Dealing with Flexibility in Assessments for Students with Significant Cognitive Disabilities

Brian Gong & Scott Marion[1]
National Center for the Improvement of Educational Assessment, Inc.
April 13, 2006

Introduction

Dealing with flexibility—or its converse, the extent of standardization—is fundamental to alignment, assessment design, and interpretation of results in fully inclusive assessment systems. Highly standardized tests make it easier to compare (performances, students, and schools) across time and to common standards because certain conditions are met that (ostensibly) reduce the irrelevant variation and support stronger inferences for interpretation. Alternate assessments on alternate achievement standards—and the corresponding instruction for these students—have come from a highly individualized tradition in which such comparisons have not been the focus. Alternate assessment (and instruction) is moving more firmly into a standards-based accountability world, due in large part to the *Individuals with Disabilities Education Act Amendments of 1997* and the reauthorization of 2004 (*IDEA*) and the *No Child Left Behind Act of 2001 (NCLB).* There is no question that NCLB has ratcheted up the pressure on states to compare validly the scores derived from alternate assessments to common content and achievement standards. NCLB aside, this is an appropriate and promising time—from policy, research, and practitioner perspectives—to reflect more deeply on what variations are good and tolerable and what variability is to be minimized in assessments for students with significant cognitive disabilities.

We want to emphasize that this paper is limited to alternate assessments on alternate achievement standards for students with the most significant cognitive disabilities. This group of students generally comprise less than one percent of the total student population and face the most profound learning challenges. Other forms of alternate assessments, particularly those based on grade level achievement standards that offer a different format for measuring the "same" content and skills found on the general assessment, raise interesting issues in terms of flexibility and standardization, as well. However, discussions of these types of assessments are not included in this paper.

These intended as well as unintended assessment variations create challenges for inferring what any individual student knows and is able to do. Evaluating the technical quality of the entire assessment system, including the alternate assessment, raises this challenge to the next level. Technical evaluations of general education assessments are dependent on aggregating large amounts of standardized data to describe such things as item quality and test reliability. Alternate assessments make this task difficult on both accounts: there are rarely large amounts of data and much of it would

[1] This work is supported, in part, by the U.S. Department of Education, Office of Elementary and Secondary Education, Enhanced Assessment Instruments Grant No. S368A040004. Opinions expressed do not necessarily reflect those of the U.S. Department of Education or Offices within it.
We acknowledge the many important contributions from the expert panels of the New Hampshire Enhanced Assessment Initiative and the National Alternate Assessment Center. Special thanks are due to Rachel Quenemoen, Rich Hill, Martha Thurlow, Jacqui Kearns, Sue Bechard, and Jim Shriner for extensive comments and feedback, which helped improve this paper. Of course, any errors and shortcomings are ours alone.

not be considered standardized by our traditional definitions and practices. In other words, these assessments rarely meet underlying assumptions necessary for use of many traditional technical evaluation methods.

Many of the challenges in evaluating the technical adequacy of alternate assessments stem from dealing with *intended variability*. More specifically, alternate assessments create difficult challenges because there is currently large flexibility in *targets* and goals in addition to flexibility in the methods and *means* used to assess. The flexibility in targets and goals in some states is due to the belief that this is the most appropriate way or even the only way to appropriately measure what these students know and can do. These states are concerned that if they do not allow flexibility in targets and goals, they might simply be measuring the extent of students' disabilities and not actual school learning. In cases where states limit the flexibility in goals and targets, they often adopt other areas of flexibility, citing the same beliefs of appropriateness. There is a long tradition in large-scale assessment on how to handle flexibility in means and methods that may be extended to the particular challenges posed by alternate assessments. However, less is known about how to deal with different targets, in large part because so much effort has gone into standardizing (reducing the flexibility) of targets.

Much of the discussion in the alternate assessment world regarding intended variability has focused on trying to associate the degree of standardization with specific forms of alternate assessment (e.g., portfolio, performance assessment, checklist). While this has a certain appeal, the simplicity of the categorization does not do justice to the types of variability found in alternate assessments (Quenemoen, Thompson, & Thurlow, 2003 – see Appendix A for excerpt to provide additional background information).

This project uses a validity lens to organize the evaluation of technical quality of alternate assessments and has benefited by drawing on the assessment triangle framework put forth in *Knowing What Students Know* (NRC, 2001). The triangle is a heuristic to describe the interactive relationship among three main components of an assessment: cognition or learning model supporting the assessment design, observations or means of collecting the assessment information, and interpretation or the methods for turning the observation data into useful score inferences. The interpretation vertex also includes methods by which we use to evaluate tests and test items. This paper focuses largely on the cognition and observation vertices of the assessment triangle, while other work from the project addresses the interpretation vertex more directly.

A key aspect of any validity evaluation is the specification of purposes of the assessment and a description of how the results will be used. Tolerability of flexibility or the converse, requirements for standardization in the assessment system, is largely dependent upon how the assessment scores will be used. We tolerate—even value—flexibility in classroom assessments when teachers are trying to determine how best to help students learn certain concepts, but this same flexibility is not usually acceptable to policy makers when holding school or students to high stakes. The alternate assessments discussed in this paper are all used as part of school accountability under NCLB and states' unique accountability systems. This would appear to raise the requirements for standardization. However, proficiency determination from the alternate assessments on alternate achievement standards is capped at one percent of the student population; with that inherent control built in, it can be argued that there is less need for standardization for alternate assessments. We return to this issue at the end of this paper.

We hope this paper provides a useful framework to inform discussion about flexibility in assessments for students with significant cognitive disabilities that will:

a. permit clarification of values and goals so fundamental policy decisions can be made regarding desired comparability,
b. support research and development work to improve assessment for this population of students, with the long-term goal that such assessment will support improved achievement,
c. help the discussion of alternate assessment approaches move beyond simply using nominal labels of familiar assessment formats (e.g., portfolio, performance, checklist or rating scales) and recognize that most alternate assessments are a blend of multiple formats with varying degrees of flexibility for different components of the system, and
d. assist in the evaluation of the technical quality of alternate assessment systems.

