

# Building Validity Evidence for Scores on a State-Wide Alternate Assessment: A Contrasting Groups, Multimethod Approach

Stephen N. Elliott, *Vanderbilt University*, Elizabeth Compton, *Boise State University*, and Andrew T. Roach, *Georgia State University*

*The relationships between ratings on the Idaho Alternate Assessment (IAA) for 116 students with significant disabilities and corresponding ratings for the same students on two norm-referenced teacher rating scales were examined to gain evidence about the validity of resulting IAA scores. To contextualize these findings, another group of 54 students who had disabilities, but were not officially eligible for the alternate assessment also was assessed. Evidence to support the validity of the inferences about IAA scores was mixed, yet promising. Specifically, the relationship among the reading, language arts, and mathematics achievement level ratings on the IAA and the concurrent scores on the ACES-Academic Skills scales for the eligible students varied across grade clusters, but in general were moderate. These findings provided evidence that IAA scales measure skills indicative of the state's content standards. This point was further reinforced by moderate to high correlations between the IAA and Idaho State Achievement Test (ISAT) for the not eligible students. Additional evidence concerning the valid use of the IAA was provided by logistic regression results that the scores do an excellent job of differentiating students who were eligible from those not eligible to participate in an alternate assessment. The collective evidence for the validity of the IAA scores suggests it is a promising assessment for NCLB accountability of students with significant disabilities. The methods of establishing this evidence have the potential to advance validation efforts of other states' alternate assessments.*

**Keywords:** alternate assessment, contrasting groups validity design, inclusive assessment

## *Definitions and Characteristics of Alternate Assessments*

During the past several years, the educational achievement of thousands of students with significant disabilities in

the United States has not been accounted for in a meaningful and statistically sound manner—even though a vast majority of these students have participated in state-level alternate as-

sessments. Alternate assessment is a generic term for a family of methods used to assess the academic performance of students with significant disabilities or limited proficiency with English. According to the U.S. Department of Education, “An alternate assessment must be aligned with the State’s content standards, must yield results separately in both reading/language arts and mathematics, and must be designed and implemented in a manner that supports use of the results as an indicator of AYP (adequate yearly progress)” (USDOE, 2004, p. 15). Alternate assessments are an important component of each state’s assessment system and, as such, are required to meet the federal regulations outlined in Title I of the Elementary and Secondary Education Act.

According to the USDOE nonregulatory document on *Alternate Achievement Standards for Students with the Most Significant Cognitive Disabilities* (August, 2005), alternate

---

*Stephen N. Elliott, Dunn Family Professor of Educational and Psychological Assessment, Peabody #59, 401 Wyatt Center, Vanderbilt University, Nashville, TN 37203-5701; steve.elliott@vanderbilt.edu.*

*Elizabeth Compton, Consultant, Southwest Regional Special Education, E-522, 1910 University Drive, Boise State University, Boise, ID 83725-1725; eberman@boisestate.edu.*

*Andrew T. Roach, Assistant Professor, Department of Counseling and Psychological Services, College of Education, Georgia State University, P.O. Box 3980, Atlanta, GA 30302; aroach@gsu.edu.*

assessments must meet standards of high technical quality—validity, reliability, accessibility, objectivity, and consistency—expected of other educational tests (i.e., *Standards for Educational and Psychological Testing*, American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999). In addition, alternate assessments must have an explicit structure, guidelines for determining which students may participate, clearly defined scoring criteria and procedures, and a report format that communicates student performances in terms of academic achievement standards. If the required standards for technical quality and use are met, then the results of alternate assessments for up to 1% of the total student population can be reported for AYP purposes.

Three approaches to alternate assessments of students with significant cognitive disabilities are prevalent across the country. These approaches are labeled Portfolio Assessment, Performance Assessment, and Comprehensive Rating Scales of Achievement. Other terms (e.g., body of work, performance tasks/events, checklist) have been used to label the various approaches to alternate assessment, but these three terms represent the consensus in the literature (Elliott & Roach, in press). Although different in name and a number of other attributes, each of these three approaches to alternate assessment, in general, requires the collection of evidence samples—e.g., classroom work products, video/audio recordings, interviews, structure observations—to characterize students' knowledge and skills that are determined to be aligned or linked to state grade level content standards. The evidence samples must then be evaluated and scored to yield data that can be summarized by a proficiency level descriptor based upon a rigorously established set of grade-level achievement standards.

