Utah’s Alternate Assessment: Evidence Regarding Six Aspects of Validity

Karen D. Hager
University of Kentucky

Timothy A. Slocum
Utah State University

Abstract: Alternate assessments are the means through which students with significant cognitive disabilities participate in accountability testing, thus measurement validity of alternate assessments is a critical aspect of state educational accountability systems. When evaluating the validity of assessment systems, it is important to take a broad view of validity that acknowledges its many facets. Evidence regarding six aspects of validity of Utah’s Alternate Assessment (UAA) is presented in this paper. These six aspects are: (a) test content, (b) performance standards, (c) administration of the test, (d) reliability of scoring, (e) stability of scores, and (f) impact of use of UAA. Results of the validity study conducted on UAA are presented along with recommendations for improving the system based on these results.

Recent federal legislation (Individuals with Disabilities Education Act [IDEA 97 and IDEA 2004]; No Child Left Behind Act of 2001 [NCLB]) has dramatically increased the importance of wide-scale assessment results, while also extending systematic accountability assessment to all students. Thus, even students with the most significant cognitive disabilities are now included in district- and state-wide testing. Prior to these legislative mandates, many students with disabilities had not been included in accountability testing (Ysseldyke & Olsen, 1999).

While it is clear that the accountability assessments designed for general education students are not meaningful for students with significant cognitive disabilities, it is not as clear what kind of testing is meaningful. States have adopted several different approaches to developing and implementing alternate assessment systems. Thompson, Thurlow, Johnstone, and Altman (2005) conducted a survey of state directors of special education and reported that 25 states use a portfolio/body of evidence system. Other approaches identified by states include completion of a rating scale/checklist (seven states), analysis of individualized education program (IEP) goals (two states), and other, defined in this survey as allowing local selection of an approach or requiring a performance assessment (seven states). Eight states reported that their system was under development/revision, and one state has not developed a statewide approach (Thompson, et al.). There is, however, considerable overlap among these categories. For example, states that identify their approach as performance assessment are categorized as other, yet 13 states that identified their approach as portfolio/body of evidence use a standardized set of performance events, tasks, or skills.

Regardless of the approach taken, states are responsible for ensuring that the assessment system they use is technically adequate. Alternate assessments must meet the same standards for high technical quality as the general assessments (Title I - Improving the Academic Achievement of the Disadvantaged, 2003), and must be consistent with nationally recognized professional and technical standards, e.g., Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association,
and National Council on Measurement in Education, 1999 [Standards]).

Only a small number of validity studies of alternate assessments have been published, and nearly all have been conducted on portfolio-based systems. For example, several studies have been conducted on Kentucky’s Alternate Portfolio system, including expert validation of the learning outcomes and performance criteria (Kleinert & Kearns, 1999), investigation of teachers’ perceptions of the benefits of including students with significant cognitive disabilities in the accountability assessments (Kleinert, Kennedy, & Kearns, 1999), and studies examining the external aspect of validity (Kleinert et al., 2000). Kleinert and Kearns asked respondents to rate the importance of learning outcomes and performance criteria for individuals with moderate and severe disabilities. On a scale of one (low importance) to five (high importance), nearly 70% of the performance criteria obtained a mean rating of 4.5 or higher, but only 40% of the academic expectations obtained a mean rating of 4.5 or higher. Kleinert, Kennedy, and Kearns reported that 55% of the teachers who responded to their survey strongly agreed or agreed that it is important to include students with significant cognitive disabilities in the accountability system. Kleinert et al. investigated whether alternate assessment scores in the last year of school were predictive of post-school outcomes. They did not find a relationship between alternate assessment scores and the measure they developed to measure post-school outcomes (r = .09). Turner et al. correlated measures of IEP quality and overall program quality with alternate assessment scores. They found a correlation of .45 (p < .001) between the measure of overall program quality and alternate assessment scores, but did not find statistically significant correlations (r = .09) between the measure of IEP quality and alternate assessment scores.

Johnson and Arnold (2004) also investigated a portfolio system, and reported evidence related to the validity of the Washington Alternate Assessment System (WAAS). They examined three aspects of the WAAS: content, teacher response process, and internal structure. Johnson and Arnold concluded that scoring procedures should be revisited because (a) teacher ability to compile a complete and accurate portfolio weighed too heavily, and (b) the generalization score received an inordinate amount of weight in comparison with the content score.

One validity study (Tindal et al., 2003) has been conducted on a performance-based system. This study was designed to assess scoring reliability and the ability of the assessment to differentiate between students with different disabilities (e.g., mental retardation and learning disabilities). Tindal and colleagues reported that results on the assessment did distinguish students with different disabilities and that pairs of scorers who provided qualitative ratings on a scale of zero to six were often within one point or less of each other (math = 89%; reading = 88%).

Given the mandate that alternate assessments must be of high technical quality, it is important to determine what evidence would support such a judgment. A brief review of the development of current validity theory and validity models, which served as the framework for the model used to evaluate Utah’s Alternate Assessment (UAA), is presented next.

Validity Theory

Validity standards were first codified in 1954 by the American Psychological Association (APA) in Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA, 1954). This document and subsequent editions of the Standards reflect the evolution of validity theory and professional consensus on the meaning of measurement validity. The 1954 Recommendations identified four types of validity (content, predictive, concurrent, and construct), each with a specific use. Validity was first identified as a unified concept in the 1985 Standards, which referred to evidence in different categories (rather than different types of validity), and clarified that a meaningful validation of a test is the result of analyzing evidence across these categories (Messick, 1993).