Enduring Issues in Alternate Assessment

Some commonly heard and sometimes contentious issues in alternate assessment are reflected in these statements:

- ✓ "These students can't be compared to regular education students; they can hardly be compared to each other because each student is so unique."
- ✓ "Without common, high expectations, these students will continue to be underserved educationally."
- ✓ "We can get a score on an assessment and include these students in accountability, but the assessment isn't even close to reflecting the same content and skills for which general education students are responsible."
- ✓ "We can assess these students on clearly identified grade-level standards, with specified relationships to the particular content and skills on which general education students are assessed."

The core challenge of these statements pivots around how much flexibility *should be* allowed, with trying to *control* how much flexibility is permitted or *interpreting results* when there is flexibility in key assessment aspects. It may be beneficial to articulate clearly what the possible sources of flexibility are, how they have been dealt with traditionally, and why they create difficulties for evaluating the technical quality in the realm of alternate assessment for students with significant cognitive disabilities.

Typical Ways of Dealing with Threats to Standardization in Assessments

In most large-scale assessment programs, standardization is maximized through policies and control processes for certain key aspects. Some key aspects are shown in Fig. 1. For example, t*he domain* to be assessed is usually fixed and specified as much as possible, often through a common set of content standards that list the target knowledge and skills. These content standards are often designated to be learned within certain time constraints, e.g., within grade 4, hence the common terms "grade level expectations." A test blueprint is used to specify the *content to be assessed* (a sample from the domain), its format, the cognitive complexity, the mix of items across the domain, the reporting units and categories, and so on. The test blueprint assumes and controls for certain commonality or standardization across test instruments and test forms, with the fundamental purpose of providing comparable testing experiences across students and testing occasions. The assessment tasks and test as a whole are subject to review in order to check, among other things, that the items and test function similarly (without unexpected variation) for individuals and subgroups (e.g., DIF). Tests are calibrated and scaled so that performance can be compared taking into account varying difficulty across items. Tests are equated to provide a basis for comparison across years, or so that test differences across years do not impede the interpretation of performance across years. Administration conditions are carefully specified. Variations of administration conditions—the well-known accommodations and modifications—are explicated because of fears that too much flexibility would threaten the validity of interpretation of scores. Large-scale assessment scoring is typically done under controlled conditions with stringent quality control checks to support the standardization of processes and the ability of scores to support a common interpretation. Machine scoring of multiple-choice responses is done in part for efficiency, and in part to make scoring "objective," i.e., reduce variability due to a range of human skill and judgment. When open-response tasks that require human scoring are used, strict protocols for maintaining standardization are employed. Evaluation criteria and rules, such as performance standards, are carefully developed and applied the same way to every performance as much as possible.

In summary, typical large-scale assessments involve a high degree of standardization of content. Where there is not standardization, there is high specification of process. Typically, large scale assessments standardize—or admit little flexibility by design—in all aspects of the testing process. With most state assessment, there is even the attempt to control administration conditions through required rules and training sessions.

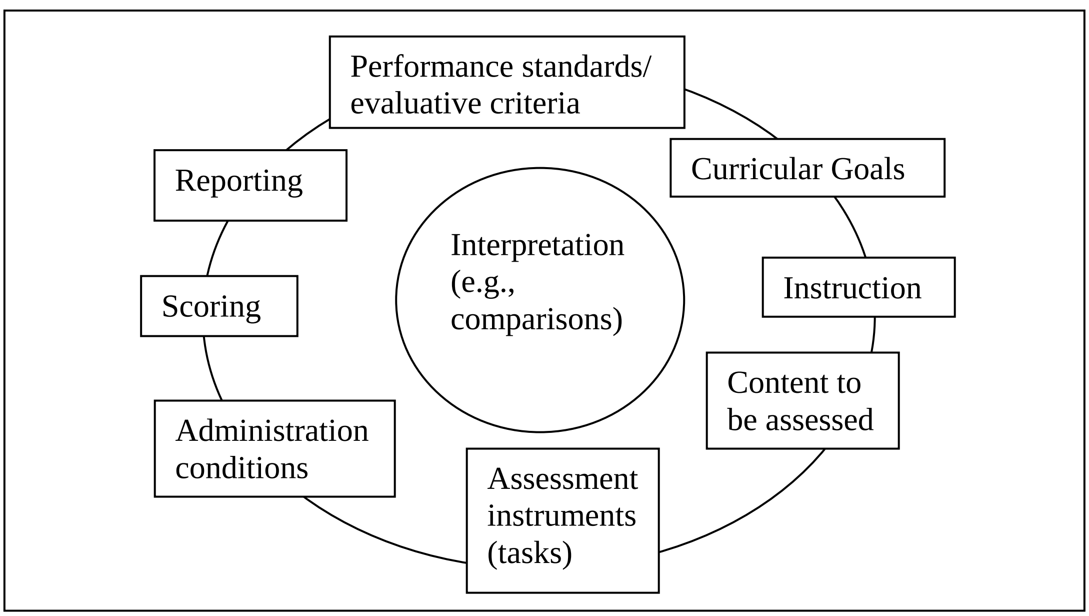

Figure 1. Typical instruction, assessment design, administration, and interpretation sequence.

For the majority of special education students participating in the same assessment as the general population, there is explicit and intense attention paid to any flexibility in the assessment. Most of the blocks are assumed to be held comparable without flexibility and any variation is scrutinized in the assessment instruments (tasks) and the administration conditions. For example, an assessment task that is modified for the visually or hearing impaired students must undergo extensive review during the item development, piloting, and psychometric analysis phases. The scrutiny over "accommodations" and "modifications" in the test administration conditions of how the test is presented, responded to, or other administrative conditions such as amount of time permitted illustrates how large-scale assessment abhors certain types of flexibility.