All alternate assessments, regardless of approach, must be demonstrated to yield reliable and valid results. Thus, a systematic plan to establish and maintain evidence about the validity of alternate assessment results must become part of each state's assessment system. Data that impact the reliability and validity of assessment decisions must be collected periodically to provide confidence in the results for users

and external reviewers. The design of alternate assessment validity studies and interpretation of the subsequent results should be strongly influenced by the technical requirements for high-quality assessments. The key outcome effectiveness data for such studies is psychometric evidence from a series of validity studies of alternate assessment scores for students with significant disabilities. The validity evaluations of alternate assessment scores must be driven by the design imperatives for technically sound assessments and interpreted against consensus standards for high-quality tests as conceptualized by classic and item response theories. All high-quality assessments should possess the characteristics of validity, reliability, and usability. Ultimately, a statement about the validity of an assessment involves an evaluative judgment of the degree to which interpretations and uses of the assessment results are justified (Kane, 2002).

#### *Research on the Validity of Alternate Assessment Results*

The validity and utility of virtually all alternate assessments has been questioned. For example, Johnson and Arnold (2004) published the first study on validating an alternate assessment portfolio. Specifically, they examine the validity of the Washington (state) Alternate Assessment System (WAAS) and concluded that the results "indicated serious shortcomings in the evidence for content, response process, and structural validity" (p. 266). They went on to identify a number of sources of invalidity for the WAAS, including (1) some portfolios did not measure the state content standards, (2) teachers' ability to assemble a portfolio contributed greatly to students' scores, and (3) a student's total portfolio score was mainly determined by the generalization skills scores, yet the basis for this score was unclear. Johnson and Arnold (2004) concluded, "It appears that the portfolio is currently more a reflection of a teacher's ability to compile a portfolio according to the guidelines than an accurate measure of a student's progress toward the IEP skill or of a program's success in getting students to access content standards. . . . (T)he alternate assessment's utility as a state-level programming tool . . . is questionable" (p. 273).

Browder, Fallin, Davis, and Karvonen (2003) expressed concerns with

performance-based approaches and suggest that the technical characteristics of these alternate assessments may negatively influence students' and schools' outcome scores. Initial data from Kentucky's efforts suggest that reliability of scores may be a source of challenge for states' portfolio-based alternate assessments (Browder et al., 2003). Challenges with the reliability of ratings were also observed by states attempting to use portfolios and performance assessments as part of their general large-scale assessment systems (e.g., Vermont and Arizona). These difficulties resulted in states' inability to publicly report assessment results (Koretz, McCaffrey, Klein, Bell, & Stecher, 1993; Tindal et al., 2003). Moreover, to demonstrate adequate alignment to state standards, performance assessments may need to include numerous tasks and work samples, resulting in an extensive and time-consuming assessment process. Browder et al.'s (2003) review also identified student risk factors (e.g., instability of student behavior or health status) as potential influences on students' alternate assessment results. In the case of on-demand performance tasks, fluctuations in student behavior or physical well-being could potentially result in inaccurate and invalid assessment results.

Extended Reading and Math Tasks, as described by Tindal et al. (2003), represent a performance task or event approach. Using curriculum-based measurement (CBM) technology, this approach consists of a continuum of tasks that measure students' basic skills in reading and mathematics. An extensive literature on the validity and utility of CBM for monitoring students' academic progress provides support for this approach (Shinn, 1995). By including assessment tasks at a range of skill levels, this alternate assessment strategy allows test users to individualize the assessment by administering only those tasks that are considered appropriate to the student's current skills and instructional experiences (Tindal et al., 2003).

To date, there only have been a few published studies where scores from alternate assessments were correlated with other existing measures of instructional programs or performance for validity analysis purposes. In one study, Turner, Baldwin, Kleintert, and Kearns (2000) examined the correlation between students' alternate

assessment performance, as measured by the Kentucky Alternate Portfolio, and two measures of education effectiveness: (a) quality of the student's educational program (as measured on the Program Quality Indicator (PQI) Checklist; Meyer, Eichinger, & Downing, 1992) and (b) the quality of the student's IEP goals and objectives. The results confirmed a moderately strong and statistically significant relationship between program quality indicators and alternate assessment performance ( $r = .50$ ), but did not support a relationship between the quality of students' IEPs and their alternate assessment scores.

A second study (Kampfer, Horvath, Kleinert, & Kearns, 2001) examined the relationship between a variety of variables and students' scores on the Kentucky Alternate Portfolio. Of particular note were the correlations between students' involvement in the portfolio process and their portfolio score ( $r = .42$ ) and between portfolio scores and the extent to which the portfolio items were "embedded" in student's daily curriculum and instruction ( $r = .37$ ). A subsequent hierarchical regression analysis grouped variables into three categories: teacher variables (years teaching, years working with the portfolio process, and participation in scoring and training activities); instructional variables (extent to which the portfolio was embedded in instruction, student involvement in the process, and teachers' perceptions of the education benefit) and time (hours spent on portfolio). This model accounted for 27.5% of the variance in students' portfolio scores with instructional variables accounting for the majority (24.1%) of this variance. Student demographic variables appropriately did not contribute to the variance in their scores on the alternate assessment.

