assessments is breadth of learning domain coverage. The question to be addressed is whether teachers (and other stakeholders such as parents) would agree that all their alternate-assessment-eligible students can appropriately be placed somewhere among the eligible content and process domain that may appear on the assessments. If not, then even the expanded standards are not sufficient to cover the full range of the learning domain as it should be realized for all students. A survey of teachers and other relevant stakeholders could address this issue.

Equivalence with the Student's Intended Curriculum and the Fundamental Accountability Mission

A universal underlying goal exists for all statewide accountability testing: the Fundamental Accountability Mission is that every student should be tested with tests that cover what that student is supposed to be learning (Schafer, 2004). For example, one context in which the mission often arises is the area of accommodations. USED (2004) and many statewide programs insist that accommodations that exist for an assessment must have been implemented for instruction; otherwise, the assessment will not match the instruction. Stated another way, this implies that accommodations used in instruction must be implemented for an assessment in order for the assessment to be valid.

The Fundamental Accountability Mission to test all students on what they are supposed to be learning implies that it should be possible to use tests to guide instruction. Like for accommodations, in the case of alternate assessments, this has implications for standardization. It may not be the case that any set of standard conditions can equivalently represent the instruction that is supposed to be given to all students. For an alternate assessment, it may actually be inappropriate to emphasize standard administration conditions for all students because they may not be consistent with students' intended instructional experiences.

Another implication of the Fundamental Accountability Mission for alternate assessments is that the instructional domain may not be oriented toward consistent display of achievement, instead opting for performance to demonstrate achievement only at optimal times for the student. This could imply that an on-demand assessment is not as valid as an assessment that allows collection of evidence when it is available.

Specifications of how data are to be collected and fidelity to specifications are validity issues whether or not observations are developed over time as in portfolios or only during on-demand testing sessions. While including transitory

demonstrations of achievement may be a challenge for on-demand approaches, ensuring that demonstrations represent actual accomplishments is similarly a challenge for portfolio approaches.

Applicability of the Achievement Level Criteria to the Assessment System

In order to make assignments to performance levels, student results need to be compared with criteria. Whatever system of capturing student results is used (e.g., a score across each content domain), the criteria need to be expressed in the same system in order to facilitate the comparison.

Consistency between Expectations and the Instructional Domain

The question of whether student achievement expectations are aligned with the instructional domain appropriate for each individual student must be considered. This may be addressed for groups of students (e.g., defined by age or grade levels) or at the individual level, such as through an IEP or similar document. Whichever approach is adopted, the decision to do it that way requires justification. Then, the process needs to be described and evaluated. Independent review by teacher committees might be a way to gather the necessary evidence. When expectations are individualized, alignment should be evaluated at the individual level; for feasibility, a sampling approach to gathering this evidence might be satisfactory.

Domain Coverage

Both the student result and the criterion should capture the important aspects of performance and be free of invalid sources of variance. This is normally straightforward for regular assessments; studies are commonly undertaken to generate convergent and discriminant evidence of validity for student scores. Criteria (e.g., cut scores) are developed through standard-setting studies that rely on the same evidence are further designed to establish content evidence for validity. Neither is necessarily straightforward for alternate assessments because the same score for different students may have different meanings depending on the students' instructional goals, and the criteria are not necessarily constant across students. At first, perhaps the best approach may be to generate outside review by stakeholders and technical experts. The reviewers could be asked to identify threats to the breadth and depth of the assessments and their referents and to identify possible sources of artificially high or low scores. Studies could then be designed to evaluate the credibility of these threats for representative samples of students.

Degree of Challenge

The process by which student content and performance expectations are generated should result in an appropriate level of challenge. When an achievement level judgment is reached, it should be possible to conclude that the student has achieved at a level that represents, for him or her, a worthy outcome. If the expectations are set too narrow or too broad, or too high or too low, neither the student nor his or her educational program will be validly assessed. This is a criterion that arises in any assessment but is especially difficult to document for alternate assessments due to individualization. Evaluating expectations using various stakeholders for a sample of students selected from throughout the achievement range could provide the needed evidence.

Alignment between Domains and Test Activities

The assessment activities on an alternate assessment need to capture (or sample) all a student must do to perform well in the intended curriculum. Do the tasks demand all relevant aspects of student production? Is the richness of the student's learning domain represented adequately in the alternate assessment? Do the tasks represent an appropriate breadth of contexts and degree of independence? These questions could begin to be evaluated using stakeholder groups.

Fairness

All assessments must be fair to all the relevant instructional programs being evaluated. In the regular assessment program, fairness across programs can be met through publication at the state level of a description of the domain of the

assessment. Teachers may be expected to provide instruction over that curriculum as a minimum. For alternate assessments, publication of the domain that is sampled may not be adequate because it is sampled differently for different students. Implications for fairness of the sampling process for individual students need to be evaluated. An operational definition of the sampling process should be developed and used as evidence of fairness. Then, stakeholders might address questions such as whether the sampling process develops sufficient breadth, depth, and challenge across content strands within programs.