Characterizing Flexibility in Alternate Assessments

Alternate assessments present a particular challenge because they allow and even encourage considerably more flexibility, and in ways that are typically not done in large-scale assessments for the regular population. This flexibility is based, often justifiably so, on articulated concerns about the appropriateness of assessment methods for these students. Alternate assessments go beyond varying administration conditions to varying a number of other assessment dimensions. These variations are often combined and maximized for individuals, rather than minimized. The following section describes the types of and degree of flexibility expected/allowed in both general and alternate assessments along nine dimensions of the assessment and accountability system. This variability is then summarized in Table 1.

1. Flexibility in the curricular goals—the content and skills students are expected to learn during a particular time span (e.g., grade level)—between students at a point in time and over time
   - General assessments: Typical standards-based approaches fix common curricular goals for all students at a point in time, based on student enrollment at a particular grade level. States vary in whether they assume students will have been taught with a common curriculum or whether they assume that curricular variation does not matter[2]. General assessments may reflect different curricula and thereby variations in curricular goals (or vice versa) through mechanisms such as end-of-course exams, out-of-level testing, curricular-targeted exams (constructed by a teacher or selected by a local educational unit), and student choice in questions. In general assessments a developmental sequence for content and curricular goals is made explicit at some level of detail.
   - Alternate assessments for students with significant cognitive disabilities: There is no consensus on a common developmental sequence of academic content and skills for the population (Kleinert, Browder, & Towles-Reeves, 2005). Individual curricular goals vary widely in content focus, scope, depth, independence, and many other important dimensions. The fundamental value underlying this approach is that maximizing individual learning requires individualization of learning goals within the general frame of academic content. The individual situations of these learners have been so specific that interpreting results in a common standards-based framework has been viewed as essentially impossible and/or irrelevant.

2. Flexibility in the instruction (learning experiences)
   - General assessments: Except for some specific curricular programs, most state leaders expect that individual teachers will instantiate the curriculum and standards in different ways. This variation in the enacted curriculum has been deemed acceptable because most believe that the knowledge and skills to be tested should transfer to situations even if that was not how students

[2] We have argued in other places (Marion, 2004) that curricular differences are an important source of variability in student performance and will not address this issue here.

were taught the material[3]. In spite of the differences in instructional experiences, students from differing instructional pathways (assuming equal quality) are expected to have equal chances of success on the assessment.

- Alternate assessments for students with significant cognitive disabilities: Students participating in alternate assessments are expected to have instructional experiences tailored to their specific learning needs. There is little expectation that these instructional and learning experiences can transfer to common assessment situations unless this has been built intentionally into the instruction (see Kleinert et al., 2005). Therefore, in many cases, the assessments are intentionally focused on students' specific learning experiences[4].

3. Flexibility in the content standards chosen to be assessed
    - General assessments: With current large-scale, standards-based assessments, all students participating in the general assessment—with or without accommodations—are assessed on the same content standards for each grade level. As a matter of fact, almost all states now base student scores on the same common item set for all students at a particular grade level.
    - Alternate assessments for students with significant cognitive disabilities: Depending on the type of alternate assessment employed, the specific content on which each student is assessed may vary widely. In many cases, the student's teacher is expected to select indicators or tasks that relate to the target content standard and these indicators would intentionally vary by student. Some states specify the content standards/strands, and teachers choose the performance indicator, while in other states, the standards and the indicators are mandated and teachers choose the activity. This could be true for checklists, portfolios, body of evidence, and perhaps even performance-task approaches.

4. Flexibility in the methods/items used to assess
    - General assessments: With the exception of students receiving accommodations, all students participating in the general assessment experience the same item formats and usually the same items. Even when a portion of the assessment counting in students' score is part of a matrix sample, the item formats across forms are usually quite similar.
    - Alternate assessments for students with significant cognitive disabilities: States typically adopt one general method to assess all students participating in the alternate assessment system. However, the specific tasks/items used for each student may vary considerably, or not at all, depending on the state. Those systems that intend to minimize this variability do so by presenting the same tasks to each student, but varying the amount of assistance—an integral part of the item—the student receives from the test administrator. This assistance can take the form of a hierarchy of verbal or physical prompts to help the student produce a response. In other cases, assistance is relative to the content, particularly how much more information is provided about the item to scaffold the students' responses. The types of assistance are generally assumed to be fairly standardized (e.g., a reasonable level of procedural fidelity), but this assumption has been difficult to assure.

---

[3] Cognitive psychologists and others have pointed out that this is not necessarily a correct assumption, but it still defines most current practice (NRC, 2001).

[4] We recognize this is not necessarily the case with some of the more "standardized" performance-based alternate assessments.

5. Flexibility in the administration conditions
   - General assessments: While there is an intention to standardize the administration conditions across the state, most state assessment programs allow local educational officials some opportunity to determine the specific schedule and testing conditions within their schools (e.g., administering the test within classrooms or a large auditorium). Generally, local officials are not permitted to vary the order or minimum length of test sessions.
   - Alternate assessments for students with significant cognitive disabilities: Most alternate assessments are administered in a one-on-one setting and therefore, by definition, there is flexibility in the administration conditions. This is true even though there are general guidelines and rules to guide the test administration. These guidelines certainly help dampen the full range of possible flexibility in administration conditions, but the one-on-one nature of the administration, especially with the allowed task or prompting/scaffolding variability, makes it doubtful that truly standardized administration conditions exist.