In summary, serious questions have been raised about using the results of statewide alternate assessments for (a) monitoring educational performance (status and/or progress) at the levels of student, classroom, school, and system and (b) making decisions about curriculum and instruction. The evidence reported in peer reviewed journals has been limited, but what has been published generally has been most negative about portfolio assessments. Performance assessment and comprehensive rating scale approaches have fared better largely because they have more aligned items, sample more discrete knowledge and

skills, and fit a quantitative validity evidence paradigm reasonably well (Roach, Elliott, & Webb, 2005).

### *Present Study*

The Idaho Alternate Assessment (IAA) was designed in 1999 and first implemented during the 2000–01 school year. The IAA is a comprehensive rating scale and was one of the first alternate assessments approved by the USDOE in 2006. The purposes of the IAA are to document students' progress towards Idaho's Academic Content Standards and to facilitate compliance with IDEA and NCLB regulations concerning the full inclusion of students with severe disabilities in statewide assessment systems. Students with a disability who meet the criteria for participation in the alternate assessment have been determined by an IEP team to be performing at a skill level significantly below that expected of grade-level peers in the general education curriculum even with appropriate accommodations.

Individual alternate assessments exist for all the content areas and grade levels identified in the Idaho assessment program. These content areas (i.e., reading, language arts, and mathematics) are all assessed using classroom-based evidence and behavior rating scale technology to document teachers' judgments of the proficiency with which a given student exhibits desired knowledge and skills. Currently, the results of the assessment are reported as one of four achievement levels: Below Basic, Basic, Proficient, or Advanced. Students achieving at the Proficient or Advanced Levels on the IAA are counted as Proficient for purposes of adequate yearly progress (AYP) calculations for NCLB reporting.

This study focused on the concurrent and discriminant validity of the inferences made about IAA test scores. Validity refers to the adequacy and appropriateness of the score interpretations made from assessments, with regard to a particular use. Criteria for evaluating the validity of score inferences from tests and related assessment instruments have been written about extensively, yet few studies about the validity of alternate assessment scores have been published. A joint committee of the American Educational Research Association, the American Psychological Association, and the National Council

on Measurement in Education in 1999 revised their comprehensive list of standards for tests that stresses the importance of construct validity and describes a variety of forms of evidence indicative of a valid test. The revised *Standards for Educational Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999) include valuable information for educators involved in testing diverse groups of students, including students with severe disabilities. In addition to the *Testing Standards*, the U.S. Department of Education (April 2004) published the *Standards and Assessment Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001*. This later document extends the *Testing Standards* and provides even more specific guidance concerning validity evidence for state alternate assessments like the IAA. For example, the Technical Quality subsection [4.1] of the Peer Review Guidance document specifically asks,

(b) Has the state ascertained that the assessments, including alternate assessments, are measuring the knowledge and skills described in its academic content standards and not knowledge, skills, or other characteristics that are not specified in the academic content standards or grade level expectations?

(e) Has the State ascertained that test and item scores are related to outside variables as intended (e.g., scores are correlated strongly with relevant measures of academic achievement and are weakly correlated, if at all, with irrelevant characteristics, such as demographics)? (p. 35)

The present study was a direct examination of the relationships between ratings on the IAA for students with significant disabilities and their corresponding ratings on two norm-referenced teacher rating scales. One measures academic competence (i.e., *Academic Competence Evaluation Scales* or *ACES*), and the other measures adaptive behavior (i.e., *Vineland Adaptive Behavior Scales*). In addition, the relationship of the IAA scores to students' demographic characteristics was also tested. Thus, the present study examined students' performances on the IAA for a representative group of students with disabilities, who according to their IEP teams were

*eligible* and actually participated in the state's alternate assessment and another group of students who had disabilities, but were *not officially eligible* for the alternate assessment. Both groups of students were assessed with an alternate assessment and their results were compared to other indirect assessments of performance, all of which were completed by students' teachers. The students who were not eligible for the alternate assessment concurrently took the state's regular assessment (ISAT) with accommodations.

This study was designed to address several of the critical elements [specifically 4.1 (b), (d), (e), and (g)] of the Technical Quality section in the USDOE's *Standards and Assessments Peer Review Guidance* document (April, 2004) and was influenced by classical test theory practices where understanding the construct(s) measured by a new test is advanced through comparisons with established tests that measure validated constructs. This methodological approach to validity evidence is commonly referred to as a multitrait multimethod (MTMM) design; however, in the assessment of students with severe disabilities it is extremely difficult, if not impossible, to use methods of assessment that do not utilize third party observations or evaluations to a significant degree. Thus, the multimethod aspect of the MTMM approach was compromised, and as a result there is the potential for more common or shared variance due to the use of the same informant to report his or her observations via rating scales.