Messick (1993) identified two major threats to validity: construct underrepresentation and construct irrelevant variance. Construct underrepresentation occurs when the assessment is too narrowly focused and fails to in-
clude important dimensions of the construct. Construct irrelevant variance occurs when variables that are not part of the intended construct are sources of reliable variance in assessment results. In Messick’s view, various forms of evidence contribute to understanding construct underrepresentation and construct irrelevant variance, and thereby to understanding the validity of test interpretation and use. Messick stated that construct validity, “Is based on an integration of any evidence that bears on the interpretation or meaning of the test scores” (Messick, 1995b, p. 742). Messick (1993) argued that social values and consequences of test use are also important aspects of validity that must be investigated when judging the validity of a test interpretation or use. While all aspects of validity must be considered, depending on the specific interpretation and use intended, evidence regarding some aspects will take priority and warrant more emphasis than others.

Concerned with the difficulty of applying Messick’s validity theory, Shepard (1993) called for simpler models of prioritizing validity questions, and suggested that gathering of evidence be organized around what the testing practice claims to do. Kane (1992) suggested an argument-based approach to evaluating validity. Kane described this approach as one in which the researcher:

(a) decides on the statements and decisions to be based on the test scores, (b) specifies the inferences and assumptions leading from the test scores to these statements and decisions, (c) identifies potential competing interpretations, and (d) seeks evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counterarguments. (p. 527)

Crooks, Kane, and Cohen (1996) combined the argument-based approach with specific validation criteria, resulting in a validation model that focuses on eight steps that link the actual testing behavior to interpretations, decisions, and impacts of testing. The eight steps of this model are (a) administration, (b) scoring, (c) aggregation, (d) generalization, (e) extrapolation, (f) evaluation, (g) decisions, and (h) impact. Crooks et al. argued that investigators should identify plausible threats to validity at each step and accumulate evidence regarding the threats in a validation process.

We used the work of Messick (1993), Shepard (1993), and Crooks et al. (1996) as a basis for developing a validity model for UAA. Acknowledging that the process of evaluating the validity of such an assessment system should be continuous and perhaps never be considered complete, the validity model used to study UAA was designed to gather as much information as possible from which to draw initial conclusions regarding the use of this assessment as part of the state educational accountability system. We considered the validity of six aspects of UAA: (a) test content, (b) performance standards, (c) administration of the test, (d) reliability of scoring, (e) stability of scores, and (f) impact of use of UAA.

Utah’s Alternate Assessment

UAA is a performance-based assessment system for students with significant cognitive disabilities. It consists of 115 tasks designed to measure independent performance of language arts and math skills that are embedded in naturally occurring routines in natural settings. The tasks are arranged in each domain in a (somewhat) hierarchical manner relative to the difficulty of the skill assessed in each task. These assessment tasks are a menu from which IEP teams select assessment targets that are relevant for individual students. The team annually determines which targets are challenging, yet attainable for an individual student. Once a specific task has been selected, however, the team must adhere to the standardized procedures. These standardized components include administration procedures, performance criteria, and scoring criteria. A brief description of specific components of UAA is provided next.

Specified Assessment Tasks

UAA consists of 115 specified assessment tasks in language arts and math. This system prescribes actual tasks, rather than identifying academic standards that must be assessed and then charging the IEP team with developing
the tasks to do so. This allows anyone familiar with UAA to understand the assessment results. The task and administration procedures are clearly articulated. This approach also allows the tasks to undergo a more rigorous review process than would be feasible if individual IEP teams were creating measurement tasks.

**Task Selection by the IEP Team**

IEP teams select target tasks for each student from the sets of language arts and math tasks. The decision to allow IEP teams to select target tasks and adapt some aspects of these tasks to individual circumstances is based on the premise that for a student with significant cognitive disabilities, the IEP team members are the most qualified people to determine challenging, yet attainable target tasks.

**Direct Observation of Student Performance**

The teacher observes the student performing the assessment tasks and scores them as they are performed. An important basis for this decision is that direct observation requires less inference than other types of scoring (e.g., rating scales or checklists that rely on interpreting performances that occurred at some time in the past).

**Established Performance Criteria**

All students are scored against the same performance criteria for a given task. The minimal level of performance necessary to make the skill useful does not vary across students; thus, UAA is designed to measure attainment of a functional level of performance on the target skills. We define a *functional level of performance* as the level at which the skill is meaningful or useful.

A four-point scale is used to determine proficiency levels on all state accountability assessments in Utah, including UAA. Levels four (*Substantial*) and three (*Sufficient*) demonstrate adequate yearly progress, and levels two (*Partial*) and one (*Minimal*) do not demonstrate adequate yearly progress. Most UAA tasks include three trials. To attain *Substantial* on UAA tasks, the student must meet all criteria on all three trials; a proficiency level of *Sufficient* requires that the student meet all criteria on two out of the three trials.

**Independent Performance**

Students must perform the tasks independently (i.e., no physical, verbal, or gesture assistance). One of the goals of this system is to measure mastery of skills that promote independence and access to more environments. If the student is only able to perform the skill with support, this does not promote independence.

**Generalized Performance**

Each task specifies how generalization should be measured. Generalization criteria are included because a skill is truly useful for a student only if it is performed whenever and wherever necessary (i.e., in natural routines). It is risky to assume that because a student performs a skill under one set of conditions, such as in the classroom with the teacher, he/she will perform that skill under other conditions (Brown, Nietupski, & Hamre-Nietupski, 1976; Stokes & Baer, 1977; Wehman, 1997). To ensure that the student is able to perform the skill in different settings, with different people, and with different materials, it is necessary to directly observe the student perform the skills under these different conditions.