6. Flexibility in the scoring
   - General assessments: Scoring variability is negligible with multiple-choice items, but is certainly a factor when open-ended items are hand scored. Even when hand scoring is done, it is usually conducted at central scoring facilities. There are often tight protocols and quality control checks established to monitor, control, and quantify this variability. As long as the variability remains within acceptable bounds and does not threaten year-to-year equating (not always addressed explicitly), it is tolerated on state assessments.
   - Alternate assessments for students with significant cognitive disabilities: The specific tasks are often scored by the teacher (or other test administrator), but can also be scored centrally, depending on the specific nature of the assessment. In fact, most of the states using portfolios or bodies of evidence have centralized, highly controlled scoring processes. When scored or rated (as with checklists) by individual teachers, additional variability is most likely introduced into the scoring process. The scoring is often dependent on the particular task and the degree of independence/assistance associated with the task, which can introduce unwanted variability into the scoring process.

7. Variance in the performance standards
   - General assessments: All students at a particular grade level are expected to be held to the same performance/achievement standards. This almost always involves some type of standard setting process enabling raw scores (or the scale score equivalent) to be converted into performance categories. Within any given year, the only variability in how the performance standards are applied to students in the general assessment is simply a function of measurement error. Across years, additional variability in how the performance standards are applied may be introduced as a result of equating error.
   - Alternate assessments for students with significant cognitive disabilities: States are permitted to introduce systematic variability into the performance standards for alternate assessments because they are permitted under NCLB to establish multiple performance standards for alternate assessments. Most states employ a single set of alternate assessment performance standards at each grade level and evaluate all alternate assessment scores against these standards. This might be a case where more variability may make more educational sense, but most states have chosen to employ a common single set of standards.

8. Flexibility in the interpretation and reporting
    - General assessments: Most states use a common, fixed reporting shell for all students (or schools, depending on the aggregation level) at a particular grade level for each subject tested.
    - Alternate assessments for students with significant cognitive disabilities: Again, most states use a common, fixed reporting shell[5] for all students participating in the alternate assessment, but there will likely be some variability in the particular cells filled in on each student's report, depending on the design of states' alternate assessment systems. For example, if students are able to be assessed on different content standards and be evaluated against different performance standards, it makes sense that there is flexibility in the reporting systems.

9. Flexibility in how scores are handled for school accountability[6]
    - General assessments: With general assessments, to the extent that participation rules are rigorously enforced, there is (or at least should be) little variation in how scores are used in school accountability calculations.
    - Alternate assessments for students with significant cognitive disabilities: In terms of NCLB accountability and as long as less than one percent of the total student body is tested on the alternate assessment, there is probably little variability in how scores are handled for school accountability. However, if more than one percent of the students participate in the alternate assessment, scores will be handled differently depending on whether the particular score is considered above or below the one percent proficiency cap. Scores considered above the one percent cap cannot be counted as proficient no matter how the student actually scored.

Table 1 (below) summarizes the amount of flexibility intentionally designed and allowed in assessments for students taking alternate assessments, special education students taking general assessments, and general education students. (This is based on judgment and not on any systematic survey or coding.)

It is clear that alternate assessments based on alternate achievement standards contain considerably less standardization than general assessments with or without standard accommodations. It is important to recognize that just has the population of students participating in alternate assessments is quite diverse, the types of approaches used to assess these students is also quite varied. Some alternate assessments can be highly structured, similar in appearance to general education assessments, while other alternate assessment programs tend to resemble the flexibility often seen with high quality classroom assessments. Much of this flexibility is intentional, especially regarding the *targets* of assessment—items 1, 3, and 7 in Table 1. These targets are usually fixed and in common for all students in other assessments. The *means and methods* of assessing—items 4, 5, 6—can also vary by design in alternate assessments to a higher degree than in other assessments. Some of this flexibility, while intended, may exceed what was expected, especially in terms of test administration.

[5] We know this to be true for all of the states for which Measured Progress is the contractor (Sue Bechard, personal communication, 2/21/06).
[6] Our focus in this paper is on school accountability. We do not know of any cases where alternate assessments are used for holding individual students accountable and we argue that no alternate assessment of alternate achievement standards should be used in situations where students could suffer consequences as a result of their performance.

	General	General with standard accommodations	Alternate on AAS
1. Flexibility in the curricular goals among students at a point in time and over time (e.g., grade-level curriculum)	Low	Low	High (individual)
2. Flexibility in the instruction (learning experiences)	Moderate	Moderate-High	High
3. Flexibility in the content standards chosen to be assessed for specific students (e.g., the standards used to develop grade-level assessment)	Low	Low – moderate	Low-high
4. Flexibility in the methods/items used to assess	Low	Low – moderate	Low-high
5. Flexibility in how the tests are administered including administration conditions	Low	Low - moderate	Moderate-high
6. Flexibility in the scoring	Low	Low	Low-high
7. Flexibility in the performance standards (evaluative criteria)	Low	Low – moderate	Low-moderate
8. Flexibility in interpretation and reporting	Low – moderate	Low – high	Moderate – high
9. Flexibility in how handled for school accountability	Low	Low - moderate	Low

Table 1. Amount of Flexibility by Design

For accountability purposes, alternate assessments on alternate achievement standards currently are handled very similarly to assessment information for other students in school accountability, e.g., proficiency scores are treated as the same. However, for student accountability, alternate assessment data are often treated as non-comparable with data from other assessments. The increased flexibility of alternate assessments, while desirable from an instructional viewpoint for students with significant cognitive disabilities may be unacceptable in some testing and accountability situations. In the next section, we provide some suggestions for ways to deal with this flexibility in order to document the technical quality of alternate assessment systems.

Dealing with Flexibility in Assessments

As described above, flexibility of many types is inevitable to some degree when constructing an assessment system; some flexibility is often desired. The challenge to having flexibility inherent in an assessment system is tied directly to the intended uses of these assessment scores. Flexibility is an important characteristic for instructionally-based assessment, but creates challenges when we try to use scores from non-standardized assessments for large-scale accountability systems. Since the validity of score inferences is intricately linked to the uses of such scores, comparability of scores is a crucial part of the validity criterion for school accountability purposes. In other words, assessment scores from any two students should be able to be fairly compared unless there are different rules for how the scores are used in the accountability system. It is not that we are focused on comparing one student's performance to another's; rather the comparability criterion is an essential component of student accountability systems. Further, comparability of interpretations is important when aggregating student scores to the school, district, or state level so that we can be confident that we are not combining apples and oranges. The challenge then, is to figure out how the flexibility in alternate assessments can be handled so that valid comparisons and aggregations can be made.