As conceptualized by Campbell and Fiske (1959), the MTMM approach allows for an integrative multivariate framework within which information about convergent and discriminant evidence is systematically gathered in a single study. All common assessment instruments in this study were teacher judgment tools using rating scale technology to formalize scores for students' performances. It was expected that students' IAA scores for the Reading, Language Arts, and Mathematics Scales would correlate minimally with demographic characteristics, but strongly with their Academic Competence Evaluation Scales (ACES) Academic Skills scores and moderately with their ACES Academic Enabler scores. Similarly, it was expected that students' IAA scale scores would correlate moderately with the Vineland Adaptive Behavior Scales

(VABS) Communication and Daily Living Skills scales, but only minimally with their Vineland Socialization scale given the academic focus of the IAA scales. Finally, it was expected that students who were not eligible for an alternate assessment would score significantly higher on the IAA than students who were eligible and their IAA scores would correlate moderately to their ISAT test results. Collectively, the evidence from testing these predictions provided new information about the validity of the IAA and, perhaps more importantly, illustrates methods for gathering evidence for the scores of new alternate assessments required by all states.

## Method

### *Participants*

Two samples of students and their respective teachers participated in this investigation. The two samples selected represented 80 randomly selected schools from 35 school districts across the state of Idaho. The first and larger sample comprised students with significant cognitive disabilities ( $N = 116$ ) who were *eligible* to take the IAA. This sample was selected to address issues concerning the concurrent validity of the IAA with the *ACES* and *Vineland Adaptive Behavior Scale*. The smaller student sample ( $N = 54$ ) included students with disabilities determined to be *not eligible* for the IAA given their disabilities, state alternate assessment participation criteria, and their present educational programs. These students typically would be administered the state's large scale assessment—the ISAT—with accommodations. This sample was selected to allow a direct test of the concurrent and discriminant validity of the IAA with the *ACES* and also the IEP teams' participation decisions.

The population of students in Idaho in 2004 with significant disabilities who qualified to take the IAA was 976. Of these qualified students, 863 took the assessment, resulting in a nonparticipation rate for this assessment of 11.6%. This nonparticipation rate is compared to the general education test where in 2004 nearly 140,000 students were eligible to participate, but less than 1% of them failed to do so.

*IAA eligible participants.* The participants who were eligible for the IAA

were 116 students identified with a severe disability in grades kindergarten through 10th from across the state of Idaho. Of the total number of students, 54 were elementary students (K-5th grade), 29 were middle school students (grades 6–8), and 33 were high school students (grades 9–10). The sample was comprised of 63 male and 53 female students; nearly the entire sample (103 or 88.8%) were Caucasian; of the other ethnic/racial groups, only the Hispanic group (6.9%) had more than one student. Of the 116 eligible students, 113 (97.4%) were officially classified having a cognitive impairment (38%), deaf/blindness (4.3%), multiple disabilities (48%), autism spectrum disorder (4.3%), or traumatic brain injury (2.6%).

Each of the participating students was assessed by his/her primary teacher ( $N = 116$ ) with consent of the students' parents. Teachers volunteered to participate and were paid a small honorarium for their time.

*Non-eligible participants.* The participants who were deemed to not be eligible for the IAA by their IEP teams included 54 students (34 male, 20 female) identified with disabilities (learning disabilities 78%, emotional disturbance 9.2%, hearing impairment 3.7%, and missing information on disability 9.2%). None of these students were officially eligible to participate in the IAA because their disabilities were not so severe as to prevent participation with accommodations in the regular statewide test. This sample of students was randomly selected and characterized as 82% Caucasian and nearly equally distributed from grades 1, 4, 8, and 10.

### *Materials*

*Academic Competence Evaluation Scales (ACES; DiPerna & Elliott, 2000).* The *ACES* was designed to measure students' skills, attitudes, and behaviors that contribute to academic competence (DiPerna & Elliott, 2000). The teacher version of the *ACES* used in this study was an 81-item questionnaire with two separate scales (Academic Skills and Academic Enablers), and each of these scales included multiple subscales. Specifically, the Academic Skills Scale includes three subscales (Reading/Language Arts, Mathematics, and Critical Thinking), and the Academic Enablers Scale included four

subscales (Interpersonal Skills, Motivation, Study Skills, and Engagement). Teachers are asked to provide two five-point ratings for each item: Frequency (or Proficiency) and Importance. Items on the *ACES* are phrased in a positive manner. Teachers rate items in terms of the level of the students' academic skills compared with grade level expectation from "1 (Far Below)" to "5 (Far Above)." Teachers rate the existence/ frequency of academically enabling skills from "1 (never)" to "5 (almost always)." Normative frequency scores according to developmental level K-2, 3-5, 6-8, 9-12 are provided for the interpretation of individual student's scores. A 3-level rating (1 = not important, 2 = important, and 3 = critical), as well as an N/O (not observed) option is available for the teacher to document the perceived importance of the skills/ behaviors on the scale.