**Assessment in Naturally Occurring Routines**

As noted above, in order for a skill to be useful it must be performed in the student’s natural routines, thus UAA tasks are designed to be administered in natural environments during typical routines. For example, task 138 (Money – uses “dollar more” strategy), should be assessed when the student is buying something (e.g., in a convenience store, in the school cafeteria). Again, similar to generalized performance, the only way to be sure the student can perform the skill in natural routines is to assess in these contexts. Therefore, this system is designed to embed the assessment tasks in students’ naturally occurring routines. This allows for greater confidence that test performance reflects actual skill use.
Acceptability of Any Response Mode

Any communication system or assistive technology the student typically uses is acceptable during the assessment. The focus of assessment is that the student makes an independent response; how the student makes the response (e.g., verbally, with a communication device) is not as important. This is another case in which the members of the IEP team are the most qualified individuals to determine the most appropriate response mode for an individual student. Clearly, the most universally understood response is the most desirable, but again, the IEP team members are in the best position to make this decision. Figure 1 shows an example UAA task.

Figure 1. Example UAA task.

The broad purpose of this study was to investigate the degree to which UAA provides valid measurement of adequate yearly
progress for students with significant cognitive disabilities. We posed one or more specific research questions for each of six aspects of UAA: (a) test content, (b) performance standards, (c) administration of the test, (d) reliability of scoring, (e) stability of scores, and (f) impact of use of UAA. The complete set of research questions is presented in Table 1.

**Method**

Two data collection strategies were used to gather evidence on the validity of UAA: (a) surveying relevant experts, and (b) videotaping and systematically observing administrations of UAA tasks to students participating in the alternate assessment.

**Survey-based Methods**

Content relevance and representativeness are traditionally assessed by expert judgment (Messick, 1995a), thus the research questions on test content and appropriateness of performance standards (research questions one through seven) were addressed by asking individuals with relevant expertise to evaluate these aspects of UAA. The 2003-2004 school year was the first year of implementation for this version of UAA and direct evidence of consequences resulting from use of this test was not yet available. Thus, evaluation of the impact of use of UAA was necessarily speculative. Therefore, research questions related to impact of test use (questions 12 – 14) were addressed by asking individuals with relevant expertise to respond to questions about potential effects of implementation of UAA.

**Survey participants.** Four different groups of experts participated in surveys: (a) special education faculty members, (b) special education teachers, (c) parents of students with significant cognitive disabilities, and (d) special education administrators. A total of 42 individuals completed surveys, with different expert groups responding to different questions. Table 2 lists each expert group, the qualifications for each group, the selection process, the research questions to which the group responded, and the total number of participants in each group.

**General procedures.** An introductory letter, written questionnaires, and self-addressed, stamped envelopes were provided to participants. The survey formats for the various research questions varied depending on the information required to answer the question. All surveys provided space for the participants to write comments. See Figure 2 for an example of the survey formats. Specific details of participants and copies of all surveys are available from the first author.

**Observation-based Methods**

Data for research questions addressing implementation fidelity, scoring reliability, and

---

**TABLE 1**

<table>
<thead>
<tr>
<th>Validity Aspects and Research Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Research Questions</strong></td>
</tr>
<tr>
<td><strong>Test Content</strong></td>
</tr>
<tr>
<td>1. Does each task fall within the domain of math or language arts?</td>
</tr>
<tr>
<td>2. How important is each task for students with significant cognitive disabilities?</td>
</tr>
<tr>
<td>3. How well does the set of tasks represent the construct of important language arts/math tasks for students with significant cognitive disabilities?</td>
</tr>
<tr>
<td>4. How well do the procedures for implementing tasks measure the named skills?</td>
</tr>
<tr>
<td><strong>Performance Standards</strong></td>
</tr>
<tr>
<td>5. How appropriate are the scoring criteria for individual trials?</td>
</tr>
<tr>
<td>6. How appropriate are the scoring criteria for determining a proficiency level based on a set of trials?</td>
</tr>
<tr>
<td>7. Is administering two tasks adequate to determine adequate yearly progress for students with significant cognitive disabilities?</td>
</tr>
<tr>
<td><strong>Administration</strong></td>
</tr>
<tr>
<td>8. How reliably do testers follow the administration procedures of the tasks?</td>
</tr>
<tr>
<td><strong>Scoring</strong></td>
</tr>
<tr>
<td>9. How reliably do testers score the tasks?</td>
</tr>
<tr>
<td>10. How stable are scores across time?</td>
</tr>
<tr>
<td>11. How stable are scores across variation in testers, settings, materials?</td>
</tr>
<tr>
<td><strong>Decisions/Impact</strong></td>
</tr>
<tr>
<td>12. What are the likely effects of implementation of UAA on curriculum and instructional decision making and planning?</td>
</tr>
<tr>
<td>13. What are the likely effects of implementation of UAA on district and state level planning (e.g., staff development)?</td>
</tr>
</tbody>
</table>
score stability (research questions 8 - 11) were collected by videotaping administrations of UAA.