UTILITY

Utility has not been very well defined in psychometric literature. Often, its use is as a catch-all for criteria that do not fit under either reliability or validity. In alternate assessments, one such criterion has to do with the effects of the assessment program on instruction.

In order to be useful to teachers, student expectations must be expressed in terms that can guide instruction. A need therefore exists for criteria for both domain and performance expectations so that they can support appropriate assessments that are aligned to them. One issue that should be addressed is whether the expectations are sufficient to convey the assessment limits appropriate for each student at the beginning of instruction in terms of content-cognition combinations that may be represented (Schafer & Moody, 2004). Put another way, the teachers should know at the beginning of their instructional activities, and from the content expectations (e.g., benchmarks or strand-level expectations in an expanded set of content standards) what is and is not fair game for the eventual assessments that will be used to judge the effectiveness of their work. Only then can the assessment system deliver on its promise both to guide and to assess instruction, thereby capitalizing on the Fundamental Accountability Mission.

SUMMARY AND RECOMMENDATIONS

Eight major points seem central to work on developing and documenting the psychometric quality of alternate assessments:

1. Test every student on what he or she is supposed to be learning. That should be the primary focus of any alignment study (or process for ensuring alignment). In other words, remember the Fundamental Accountability Mission.

2. State each inference that is to be supported specifically in a psychometric evaluation of the validity of any assessment or assessment program. Elaborate each statement to address all relevant questions stakeholders may have, such as instructional implications, implications for certification of achievement, and institutional implications. Collect validity evidence for each inference separately in such a way to evaluate the assumptions that are necessary for the inference.

3. The most important inference to focus on for statewide assessments is that of assignments to achievement levels. This is also true for alternate assessments. Other inferences may also be important depending on the context. Unintended inferences may need study too, even if only to show that they are invalid.

4. For all assessments, and especially for alternate assessments, evaluate the reliability and validity for the student's assessment results, for their referents, and for the process by which they are compared.

5. Assessments of reliability that focus on evidence across examinees (e.g., variance components, correlations) are probably not going to be useful in studying alternate assessments. Instead, document the consistency of the score and of its referent independently, as well as the process of making the comparison, and focus on the process as it occurs at the individual student level.

6. The student's instructional domain must be consistent with criteria for alternate achievement standards and the student's

alternate assessment must be aligned with that instructional domain. Researchable aspects of these criteria are: (1) is the breadth of allowable individual content expectations sufficient to include the appropriate instructional domain for each student, (2) does the instructional domain provide access to all aspects of the regular curriculum, (3) does the student's alternate assessment align with the student's instructional domain, and (4) do the student's performance expectations represent the highest possible achievement standards that are consistent with the student's instructional domain.

7. On-demand assessments and their associated need for standardization may not be crucial for alternate assessments. Indeed, it may even be best to evaluate maximal rather than typical student performance. Decisions about whether to require standardized, on-demand data collection or to generate data less formally might best be made at the individual student level (i.e., individualized).

8. Include within the criterion of utility, how well the assessment system provides explicit instructional focus for the teacher. Consider this recommendation as a companion to the Fundamental Accountability Mission.

REFERENCES

Almond, P. & Bechard, S. (2005). Alignment of two performance-based alternate assessments with combined content standards from eight states through expanded benchmarks. Presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Almond, P., Filbin, J, & Bechard, S. (2005). Reliability in two types of alternate assessments for students with significant cognitive disabilities: Administration fidelity, rater agreement, and internal consistency. Presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Bechard, S. (2005). Developing alternate assessments using expanded benchmarks from a nine state consensus framework in reading, writing, mathematics, and science. Presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Schafer, W. D. (2004). Review [of Tindal, G. & Haladyna, T. M. (Eds.), *Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*. Mahwah, NJ: Lawrence Erlbaum Associates]. Contemporary Psychology, 49, 622-625.

Schafer, W. D. & Moody, M. (2004). Designing accountability assessments for teaching. *Practical Assessment, Research & Evaluation*, 9(14).

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice,* 22(4), 26-33.

United States Department of Education (2004). Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001. Washington, D. C.: U. S. Department of Education.

Acknowledgements

This paper was partially funded by the Maryland State Department of Education (MSDE) through the Maryland Assessment Research Center for Education Success (MARCES) at the University of Maryland. The author is indebted to Sue Bechard, Beth Cipoletti, Michael Harmon, Robert Lissitz, and Heather Mann for helpful comments on earlier drafts. The opinions expressed are those of the author and not necessarily those of MARCES, MSDE, or any of the individuals who shared their comments during

preparation.

Citation

Schafer, William D. (2005). Technical Documentation for Alternate Assessments. *Practical Assessment Research & Evaluation*, 10(10). Available online: http://pareonline.net/getvn.asp?v=10&n=10

Author

William D. Schafer is Affiliated Professor (Emeritus), Maryland Assessment Research Center for Education Success, Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD. He specializes in assessment and accountability.