The internal consistency index coefficient  $\alpha$  for *ACES-Teacher* has a mean of .99 on the Academic Skills and Academic Enablers scales across grade levels. The test-retest reliability of *ACES-Teacher* form over two to three weeks' interval ranges from .88 to .97. The reported standard error of measurement for the Academic Skills scale ranges (depending on grade level) from 2.5 to 3.1 and for the Academic Enablers scale ranges from 3.6 to 4.7. The developers also examined validity evidence based on test content, internal structure, relationship with other variables, and the consequence of testing. Two subscales, Academic Skills and Academic Enablers, were derived from factor analysis. In relation to other standardized achievement and behavior measures, *ACES* also demonstrated solid evidence for the convergent, discriminate, and test-criteria validity.

*Vineland Adaptive Behavior Scales* (Sparrow, Balla, & Cicchetti, 1985). The *Vineland* (VABS) was designed to assess individuals with and without disabilities from birth to adulthood in four domains: Communication, Daily Living, Socialization, and Motor Skills. The Classroom edition form used in this study has 244 items. The *VABS* is a widely used instrument and was standardized on 3,000 individuals ranging in age from birth to 19 years and representative of a diverse demographic population. Reliability of the *VABS* is adequate

for the overall four domains, but poor for some of the subscales within each domain. Median split-half reliability coefficients across ages range from .83 for Motor Skills to .90 for Daily Living Skills. Inter-rater reliability for the domains is lower and ranges from .62 to .78. The standard error of measurement for the various scales ranges (depending on age) from 3.4 to 6.6. Validity data in the test manual is limited, but a number of studies since its publication have established that it correlates strongly with other measures of adaptive behavior but much lower with measures of intelligence (Witt, Elliott, Daly, Gresham, & Kramer, 1998).

*Idaho Alternate Assessments* (Idaho Department of Education, 1999). The *IAA* for students with significant disabilities is a set of teacher rating scales focusing on the content areas of reading, language arts, and mathematics. The item content for the *IAA* was developed by an item development workgroup of special educators. The content was determined to be well aligned with the state's general education content standards (Roach, 2003). Teachers are required to rate evidence collected for *IAA* items that align with students' IEP objectives using two criteria: achievement level (four levels: Nonexistent/Beginning, Emerging, Developing, and Generalized) and progress level (four levels: Beginning, Little, Good, and Excellent). These two dimensions of performance, Achievement and Progress, are integrated into a 16-point scoring framework that ranges from a Nonexistent-Beginning to Generalized-Excellent. In the present study, however, we primarily examined the relationships among *IAA* achievement level scores and achievement or proficiency scores on the other measures to enhance the comparability of the nature of the score comparisons. For these comparisons, raw scores on the *IAA* subscales of Reading, Language Arts, and Math were used. (Note that the Achievement Level Ratings correlate highly with Progress Level Ratings: for Reading  $r = .98$ , for Language Arts  $r = .94$ , and for Mathematics  $r = .97$ ). For the analysis where we compared the scores of eligible to not eligible students, we used the combined *IAA* (achievement and progress) scores.

The *IAA* system consists of the following components:

- Reading—Idaho Alternate Assessment Grades K-10. This form of the *IAA* is designed to measure reading, listening, and viewing skills of students chronologically placed in grades 3 through 11. The rating form consists of 12 items.

- Language—Idaho Alternate Assessment Grades 2-10. This scale contains six (6) alternate knowledge and skill items for the Idaho Achievement Standards in writing and speaking. The emphasis is on how a student learns to be a successful expressive communicator in writing and speaking.

- Mathematics—Idaho Alternate Assessment Grades 2-10. This scale contains eighteen (18) alternate knowledge and skills for Idaho Achievement Standards in the categories of number sense, computation, reasoning and problem solving, measurement, geometry, and math models and functions. The emphasis is on using the basic concepts of numbers in functional daily and vocational skills.

The reliability of the *IAA* rating scales has been established using both coefficient alphas to indicate the degree of internal consistency and also inter-rater reliability of item-level ratings, that is, how well first and second raters agreed when they rated the proficiency level of the evidence provided for each test item. The *IAA* Administration Guide (undated) states that second raters are required and that "at a minimum, an agreement of 80% must be reached between the two raters" before a rating is considered reliable. The coefficient alphas for the various scales ranged from a low of .84 for Writing to a high of .94 for Mathematics. Based on these coefficients, the reported standard error of measurement for Reading is 2.2, for Writing is 2.6, and for Mathematics is 2.6. The inter-rater agreements ranged from a low of 10% to a high of 100% with the mean inter-rater agreement approaching 85%. For the students in this study, the mean inter-rater agreement across each scale was 92.5%.