**Videotape participants.** Thirteen teachers participated in the videotaped portion of this study. The teacher participants came from three school districts: an urban district, a suburban district, and a rural district. The special education director in each district provided a list of all special education teachers whose caseloads included students with significant cognitive disabilities. The researcher contacted each teacher by phone, confirmed that the teacher was administering UAA, explained the study, and invited the teacher to participate. Participants included all teachers in the identified districts who had at least one student participating in UAA and who agreed to participate in this study. Seventeen teachers were invited to participate, and thirteen consented. Five teachers were in elementary schools, five in middle schools, and three in high schools.

Twenty-six students participated, one to three students from each participating teacher’s caseload. All students eligible to participate in UAA, for whom consent was obtained, and who attended school during the testing window were included.

**General procedures.** Administrations of UAA tasks were videotaped during the testing window from the middle of April through May, 2004. Eight special education undergraduate and graduate students were trained to videotape UAA task administrations. Training included an overview of UAA, use of the digital camcorders, taping protocol (e.g., labeling of video cassettes, filming all materials used in trials), and protocol for entering schools and classrooms. The videographers taped six to twelve trials for each student participant.
Tasks videotaped. For each participating student, the basic UAA administration of one math task and one language arts task was videotaped. UAA tasks typically require three trials for each task and all trials were videotaped. In addition, it was necessary to perform additional trials for the research questions on score stability (research questions 10 and 11). Therefore, in addition to videotaping each student participant’s initial performance of the two UAA tasks (one language arts and one math), test-retest and generalization trials were also arranged and videotaped.

Validity of Utah’s Alternate Assessment / 151
tion of generalization is only relevant once the student has demonstrated mastery (i.e., a proficiency level of *substantial*) on a UAA task. Without this demonstration, there would be no basis for predicting the student has the skill necessary to succeed in additional trials. Therefore, generalization was assessed only on tasks on which the student had attained a proficiency level of *substantial*.

The trials that were videotaped were determined according to the following procedures:

- Each student participant performed one language arts task (three trials) and one math task (three trials). These tasks constitute the basic UAA administration.
- If the student achieved a score of *substantial* on both tasks, one task was randomly assigned to test-retest assessment and the other to generalization assessment.
- If student achieved a score of *substantial* on only one task, that task was assigned to generalization assessment and the other to test-retest.
- If student did not achieve a score of *substantial* on either task, one task was randomly selected for test-retest.

Test-retest and generalization tasks were administered two days to four weeks after initial tasks. These tasks were conducted within a few days for 20 of the 26 participants. One district had completed all UAA administrations before permission was obtained from the district to participate in this study. Therefore, the initial tasks could not be videotaped for these students. However, test-retest and generalization tasks were videotaped. The time period between tests was one to four weeks for these six students. The tasks were videotaped in a range of locations including classrooms, lunchrooms, football fields, and convenience stores. Videographers followed the students to any location in which they performed UAA tasks.

Evaluating videotapes. The research question on reliability of test administration procedures (research question eight) was addressed by viewing each videotape and completing a researcher-developed checklist (see Figure 3). Five aspects of administration were assessed for each trial: (a) fulfilling the critical function of the task, (b) assessing within a natural routine, (c) presenting the prescribed directions to the student, (d) presenting the directions only one time, and (e) refraining from prompting the student during the assessed performance. To gather data on reliability of scoring (research question nine), the researcher viewed each task administration and completed the task’s assessment record without prior knowledge of the teacher’s scoring of the task.

To assess test-retest reliability (research question 10), the researcher viewed a second administration of a subset of the tasks viewed for scoring reliability. The assessment record for each task was completed and compared with the assessment record that had been completed for scoring reliability on the first administration. The same procedure was used to assess generalization (i.e., stability of scores across variation in testers, settings, and/or materials - research question 11). For this question, however, the second administration differed across one or more of these variables.

Interobserver Agreement

The researcher systematically observed videotaped administrations of math and language arts tasks administered to 26 students. Five aspects of implementation fidelity were scored for each task using a researcher-developed checklist. This checklist was also completed by a second observer on one set of videotaped trials for each of ten randomly selected students (38% of student participants). Half of these trials assessed a math task, and half assessed a language arts task. Interobserver reliability was calculated separately for each step identified on the implementation fidelity checklist, with the following results: fulfilled critical function (96%), assessed within typical routine (90%), acceptable directions (100%), directions presented only once (95%), and refrained from providing prompts (90%).

Results

Test Content

The four research questions related to test content were addressed through surveys completed by relevant expert groups. University faculty members (faculty) were asked to state whether each task assessed a skill within the identified domain of math or language arts.
Across all 115 tasks and four respondents, 85% of the responses indicated that the tasks did assess a skill in the identified domain.

Regarding the importance of the skill measured by each task for the students taking alternate assessment, ratings were provided by faculty, special education teachers (teachers), and parents of students with significant cognitive disabilities (parents). Raters read the name and a short description of each task and responded by indicating whether the task was important, minimally important, or not important. Across all 115 tasks and 28 raters, 78% of the ratings were Important, 17% were Minimally Important, and 5% were Not Important. Comparing the two domains, 71% of the math ratings, and 87% of the language arts ratings were Important. These same groups were asked to rate how well the group of tasks represented the target construct: important math/language art tasks for students with significant cognitive disabilities. The response options and results were as follows: (a) Yes, the set represents the construct very well. Few addi-

---

Observer: ____________________________

Student: _______ Teacher: _______ Task: ___ Original ___ Gen ___ Mx: ___

Trial 1 ___ Score: 1 ___ 0 ___ Student score on trial (by teacher rating) ___

Tester: _______ Location: _______ Other: _______

Comments:

Setup/typical routine: Yes ___ No ___ UK ___

Setup/no distractions: Yes ___ No ___ UK ___

Matches task description (e.g., appropriate materials, accurately interprets critical function of task)

Yes ___ No ___ UK ___

Instructions/acceptable: Yes ___ No ___ UK ___

Instructions/not repeated: Yes ___ No ___ UK ___

Prompts/acceptable: Yes ___ No ___ UK ___

Information above recorded for each trial.