There is a long tradition in large-scale assessment on how to handle flexibility in means and methods that may be extended to the particular challenges posed by alternate assessments, such as strict adherence to specific training protocols. However, less is known about how to deal with different curricular and assessment targets, in large part because so much effort has gone into standardizing (reducing the flexibility) of targets. In order to evaluate the technical quality of alternate assessments, researchers must first figure out how to characterize and handle both the intended and unintended flexibility. We discuss three general approaches for trying to ensure comparability among non-standardized assessment results even when the assessments are based on different curricular and assessment targets.

How can we determine how related one assessment is to another, in terms of content? Three ways to support interpretation of assessment results as comparable are outlined below and then explained in more detail following the brief introduction.

1. Establish construct comparability based on similar content – for example, one assessment item taps the same construct as another assessment item. This may be based on a content and/or cognitive analysis.
2. Establish comparability based on similar or compensatory functionality – distributional requirements often specify profiles of performance will be treated as comparable; total scores based on a compensatory system do similarly.
3. Establish comparability based on judgments of relatedness or comparability – disciplined judgments may be made to compare almost anything in terms of specified criteria (e.g., is this bottle as good a holder of liquid as this glass is?). Decision-support tools and a common universe of discourse undergird such judgments.

Some proponents of alternate/authentic assessment have embraced **construct comparability**. A typical belief for this position is that a student has the same cognitive knowledge and skills, and could demonstrate them if only the proper assessment could be devised. Technology is often used as an example supporting this viewpoint. A related viewpoint is that the same construct largely can be demonstrated, but with some different conditions. Those conditions are typically thought of as accommodations/modifications, but could encompass all the sources of flexibility described previously. A research task would be to array this range of performances, with the specified conditions, and then decide what is "comparable" in terms of content and skills. Essentially, this would require a conceptual map of the domain (including a developmental sequence or matrix), and a way to assess the alignment of the assessment task/response to the domain.

Functional comparability requires a judgment of things that are somewhat related. The judgment may be based on a conceptual analysis and/or on empirical analyses. A portfolio assessment system and certain types of observation checklists often include, at least implicitly, assumptions about functional comparability. Alternate assessments for students with significant cognitive disabilities rarely share this type of functional comparability with general assessments, because the two populations typically do not share overlapping tasks or instructional experiences. However, comparability of two students' alternate assessment scores in many systems is dependent, in part, on functional judgments. The state, school, and/or teacher have determined that students' different portfolio entries, for example, are comparable for school accountability purposes. Of course, states using assessment approaches where all students attempt the same tasks are not necessarily concerned about functional comparability. However, in many of these situations, the "same tasks" are altered by the amount of assistance provided by the test administrator and so in these cases assessment professionals must be concerned about construct similarity[7]. It is clear that we need to develop a better understanding regarding the interaction among the task, degree of scaffolding, and the target construct.

Judgment-based decisions about comparability enable us to make comparisons and judgments, and to interpret things that are less similar than the two previous categories. Alternate assessment achievement levels (performance standards) are typically in this vein—we decide as a policy (not because of content or functional similarity) that Proficient on the alternate assessment shall be treated as equivalent as Proficient on the regular assessment for certain purposes (e.g., school accountability).

In the complicated world of state assessment and accountability policy, the actual comparability decisions are often a combination of these three approaches. While this might be desirable, we should be very clear when we are operating in one approach versus another. Undesirable flexibility, on the other hand, should be identified and minimized by design and vigilant control over implementation processes. This will require policy decisions about what is "undesirable," of course.

Both the construct and functional comparability approaches are attempting to deal with differences in curricular and/or assessment targets. Two specific processes—content alignment and cognitive analysis—provide a means for helping to ensure the validity of either the construct or functional comparisons.

Content Alignment Process – One way would be to locate the two assessments on a content map. If we had a map that laid out all the content standards in a developmental sequence along one axis, and all the variations of simple to more complex (or whatever dimension we thought captured the important flexibility) along another dimension(s), then theoretically we could locate any two assessments on this map and show the relationship between them. The discussion in the field about alignment methodologies for alternate assessment on alternate achievement standards points to the need for this type of matrix (NHEAI/NAAC Expert Panel, 2006). Describing the relationship of the content for an alternate assessment in relation to the map would then allow one to make a more value-based judgment about whether the relationship between the alternate and general assessment scores or between two alternate assessment scores were close enough to constitute either "on-grade performance but with simplified content and skills" or equivalent in some important ways. Massachusetts has tried to do this in a very basic way by scoring alternate assessments on "closeness to grade level performance" (Wiener, 2006). Other commonly used alignment procedures (e.g., Webb, WestEd) provide similar ways of examining the relationship between the content standards for the alternate assessments and those guiding the general assessment.

---

[7] One could argue that certain assessments appearing to administer the "same" items to all students are actually allowing students to answer "different" items because of scaffolded assistance provided by the test administrator.

Cognitive Analysis – Cognitive models that specify performance differences in terms of mental models and developmental acquisition or elaboration of those models are an alternative to a content alignment process, although both seek to specify a universe of knowledge and skills. Developmental psychologists often assume that the mental models develop somewhat independent of curriculum, or that the stages/models are invariant or stable (e.g., Piagetian or neo-Piagetian models). These invariant characteristics go beyond surface variations to deep explanatory similarities and differences. On the other hand, cognitive and sociocultural researchers theorize that people develop models as a result of their learning (and social) interactions and experiences (e.g., curriculum and instruction). Recent discussions about cognitive models of students with significant cognitive disabilities touched on the need for this type of cognitive model specification so that valid inferences can be drawn from assessment results (NHEAI/NAAC Expert Panel, 2006). As a fundamental aspect of the assessment triangle (NRC, 2001), the specification of cognitive models will allow for more sound judgmental comparisons to made among various assessment scores.