### *Procedures*

The second author coordinated participant recruitment and data collection during the spring 2003 and 2004 statewide assessment window in May. The academic competence and adaptive behavior ratings were collected during the week before or just after the completion of the *IAA*. The *ACES* and *Vineland* scales were scored by the researchers. Because the *IAA*

is completed on-line and simultaneously scored, teachers were not aware of the proficiency level their students' achieved on the *IAA* prior to the completion of the *ACES* or *Vineland* scales.

### Data Analyses and Interpretation Guidelines

Given the nature of the research predictions, the data analyses primarily involved correlation and regression analyses to provide quantitative indices of the relationships among various academic skills and related behaviors. We also conducted some descriptive and chi square analyses to facilitate the comparison of the average *IAA* ratings of the *eligible* and *not eligible* groups of students. The evidence for the concurrent validity question was correlations between the *IAA* and the established norm-referenced assessments (i.e., *ACES* and *Vineland Adaptive Behavior Scales*). In general, we attended to correlations between the *IAA* scales and the composite or total scale scores for the *Vineland* and *ACES*, respectively, because the scale scores are more reliable than the scores for the various subscales. We also looked for the correlations between the same scales to be similar across

grade clusters, thus providing a form of within-study replication. As a guide to interpreting the magnitude of these correlations, we characterized correlations below .30 as weak, correlations between .30 and .60 as moderate, and correlations greater than .60 as strong. In addition to the magnitude of the correlations, we provide information about the statistical significance of the correlations to facilitate comparisons and clarify the probability that such correlations could occur by chance.

The examination of the ability of the *IAA* to discriminate between the eligible and not eligible group of students was tested with a logistic regression analysis. We expected this analysis with *IAA* total scores would yield accurate identification of participants in over 90% of the cases given the content being assessed and the known competencies of the students. We also predicted that this classification accuracy would not vary when demographic information was added to the regression formula.

### Results

The means and standard deviations for the three measures of student functioning are provided in Table 1. First,

it should be noted that all the domain and composite scores on the *Vineland* for students at each grade cluster levels were determined to be at the "Low Adaptive Behavior Level" as defined in the *Vineland Technical Manual* (p. 81). This level characterizes the lowest 2% of the students in the standardization sample of the *Vineland*. In a similar vein, the means for the Academic Skills Scale and Subscales were all in the lowest decile of the *ACES* standardization sample. Some of the Academic Enabler scores (i.e., Interpersonal Skills and Motivation), particularly for middle and high school students, however, were in the 2nd or 3rd deciles. Collectively, this norm-referenced data from the *Vineland* and *ACES* indicated that the sample of students who were *eligible* for the *IAA* were students with generally the lowest levels of adaptive behavior and academic skills when compared to national representative standardization samples of peers in their age ranges.

A second observation from the rating scores in Table 1 is that the mean performances of students on the *IAA* increased incrementally across the three grade clusters. This was an expected outcome given that opportunity to learn and perform the core skills measured

**Table 1. Means and Standard Deviations for All Measures for Students Eligible to Participate in the *IAA***

	Elementary Students (N = 50) Mean (SD)	Middle School Students (N = 27) Mean (SD)	High School Students (N = 32) Mean (SD)
Academic Competence Evaluation Scales			
Academic Skills Total	35.78 (8.91)	32.70 (7.57)	35.54 (7.86)
Reading/Language Arts	12.84 (3.31)	12.48 (2.61)	35.54 (7.86)
Mathematics	9.29 (3.08)	8.59 (1.39)	9.04 (2.32)
Critical Thinking	13.46 (3.77)	11.63 (4.38)	13.71 (4.53)
Academic Enablers Total	84.66 (23.83)	95.67 (35.33)	112.63 (31.35)
Interpersonal Skills	30.80 (7.61)	33.00 (8.83)	35.97 (8.41)
Engagement Skills	14.59 (5.97)	16.59 (7.94)	21.21 (7.04)
Motivation	20.83 (6.81)	23.15 (11.28)	27.23 (9.86)
Study Skills	19.69 (7.90)	23.71 (11.28)	27.23 (9.86)
Vineland Adaptive Behavior Scale—Classroom Edition			
Communication Domain (Receptive + Expressive + Written)	49.89 (16.70)	41.28 (16.43)	53.32 (18.28)
Daily Living Skills Domain (Personal + Domestic + Community)	51.25 (14.85)	51.44 (21.84)	64.54 (17.47)
Socialization Domain (Interpersonal + Play + Coping)	61.04 (15.88)	59.54 (23.36)	76.13 (19.83)
Adaptive Behavior Composite	52.97 (14.06)	49.96 (17.35)	65.27 (14.77)
Idaho Alternate Assessment (Raw Achievement Rating Total)			
Reading (12 items; raw score range 0–36)	14.37 (6.95)	16.72 (9.31)	18.75 (7.69)
Language Arts (6 items; raw score range 0–18)	5.00 (3.30)	6.90 (4.37)	7.00 (3.97)
Mathematics (18 items; raw score range 0–54)	12.41 (9.12)	18.10 (12.98)	20.94 (9.50)