Information below recorded across set of 3 trials.

Generalization: Yes ___ No ___

The following not scored for implementation fidelity:

Assessment record completed accurately: Yes ___ No ___

Proficiency level completed accurately: Yes ___ No ___

Proficiency level (by teacher rating):

UK = unknown; cannot tell from videotape

Figure 3. Procedural fidelity and scoring reliability checklist.
tional tasks needed (43% math, 75% language arts); (b) Somewhat, the set represents the construct reasonably well. Some additional tasks needed (38% math, 19% language arts); and (c) No, the set does not represent the construct. A substantial number of additional tasks are needed (19% math, 6% language arts).

The final research question evaluating test content was addressed to teachers. They read the entire task sheet (i.e., the title, description, setup, generalization, instructions, examples of the task, scoring criteria, and use of prompts), and responded “Yes” or “No” as to whether the procedures would adequately measure the skill. A sample of tasks was read by the teachers because the same format and general procedural approach are used for all tasks. Overall, 41% of the math tasks and 50% of the language arts tasks were evaluated, with each teacher reading one set of 12 to 14 tasks, and each set evaluated by 3 to 6 teachers. The procedures for 69% of the tasks were judged to be adequate to measure the task by all raters evaluating them. An additional 27% of the tasks were judged adequate by all except one rater.

Performance Standards

The three research questions addressing performance standards were also posed to the expert groups. UAA includes three criteria for trials: (a) the performance must be independent, (b) it must meet a specified performance level, and (c) it must be generalized across relevant variables (i.e., settings, people, sets of materials). Responding to each criterion separately, faculty and teachers rated the appropriateness of using these criteria for all UAA tasks. A majority of teachers (61%) indicated the independence criterion is appropriate, while faculty were split evenly across the ratings. All faculty and 67% of the teachers agreed with having a specified performance level. Finally, 83% of the teachers and half the faculty members agreed that the generalization criterion is appropriate. The other half of the faculty indicated more generalization should be required.

The same participants who rated the trial criteria rated the proficiency level criteria (i.e., 3/3 trials required for substantial proficiency, 2/3 for sufficient proficiency, 1/3 for partial proficiency, and 0/3 for minimal proficiency) as appropriate, too high, or too low. Overall, 77% of faculty and teacher participants indicated the criteria for assigning proficiency levels are appropriate; 9% stated they should be higher, and 9% stated they should be lower.

UAA procedures require that each student perform one math and one language arts task. The final question regarding performance standards asked if performance on two tasks is sufficient to determine adequate yearly progress for students with significant cognitive disabilities. Faculty, teachers, and parents responded to this question. All faculty and most parents (67%) responded that two tasks are not sufficient to determine adequate yearly progress; while the majority of teachers (88%) indicated administering two tasks is sufficient to determine adequate yearly progress.

Administration

To determine how reliably testers follow the administration procedures of the tasks, 234 videotaped trial administrations (127 math trials and 107 language arts trials) were systematically observed. A researcher-developed checklist (see Figure 3) was completed for each set of trials. Five aspects of each trial were assessed: (a) fulfilling the critical function of the task, (b) assessing within a natural routine, (c) presenting the prescribed directions to the student, (d) presenting the directions only one time, and (e) refraining from prompting the student during the assessed performance. Each trial was rated as meeting or not meeting the criteria on each aspect of administration. Technical difficulties (e.g., poor audio or video quality) prevented scoring of one specific aspect in 19 trials (8%). The aspect was scored as unknown in these trials. Figure 4 shows results for each of these five aspects of correct administration of trials. Twenty-nine math trials (23%) and 16 language arts (15%) trials did not meet the criteria for fulfilling the critical function of the task. In these trials, the student was asked to perform a task other than the identified task. Examples include the following:

(a) the identified task was responding to gestures, but the tester asked the student to
demonstrate asking for things politely;
(b) the identified task required exchanging different denominations of currency for the same total amount (e.g., a five-dollar bill for five one-dollar bills), but the tester asked the student to add various amounts of money and come up with a total;
(c) the identified task was rote counting, but the tester asked the student to skip count by twos; and
(d) the identified task was matching pictures to printed words, but the tester provided the student with words to match to words.

Since these trials that did not fulfill the critical function of the task were so far from the intended tasks, assessing the other aspects of reliable administration (e.g., presenting the prescribed directions, refraining from prompting the student during the assessed performance) would not be relevant. Therefore, the remaining aspects of reliable administration were assessed only for the 98 math trials and 91 language arts trials that fulfilled the critical function of the assessment task.

In the area of language arts, each aspect of administration was performed with fidelity on at least 86% of the trials. Fidelity on math tasks was somewhat lower and less consistent. At least 92% of the trials showed acceptable directions and directions presented only once. However, only 76% of the math tasks were administered within the natural routine and 78% were unprompted. An example of a task that was not administered within natural routines is assessing the “dollar more” strategy in the classroom, instead of assessing this task when the student is actually purchasing something at the school store or a neighborhood convenience store. Examples of unacceptable prompting include pointing to an error in written math facts and asking, “Did you check this one?”