In addition to the two approaches discussed above, there are also adaptations of traditional psychometric methods that can be used to help create some degree of comparability. These also rely, at least in part, on one or more of the general comparative methods discussed previously.

"Scaling and Equating" – The purpose of scaling and equating is to ensure that similar inferences can be derived from similar test scores both within and across years. Almost all alternate assessments for students with significant cognitive disabilities are on different scales than the state's general education assessment, so the challenge is to ensure that valid comparisons can be made for multiple alternate assessment scores within the same state. Depending on the nature of the specific alternate assessment, some fairly traditional scaling methods may be used to ensure comparability of inferences for students with similar scores. But for other systems, more qualitative or judgmental solutions to scaling will have to be employed to help create comparable inferences, i.e., the scores are functionally equivalent.

Alternate assessments pose particular challenges for year-to-year "equating," the family of techniques used to ensure comparability of inferences across years. Most general education equating designs rely on having a portion of the full set of test items administered to both groups of students (across years). Some alternate assessment designs eliminate the need for equating by administering exactly the same test across years. Assuming that proper testing practices are followed (e.g., no inappropriate teaching to the test), scores across years in these situations are already "equated." Other testing designs allow teachers or others to create unique tasks for students tested each year and to administer unique sets of tasks to individual students. In these cases traditional equating designs will not be useful. Rather, judgmental methods—that is judgments against specified standard similar to comparing writing samples using a common rubric—will have to be employed to satisfy the need for year-to-year or student-to-student comparability.

"Standard setting" – The purpose of standard setting is to determine how various performances are valued. Typical achievement standards provide a way to say that similar scores (assuming they do not cross cutscore boundaries) are valued equally, even if those similar scores were based on different skill profiles. Most states establish a single set of alternate achievement standards, but there is no prohibition to establish multiple alternate achievement standards if the state determined that it was important to value different performances (in some absolute sense) equally or at least functionally similar. Most standard setting methods rely on examining patterns of item responses or student score

profiles from large numbers of students completing the same assessment. Alternate assessments challenge these common methods because the cutscores are generally established based on relatively few students, therefore more judgmental approaches than might be the case with general education assessment will have to be employed to set alternate achievement standards. These judgmental methods should incorporate notions of functional equivalence to clarify the values associated with different performances.

Implications for evaluating technical quality

Evaluating the technical quality of alternate assessment systems requires drawing on existing psychometric and evaluation techniques as well as modifying existing approaches or inventing new ones. When asked to document the technical quality of alternate assessment systems, the first reaction of measurement professionals is that the small numbers of students and relatively variable nature of the assessment system makes this an almost impossible task. The apparent non-standardization characteristic of many alternate assessments appears overwhelming when thinking about applying traditional psychometric techniques. The purpose of this paper is to help sort out the variability or flexibility inherent in most alternate assessment systems.

Most of the increased levels of flexibility in alternate assessment system can be dealt with in evaluations of technical quality through a variety of judgmental methods. However, flexibility in learning goals and in selecting the content standards to be assessed poses more significant challenges for aggregating results and then evaluating the technical quality of the inferences from student and school scores.

We have tried to describe in this paper the various areas of flexibility and standardization in state alternate assessment programs. As documented in the paper, different assessment forms and alternate assessment programs have wide ranges of standardization, usually for educationally defensible reasons. By discussing the degree of standardization/flexibility for the component parts of the assessment system, we have tried to provide a framework to help state policy makers consider where they might like to create more standardization and for what aspects of the system they would like to keep flexible.

The one area that is perhaps the most challenge, in terms of technical quality evaluations, involves the variability of content assessed for students at a given grade level. We are not recommending fixing the content domains assessed for all students, although some programs do this successfully, but we are recommending that states consider methods for addressing the apparent lack of comparability among assessment domains for individual students within a given state alternate assessment program. Creating more comparable score inferences is not something that can be accomplished quantitatively, but will require the use of systematic judgmental methods. The work in content alignment can provide some useful tools and procedures to help with this issue. We also thinking getting better at cognitively mapping the tasks and performances—along the multiple dimensions of performance—can help provide a systematic method for comparing tasks and student performance.

Many people are justifiably concerned that the flexibility associated with alternate assessments is a barrier to fairly evaluating the technical quality of these assessments. The current work of the New Hampshire Enhanced Assessment Initiative is leading to some techniques and procedures to help address some of the traditional areas of technical evaluations such as reliability, scaling and equating, and standard setting. In the language of the assessment triangle (NRC, 2001), these techniques are focused on the interpretation vertex of the triangle. This paper has tried to contribute to that work by beginning the discussion about the sources flexibility and expectations for standardization associated with the other two vertices—cognition and observation.

<u>What should states do now?</u>

State assessment leaders and other key policy makers should first be very clear about the purposes and uses of the assessment scores. Depending on the highest priority purposes, flexibility for many aspects of the assessment system may be vital for fulfilling these purposes, such as improved classroom instruction. It is this clarity of purpose that will allow state leaders to properly evaluate the flexibility in their system.

State alternate assessment leaders should be clear and explicit regarding where flexibility is intended compared with where unintended flexibility may become part of the system. The worksheets (developed by Marion, Quenemoen, & Kearns, 2006) found in Appendix B are designed to assist states with identifying the intended and unintended sources of flexibility in the assessment system. Once these worksheets are completed, state leaders and test designers should clarify what types of judgmental methods can and will be used to try to facilitate comparability, if desired, among student scores.