**Table 2. Correlations of IAA Subscale Scores with Student Demographic Variables**

	IAA Reading	IAA Language Arts	IAA Mathematics
Students Sex	.22*	.21*	.23*
Students Race	-.02	-.02	.05
Students Grade	.20	.17	.24*

\* $p < .05$ .

by the *IAA* increase with years of schooling.

*Evidence to Support the Concurrent Validity of the IAA Scores*

The initial evidence of interest concerned the fundamental relationships among the *IAA* scales and student demographic variables of sex, race, and grade (Table 2). The low correlations between these demographic variables and scores on the *IAA* subscales, especially for the race variable, suggested that the alternate assessment is primarily measuring constructs other than those represented by the demographic variables. This, in turn, suggests little or no bias due to demographic variables.

The featured relationships between the constructs measured by the *IAA* and the established measures of academic competence and adaptive behavior are

highlighted in Tables 3–5 for reading, language arts, and mathematics, respectively. The data in these tables have been broken down into three grade clusters—elementary, middle school, and high school—and the boldface correlations are of primary interest because they represent the strength of the relationship between the most robust measures of the constructs that each instrument assesses.

Although there is some variability in the correlations across the content areas, the *IAA*-to-*VABS* Adaptive Behavior composite and the *IAA*-to-*ACES* Academic Enablers total correlations were consistently larger (i.e., mean  $r$  values between adaptive behavior and academic enabler totals are nearly twice as large as the mean  $r$  for academic skill totals) than the correlations for the *IAA*-to-*ACES* Academic Skills total. A closer look at the cor-

relations between the various *ACES* Academic Skills subscales and the *IAA* content areas of Reading, Language Arts, and Mathematics, however, indicates some rather inconsistent patterns. These data suggest that the *IAA* shares some significant variance with measures of adaptive behavior and academic enablers and to a noticeably lesser degree with a measure of academic skills.

In Table 6, we present the combined correlations among the *IAA*, *ACES*, and *VABS* for students in grades 3–10, the group that would be relevant to NCLB accountability analyses. The pattern of relationships observed in Tables 3–5 is sharpened through the integration of student groups from all grade levels. Thus, Table 6 highlights the fact that the *IAA* Reading, Language Arts, and Mathematics scales all share significantly more variance—over twice as much—with the measures of adaptive behavior and academic enablers than with measures of academic skills. Transformations to  $z$  scores of the correlations between the *IAA* and each of the total scale scores on the *ACES* and *VABS* within Reading and Mathematics also indicated that the correlations between the *IAA*-Academic Enablers and the *IAA*-Adaptive Behavior pairs were always statistically larger (with  $z$  scores

**Table 3. Correlation Matrix for the IAA-Reading, Vineland Adaptive Behavior Scale, and Academic Competence Evaluation Scales for IAA Eligible Students**

	IAA-Reading		
	Elementary Students ( $N = 50$ )	Middle School Students ( $N = 27$ )	High School Students ( $N = 32$ )
Academic Competence Evaluation Scales			
Academic Skills Total	<b>.35*</b>	<b>.40*</b>	<b>.24</b>
Reading/Language Arts	.27	.29	.16
Mathematics	.22	.15	.27
Critical Thinking	.41**	.47*	.15
Academic Enablers Total	<b>.37*</b>	<b>.64**</b>	<b>.44*</b>
Interpersonal Skills	.36*	.52**	.55**
Engagement Skills	.42**	.51**	.48*
Motivation	.22	.64**	.36
Study Skills	.42*	.66**	.34
Vineland Adaptive Behavior Scale—Classroom Edition			
Communication Domain (Receptive + Expressive + Written)	.37**	.61**	.42*
Daily Living Skills Domain (Personal + Domestic + Community)	.33*	.63**	.29
Socialization Domain (Interpersonal + Play + Coping)	.31*	.73**	.47**
Adaptive Behavior Composite	<b>.36*</b>	<b>.70**</b>	<b>.25</b>

\* $p < .05$ .

\*\* $p < .01$ .