Scoring

UAA includes an assessment record designed specifically for scoring each task. The researcher viewed each videotaped trial and completed the appropriate assessment record. Technical difficulties (e.g., poor audio or video quality) in 35% of the videos prevented scoring by the researcher, and these trials were not included in the scoring reliability set.
The overall reliability of scoring was 89%, with 96% attained on math trials and 84% on language arts trials.

Score Stability

The stability of the scores across time (i.e., test-retest reliability) and across testers, settings, and/or materials (i.e., generalization) was also evaluated. The three trials required for a task administration of UAA are referred to as a test; the initial test plus the test-retest or generalization test are referred to as a set. Fourteen sets were assessed for test-retest reliability and 12 sets were assessed for generalization. Proficiency levels on nine of the 14 test-retest sets (64%) were the same on both the initial and the second tests; nine of the 12 generalization sets (75%) also showed exact matches. Thus, scores were stable for 69% of all sets. The adequate yearly progress rating was stable on 23 of the 26 sets (88%) assessed for stability.

Decisions/Impact

The potential consequences of implementing this assessment system were evaluated by asking district special education administrators to respond to open-end questions regarding the likely effects of implementation of UAA on curriculum and instructional decision-making/planning and the likely effects on district and state level planning. They were asked to provide both positive and negative potential outcomes. Participants each identified one to four effects on curriculum and IEP development that they considered to be likely. Sixteen of the identified effects were positive (64%), five were negative (20%), and in four responses (16%) it was unclear whether the respondent considered the outcome positive or negative. These responses are summarized in Table 3. When asked about the potential effects on district and state level planning, the need for more training in the core curriculum and on teaching for generalization was noted (see Table 4). Finally, the participants were asked what changes had already taken place that they attributed to UAA implementation. Responses included that all students are now included in testing and teachers and administrators feel more accountable for student progress, the district is more supportive of obtaining appropriate curricular materials, and teachers are recognizing the correlation

<table>
<thead>
<tr>
<th>Positive effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>• May improve IEP goal writing skills (e.g., include more functional academic goals, include more specific criteria, guide data collection) (N = 5)</td>
</tr>
<tr>
<td>• May increase teaching generalization of skills (N = 4)</td>
</tr>
<tr>
<td>• May increase data collection (N = 3)</td>
</tr>
<tr>
<td>• Process may generalize to areas outside UAA (e.g., development of all IEP goals) (N = 2)</td>
</tr>
<tr>
<td>• May increase focus on academics (N = 1)</td>
</tr>
<tr>
<td>• May increase access to general curriculum (N = 1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>• UAA may drive the IEP (N = 3)</td>
</tr>
<tr>
<td>• May emphasize academics over life skills (N = 2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effects not identified by participants as positive or negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Cues teachers to include math and reading goals on IEPs (N = 1)</td>
</tr>
<tr>
<td>• Small districts have difficulty serving students with significant cognitive disabilities (N = 1)</td>
</tr>
<tr>
<td>• Instruction may be directed toward UAA tasks (N = 1)</td>
</tr>
<tr>
<td>• May require better collaboration between special and general education (N = 1)</td>
</tr>
</tbody>
</table>
between data and curriculum. These responses are summarized in Table 5.

**Discussion**

Validity is viewed as a unified concept, yet it is useful to differentiate specific aspects in order to focus on issues that might be overlooked when viewing it as a whole (Messick, 1995a). Six specific aspects of UAA were addressed in this validity examination. On one hand, the evidence for each aspect must stand on its own; yet on the other hand all aspects must be evaluated together to judge the overall validity of UAA for its intended purposes as part of a state educational accountability system. It is necessary that each aspect (e.g., test content, scoring) be judged adequate, but examining any aspect alone is clearly not sufficient for the assessment as a whole to be judged as having a high degree of validity. For example, attaining very high scoring reliability would not compensate for inadequate test content. Each aspect that was examined is briefly discussed next, followed by an overall analysis of the validity of UAA.

The four research questions that examined the validity of UAA’s test content provide an overall indication that the task sets do consist of important math and language arts tasks, and there is strong agreement that the measurement procedures are adequate to assess the skills. There is also strong agreement that the sets of tasks are representative of the constructs. While these ratings were encouraging, they were not unanimous. Thus, they raise issues regarding retaining or deleting tasks based on relevance to the construct and importance for the target population that must be addressed. These questions suggest a need for on-going research that includes a strategy to continue to gather feedback from relevant constituencies. Another important consideration here is that the tasks must be aligned to extended standards. Utah had not yet developed extended standards with which to align the tasks at the time this system was implemented, thus the expert judgment process was used. Once extended standards are in place, it will be critical to complete an alignment process.