Over the next months and years, the work of projects like the New Hampshire Enhanced Assessment Initiative (NHEAI) and the National Alternate Assessment Center (NAAC) will continue to expand our understanding of practical and educationally sound solutions to the challenges of fully inclusive assessment systems. Lessons learned from these and other efforts can help define areas for improvement of entire assessment systems.

References

Marion, S. F. (2004). *Psychometric Concerns When Measuring Advanced Knowledge*. Unpublished doctoral dissertation. Boulder, CO: University of Colorado.

Marion, S., Quenemoen, R., & Kearns, J. (2006). *Inclusive Assessment System Options: Degree of Standardization and Flexibility Worksheets*. Working papers from NHEAI/NAAC Collaborative Projects.

NHEAI/NAAC Expert Panel. (2006). Expert Panel Meeting Notes, Alexandria, VA, February 2-3, 2006.

Kleinert, H., Browder, D., & Towles-Reeves, E. (2005). The assessment triangle and students with significant cognitive disabilities: Models of student cognition. National Alternate Assessment Center, Human Development Institute, University of Kentucky, Lexington. (PDF File)

National Research Council, Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Quenemoen, R., Thompson, S. & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring criteria*(Synthesis Report 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/nceo/OnlinePubs/Synthesis50.html

Wiener, D. (2006). *Alternate assessments measured against grade-level achievement standards: The Massachusetts "Competency Portfolio"* (Synthesis Report 59). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis59.html

Appendix A

The following information is excerpted from Quenemoen, Thompson, & Thurlow (2003)[8] to help provide additional background material regarding the various types of alternate assessment. The important message from this report is that these categories of assessment forms are not mutually exclusive.

Alternate Assessment Approaches: Not Mutually Exclusive Categories

In general, the alternate assessment approaches defined in Table 1 go from a basic methodology of student-by-student individually structured tasks (portfolio assessment) to highly structured common items or tasks completed by every student (traditional tests) as you read down the table. These approaches are not mutually exclusive categories, and as state practices are examined, it is clear that a great deal of overlap in methods occurs.

Portfolio Overlap with IEP Linked Body of Evidence

The "portfolio" approach typically requires the gathering of extensive samples of actual student work or other documentation of student achievement. Very often, this work is in response to teacher-developed tasks with teacher-defined linkage to content standards, thus the evidence varies dramatically from one student to the next. It is the standardized application of scoring criteria to the varied evidence that results in comparability across students. "IEP linked body of evidence" approaches as defined here also may require extensive sampling of work, have similar scoring criteria, and apply them in similar ways to the portfolio approach. However, in this report, the states using a portfolio approach require extensive student products; the state that uses an IEP linked body of evidence has more focused evidence requirements, related specifically to the skills and knowledge defined in the student's IEP, and the documentation of progress in the IEP process. In general, the distinguishing characteristics between "portfolio" approaches versus "body of evidence" approaches tend to be, for the purpose of this report:

1. the amount of evidence required is more for portfolio, less for body of evidence;
2. the degree of state provided definition of what specific content is measured is less with portfolios, and there is more state provided definition of specific content for a body of evidence; and
3. the degree of IEP linkage is less for portfolio and more for a body of evidence.

(A complicating variable is how advanced a state is in implementing standards-based IEP planning, thus the IEP linkage to alternate assessments may be "pushing the envelope" of standards-based reform for students with disabilities. That discussion is beyond the purpose of this report, but will be increasingly important as alternate assessment evolves.)

[8] Quenemoen, R., Thompson, S. & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring criteria*(Synthesis Report 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/nceo/OnlinePubs/Synthesis50.html

Table 1. Definitions of Alternate Assessment Approaches Discussed in this Paper

Portfolio: A collection of student work gathered to demonstrate student performance on specific skills and knowledge, generally linked to state content standards. Portfolio contents are individualized, and may include wide ranging samples of student learning, including but not limited to actual student work, observations recorded by multiple persons on multiple occasions, test results, record reviews, or even video or audio records of student performance. The portfolio contents are scored according to predefined scoring criteria, usually through application of a scoring rubric to the varying samples of student work.
IEP Linked Body of Evidence: Similar to a portfolio approach, this is a collection of student work demonstrating student achievement on standards-based IEP goals and objectives measured against predetermined scoring criteria. This approach is similar to portfolio assessment, but may contain more focused or fewer pieces of evidence given there is generally additional IEP documentation to support scoring processes. This evidence may meet dual purposes of documentation of IEP progress and the purpose of assessment.
Performance Assessment: Direct measures of student skills or knowledge, usually in a one-on-one assessment. These can be highly structured, requiring a teacher or test administrator to give students specific items or tasks similar to pencil/paper traditional tests, or it can be a more flexible item or task that can be adjusted based on student needs. For example, the teacher and the student may work through an assessment that uses manipulatives, and the teacher observes whether the student is able to perform the assigned tasks. Generally the performance assessments used with students with significant cognitive disabilities are scored on the level of independence the student requires to respond and on the student's ability to generalize the skills, and not simply on accuracy of response. Thus, a scoring rubric is generally used to score responses similar to portfolio or body of evidence scoring.
Checklist: Lists of skills, reviewed by persons familiar with a student who observe or recall whether students are able to perform the skills and to what level. Scores reported are usually the number of skills that the student is able to successfully perform, and settings and purposes where the skill was observed.
Traditional (pencil/paper or computer) test: Traditionally constructed items requiring student responses, typically with a correct and incorrect forced-choice answer format. These can be completed independently by groups of students with teacher supervision, or they can be administered in one-on-one assessments with teacher recording of answers.

Adapted from Roeber, 2002.