Three research questions addressed the performance standards of UAA. Independent performance is one criterion because when students can perform more behaviors independently, they will have more control over the events in their lives and more vocational and living options will be available to them. This rationale was accepted by more than half of all respondents. Teachers who indicated this criterion was too high, also tended to respond that specified performance levels was also inappropriate. It appears that these teachers feel these criteria should be established individually for each student. The one university faculty member who indicated this criterion is too low, commented that maintenance should also be required, and one teacher also referred to maintenance as being an important indicator. Whether the students continue

---

**TABLE 4**

Likely Effects of Implementation of UAA on District and State Level Planning

<table>
<thead>
<tr>
<th>Likely effects of implementation of UAA on staff development/inservice training</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Little/no impact ($N = 4$)</td>
</tr>
<tr>
<td>• Annual UAA training will be necessary: for special education teachers ($N = 3$) and for general education teachers ($N = 2$)</td>
</tr>
<tr>
<td>• Increased training on the core curriculum ($N = 2$)</td>
</tr>
<tr>
<td>• Need to focus on generalization ($N = 1$)</td>
</tr>
<tr>
<td>• Unsure ($N = 1$)</td>
</tr>
</tbody>
</table>

---

**TABLE 5**

Changes that Administrators Attribute to UAA Implementation

<table>
<thead>
<tr>
<th>Changes that have already taken place that are attribute to implementation of UAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Teachers and administrators feel more accountable for student progress ($N = 2$)</td>
</tr>
<tr>
<td>• District is more supportive of obtaining curricular materials ($N = 2$)</td>
</tr>
<tr>
<td>• All students are now included in testing ($N = 1$)</td>
</tr>
<tr>
<td>• More focus on the core curriculum ($N = 1$)</td>
</tr>
<tr>
<td>• Teacher attitudes toward UAA have improved ($N = 1$)</td>
</tr>
<tr>
<td>• Teachers are choosing tasks that can be tested instead of IEP goals ($N = 1$)</td>
</tr>
</tbody>
</table>
to perform the behaviors measured on UAA after the testing period is clearly a factor in the validity of UAA. It was not addressed in this study, but is a relevant issue for future research. When UAA has been implemented for several years, there will be opportunities to systematically investigate the degree to which skills initially demonstrated on UAA are retained and used across time.

There was a very high degree of support for providing a performance criterion. More than one-quarter of the teachers (28%) disagreed with the concept of a standardized performance level (with some commenting that performance levels must be individually established for each student), but none of the faculty disagreed with this concept. This research question addressed the concept of having a specified performance level for each task, but did not address whether the particular performance levels provided were appropriate. Overall, the concept of requiring students to meet specified performance levels was supported (73% of the participants agreed with the concept), but future research should address the appropriateness of the performance level required for each task.

In order to demonstrate adequate yearly progress, the student must attain Substantial (meet all criteria on all three trials) or Sufficient (meet all criteria on two out of the three trials) on UAA tasks. The rationale for these scoring criteria is that students should be required to meet all criteria on at least two trials because this constitutes some evidence of generalization. Overall, this approach to determining proficiency levels received strong support, and changes do not appear warranted.

UAA requires that IEP teams select one language arts task and one math task for each student. The validity question this raises is one of construct underrepresentation - whether a meaningful judgment of student achievement can be made from this relatively small sample of behavior. Faculty, teachers, and parents were asked whether performance on two tasks is a sufficient basis from which to determine adequate yearly progress. In response to this question, 100% of the faculty participants and 67% of the parent participants stated that it is not an adequate measure. These two participant groups clearly have strong reservations about this aspect of UAA. The response from teachers, however, was quite different. Eighty-eight percent of the teachers responded that administering two tasks is an adequate measure. This question elicited very strong responses on both sides. A potential explanation is that teachers likely have a great deal of information on student academic achievement in addition to UAA (thus feeling the small sample from UAA is sufficient), but the fact remains that only the information from UAA is used to determine adequate yearly progress. This being the case, the negative judgment of this aspect of UAA given by faculty and parents must be taken very seriously.

A related question that was not addressed in this study is the appropriateness of the tasks selected for each student. The IEP team selects tasks for each student and the results of the student’s performance on these tasks are used to determine whether the student has demonstrated adequate yearly progress. However, no evidence has been gathered regarding the validity of the IEP teams’ choices of tasks for individual students. For example, no baseline data are collected to determine whether or not the student can already perform the task at a mastery level when tasks are selected. This is a critical issue that potentially threatens the validity of UAA and should be investigated in future research.

Assessment tasks must be administered correctly in order for the results to reflect the focal construct. The students’ performances can be greatly influenced by how closely the teacher adheres to the administration procedures (Crooks et al., 1996). If there are aspects of test administration for which the fidelity is judged to be unacceptably low, steps must be taken to improve it. There are no standard criteria for evaluating fidelity of test administration, but the current study and others addressing the validity of performance-based assessments may begin to provide criteria. As a starting point, a standard of 85% fidelity seems reasonable as an overall criterion.

Five critical aspects of task administration were assessed: (a) fulfilling the critical function of the task, (b) assessing within a natural routine, (c) presenting the prescribed directions to the student, (d) presenting the directions only one time, and (e) refraining from prompting the student during the assessed
performance. In practice, fulfilling the critical function of the task means that what the teacher had the student do matches the description of the task. If the teacher completed the assessment record for task 145 (completes math problems with addition – symbolic form), did the student in fact complete addition problems? If the student did something other than complete addition problems, then the score reported will appear to provide information on the student’s ability to complete addition problems, yet it will not actually do so. Varying too much from the specified directions may make the task easier or harder. The directions are to be provided only once because in both school and work environments individuals are expected to respond to instructions the first time they are presented. The final aspect, refraining from prompting the student, is important because prompting would violate the independence criterion.

The aspect of implementation that raises the greatest concern is that only 77% of the math trials and 84% of the language arts trials fulfilled the critical function of the task. In some cases, the student appeared to perform a task that was more difficult than the identified task, but regardless, the performance did not correspond with the identified task. If meaningful inferences are to be made about student achievement, it is critical that the student perform the identified task.

Potential explanations for the low fidelity include (a) instructions for administration of assessment tasks were not written clearly enough for teachers to interpret, (b) this is the first year of implementation of this form of UAA, so questions may not have come up until teachers were actually administering the tasks, and (c) training on implementation of this assessment system was not adequate. All three may well have contributed to problems observed in fidelity of test administration because each interacts with the others. For example, more clearly written tasks may lessen the training requirements, and more in-depth training may enable teachers to better administer the tasks as written. Clearly, fidelity of test administration is a significant weakness in the validity of UAA. A combination of changes including clarification of the written materials, improvement in training, and provision of other supports would likely improve fidelity and the overall validity of UAA. In addition, fidelity of test administration should be monitored closely as these changes are put into place.

Overall, scoring reliability does not appear to be a serious concern for UAA (overall reliability of scoring was 89%), although it should continue to be monitored. Procedures could be incorporated into the training sessions to increase scoring reliability. For example, viewing videotapes of UAA administrations during training sessions and discussing the specific details regarding why a trial does or does not meet the identified criteria may be useful for increasing scoring reliability.

One purpose of assessment is to make inferences to behavior beyond the specific instances observed during assessment. Evidence that it is reasonable to make these inferences must be collected and evaluated (Messick, 1995a). UAA tasks consist of three trials assessing a single skill, under varying conditions. The results for both test-retest and generalization allow reasonable confidence that one set of trials is appropriate for judging the student’s achievement in that skill.

It is important to assess the consequences of any assessment system to identify evidence of positive and negative consequences. There is clearly potential for both positive and negative consequences from implementing alternate assessment, but without careful evaluation these consequences may not be identified and attributed to the assessment system. It is important to identify the consequences that result from the assessment system so that the assessment can be modified as necessary to optimize the positive consequences and minimize the negative consequences. At the time of this research, UAA had only been in place for one school year, so it was difficult to gather direct evidence of consequences.

Examination of the potential positive and negative consequences of use of UAA offered several insights. The numerous comments that referred to UAA’s anticipated impact on the specificity of IEP goals (5 comments), the emphasis on teaching for generalization (4 comments), and the amount of data collection (3 comments) were unanimously positive. Three participants indicated it would be a positive outcome if the criteria for successful performances (independence, a specified per-
formance level, and evidence of generalization) were applied to other IEP goals and if this type of data collection was used across the curriculum. These may be examples of the test driving the curriculum in a positive way.

The most common consequences identified by respondents referred to potential for UAA to drive curriculum, IEP development, and allocation of instructional time. Interestingly, these consequences were interpreted by some participants as positive and by other participants as negative. Numerous comments suggested that UAA may drive a shift in curricular focus in the direction of greater emphasis on academics (i.e., general curriculum). However, respondents differed on whether this shift would be a positive development and whether increased emphasis on academics comes at the detriment of life skills. For example, one respondent suggested a positive effect of “increased focus on academics” and another hoped it would promote increased access to the general curriculum; in contrast, two respondents expressed concern that UAA would cause educators to “emphasize academics over life skills.” Similarly, with respect to IEP development, several respondents anticipated “improved IEP writing including more functional academic skills” and another suggested that UAA could “cue teachers to include math and reading goals.” On the other hand, three respondents saw as negative that “UAA may drive the IEP,” and expressed concern that the IEP would be less individualized. Thus, the difference between positive and negative implications of UAA appears to be largely a matter of the respondent’s judgment about whether such a shift would come at the expense of life skills and their values with respect to academic versus life skills curriculum for students with significant cognitive disabilities.

The potential for academic skills to displace functional life skills is a serious concern that should be addressed. It is critical to ensure that the most important outcomes of the educational program are assessed because there is a potential for these assessment systems along with the contingencies of legislation (e.g., NCLB) to impact the curriculum and IEP. It has been noted “what gets tested, gets taught” (Burgess & Kennedy, 1998). The question, of course, is what are the most important outcomes of educational programs for students with significant cognitive disabilities? Functional life skills have always been a central part of the curriculum and IEPs for students with significant cognitive disabilities (Spooner & Test, 1994) and their complete omission from the accountability testing for these students is something that must be critically examined. If teaching life skills will not improve scores on alternate assessments, instructional time that has been devoted to these skills may be reallocated to teaching skills that directly improve alternate assessment scores. There should be discussion about the relative importance of various domains and how to balance academic and functional life skills to best meet the needs of students participating in alternate assessment.

Conclusions

This study evaluated the validity of UAA for its use as part of a state educational accountability system. Data collection and analysis focused on six aspects of UAA. The strongest of these aspects of validity are test content, scoring reliability, score stability, and impact of test use. Fidelity of test administration clearly needs to be improved. In addition, many respondents expressed serious reservations regarding the validity of using a sample of only two tasks for determining whether students demonstrate adequate yearly progress. These aspects of UAA can be modified and strengthened without disrupting the basic structure of the assessment.

Test validation is a continuing process of gathering and evaluating evidence to make a judgment of the degree to which the test accomplishes specific purposes. The evidence presented in this study provides strong support for the validity of some aspects of UAA and identified other aspects that could be strengthened further. As changes are made in UAA based on this study and others like it, evidence will again need to be gathered and evaluated to assess the effects of these changes. Test validation is an ongoing process, and this study provides a strong foundation on which future studies can be based. Further, this study provides a model that can be applied to studying the validity of alternate assessments in other states.
References


Received: 11 October 2006
Initial Acceptance: 15 December 2006
Final Acceptance: 1 August 2007