Body of Evidence Overlap with Performance Assessment

The "body of evidence" tendency toward more focused and specific evidence in turn reflects a similarity with the least structured specific "performance assessment" approaches in other states. That is, some performance assessment approaches define the skills and knowledge that must be assessed for each student, but they still allow the test administrator to structure a task that the student will perform to demonstrate the skills and knowledge. The most structured body of evidence approaches tend to be very similar to the least structured performance assessments. In other words, a state may require in a performance assessment OR a body of evidence that a student demonstrate his or her reading process skills by identifying facts, events, or people involved in a story. How the student demonstrates those skills will vary, and the task could involve, for example:

- ✓ requiring that a student use a switch to provide different sound effects corresponding to characters in a story, whether read by the student or teacher;
- ✓ having a student look at pictures to identify favorite and least favorite parts of a story that was read aloud; or
- ✓ a student reading a simple story and then making predictions of what will happen next using clues identified in the text.

As a further source of individualized tailoring in either a highly structured body of evidence or a loosely structured performance assessment, each of these tasks could allow for varying levels of teacher prompting, and thus scoring criteria could include the criterion of the level of prompting/degree of independence. Where the approaches differ is that a body of evidence approach generally requires submission of the student evidence for scoring; a performance assessment approach typically involves the test administrator or teacher scoring student work as it occurs.

Other states that define their approach as a performance assessment provide a high degree of structure and specifically define the task (e.g., having a student look at pictures to identify favorite and least favorite parts of a story that was read aloud, with provided story cards and materials). Yet they typically allow variation in the degree of prompting (ranging from physical prompts to fully independent responses), or in the methods of student responses (from use of picture cards vs. verbal response for example). Even states that use common performance assessment tasks for their required alternate assessment for students with significant disabilities tend to use multiple scoring criteria more similar to portfolio or body of evidence approaches, as compared to simple recording of correct or incorrect responses used in checklist or traditional test formats.

<u>Performance Assessments Overlap with Checklist and with Traditional Test Formats</u>

Most "checklist" approaches ask the reviewer to record whether a student has demonstrated the skill or knowledge. These may include a judgment on degree of independence or generalization as well as accuracy of skill performance, but the judgments may simply reflect accuracy. The difference between checklists and performance assessment approaches where the test administrator scores the performance is that the checklist approach relies on recall and not on actual on-demand student performance. By contrast, "traditional test" formats require the on-demand performance of skills and knowledge, on a specified item, with built in connection to content standards and with accuracy (or "right/wrong") as the primary criterion. The test administrator records student performance as right or wrong, and no further scoring is necessary. This approach is the most similar to the testing approaches most adults have experienced described in the opening section of this report.

Appendix B:

Inclusive Assessment System Options by Degree of Standardization/Flexibility

Scott Marion., Center for Assessment

Rachel Quenemoen, National Center for Educational Outcomes

Jacqui Kearns, NAAC, University of Kentucky

February 2, 2006

Directions: For each of the options in your state assessment system, please indicate the degree of standardization/flexibility for each component of the assessment/accountability system listed in 1-11 below. Delete any options that do not apply to your state system; add options that you offer that are not described here.

Please use 1-5 scale to indicate your ratings while keeping the following anchor points in mind:

5: highly flexible (along the lines of most diagnostic assessments—focus is on maximizing information about individual students' knowledge and skills)

4: quite flexible (an interest in comparability, but more of a focus maximizing having students demonstrate knowledge & understanding)

3: moderately flexible (a balance between comparability and individualization)

2: quite standardized (along the lines of most state general education statewide assessments—focus is certainly on comparability and aggregation, but allows some flexibility in administration conditions and presentations)

1: highly standardized (along the lines of NAEP, SAT, etc.—focus is on the comparability and aggregation of assessment results)

Briefly summarize who is eligible to participate in each option at the top, as well as the understanding of how these students demonstrate their learning as it relates to the design of each assessment option.

Assessment/Accountability System Components (See Gong and Marion typology paper for examples of each)	**State Inclusive Assessment System Options**					
	General Assessment	**Gen Assmt w/ standard accommodations**[9]	**Gen Assmt with modified administration**[10]	**Alternate assmt format on modified ach. standards**[11]	**Alternate Assmt on grade level achievement standards**	**Alternate Assmt on alternate achievement standards**
1. WHO is assessed by this option (summary of participation guidelines for this option)						
2. WHY this group needs this option (how the group develops proficiency in the domain, how they evidence their learning, it as it relates to this option)						
3. Curricular goals to be assessed across population						
4. Test specifications/ choice of assessment format to use						
5. Item/task development						
6. Administration conditions – timing, presentation and response options						
7. Scoring process						
8. Performance standards (evaluative criteria)						
9. Reporting						
10. How scores are handled for student accountability						
11. How scores are handled for school accountability						

[9] Although the scores from an assessment taken with standard accommodations has, in most cases, been determined to mean the same as results from the assessment taken without accommodations, there are important measurement considerations that affect the degree of standardization. Thus, in this table, they are listed separately.
[10] This category is part of the proposed new 2% flexibility
[11] This category is also part of the proposed new 2% flexibility.

Alternate Assessment Design Elements by degree of standardization vs flexibility

Directions: For this specific profile, please indicate the degree of standardization/flexibility for each component of the assessment/accountability system. Please use 1-5 scale to indicate your ratings while keeping the following anchor points in mind:

1=highly flexible (this would be equivalent to most classroom instruction and assessment systems)
2=less flexible than 1
3=moderately standardized with some flexibility
4=less standardized than 5
5=highly standardized (along the lines of NAEP, SAT, etc.)

Assessment/Accountability System Components	**Degree of standardization/ flexibility (1-5 scale)**	***Evidence*** ***What protocols, training, audit procedures, or other data are sources of evidence regarding the stated degree of standardization and/or flexibility?***
1. Curricular goals to be assessed across population		
2. Test specifications/choice of assessment format to use		
3. Item/task development		
4. Administration conditions – timing, presentation and response options		
5. Scoring process		
6. Performance standards (evaluative criteria)		
7. Reporting (add examples)		
8. How scores are handled for student accountability		
9. How scores are handled for school accountability		