**Table 4. Correlation Matrix for the IAA-Language Arts, Vineland Adaptive Behavior Scale, and Academic Competence Evaluation Scales for IAA Eligible Students**

	IAA-Language Arts		
	Elementary Students (N = 18)	Middle School Students (N = 25)	High School Students (N = 19)
Academic Competence Evaluation Scales			
Academic Skills Total	<b>-.08</b>	<b>.32</b>	<b>.19</b>
Reading/Language Arts	-.05	.30	.05
Mathematics	-.13	.19	.18
Critical Thinking	.05	.31	.15
Academic Enablers Total	<b>.20</b>	<b>.42*</b>	<b>.41</b>
Interpersonal Skills	.14	.36	.38*
Engagement Skills	.42*	.43*	.45*
Motivation	-.01	.36	.13
Study Skills	.19	.40	.38
Vineland Adaptive Behavior Scale—Classroom Edition			
Communication Domain (Receptive + Expressive + Written)	.30	.54**	.29
Daily Living Skills Domain (Personal + Domestic + Community)	.30	.41*	.26
Socialization Domain (Interpersonal + Play + Coping)	.16	.51**	.26
Adaptive Behavior Composite	<b>.27</b>	<b>.60**</b>	<b>.27</b>

\* $p < .05$ .

\*\* $p < .01$ .

**Table 5. Correlation Matrix for the IAA-Mathematics, Vineland Adaptive Behavior Scale, and Academic Competence Evaluation Scales for IAA Eligible Students**

	IAA-Mathematics		
	Elementary Students (N = 18)	Middle School Students (N = 25)	High School Students (N = 19)
Academic Competence Evaluation Scales			
Academic Skills Total	<b>.20</b>	<b>.39*</b>	<b>.28</b>
Reading/Language Arts	.26	.26	.29
Mathematics	.07	.19	.35
Critical Thinking	.11	.45*	.06
Academic Enablers Total	<b>.35</b>	<b>.79**</b>	<b>.22</b>
Interpersonal Skills	.16	.66**	.26
Engagement Skills	.53**	.64**	.27
Motivation	.22	.81**	.07
Study Skills	.44*	.72**	.25
Vineland Adaptive Behavior Scale—Classroom Edition			
Communication Domain (Receptive + Expressive + Written)	.60**	.75**	.68**
Daily Living Skills Domain (Personal + Domestic + Community)	.58**	.77**	.59**
Socialization Domain (Interpersonal + Play + Coping)	.55**	.79**	.48**
Adaptive Behavior Composite	<b>.60**</b>	<b>.78**</b>	<b>.50*</b>

\* $p < .05$ .

\*\* $p < .01$ .

**Table 6. Correlation Matrix for IAA Subscale Scores with ACES and Vineland Scores for IAA Eligible Students in Grades 3–10 for NCLB Reports**

	IAA Reading (N = 92)	IAA Language Arts (N = 91)	IAA Mathematics (N = 90)
ACES			
Academic Skills Total	.30**	.19	.29**
Reading/Language Arts	.22*	.16	.30**
Mathematics	.21*	.17	.28**
Critical Thinking	.28*	.14	.18
Academic Enablers Total	.55**	.40**	.60**
Interpersonal Skills	.51**	.35**	.45**
Engagement Skills	.51**	.47**	.57**
Motivation	.52**	.29**	.52**
Study Skills	.51**	.37**	.58**
Vineland Adaptive Behavior Composite	.59**	.45**	.75**
Communication Domain	.59**	.53**	.78**
Daily Living Skills Domain	.57**	.45**	.76**
Socialization Domain	.61**	.44**	.70**
Motor Skills Domain	.55**	.47**	.65**

\* $p < .05$ .

\*\* $p < .01$ .

greater than 1.96, which is required for a  $p < .05$ ) than the IAA-Academic Skills correlations.

*Evidence to Support the Discriminant and Consequential Validity of the IAA Scores*

To examine the ability of the IAA to differentiate between groups of students who were *eligible* for an alternate assessment and those *not eligible*

for it, we established key characteristics for the two groups on a common measure with national norms. Table 7 provides the means and standard deviations of ACES ratings by teachers of students with disabilities who did not qualify to take the IAA, but instead took the statewide achievement test (i.e., *ISAT*). The means for ACES Academic Skills total and Academic Enablers total are much higher than those of the *eligible* group (Table 8) and place

these students in the 2nd decile and 4th decile, respectively. The *not eligible* group's mean score on the IAA scales also was much higher than that of the *eligible* counterparts. These differences in the two groups were also evident when their IAA scores (achievement and progress combined) were transformed to proficiency levels used for NCLB reporting. Table 9 provides information about the groups' actual IAA scores, the percentage of students in

**Table 7. Means and Standard Deviations for All Measures for Students Not Officially Eligible for the IAA**

	Elementary Students (N = 20) Mean (SD)	Middle School Students (N = 10) Mean (SD)	High School Students (N = 14) Mean (SD)
Academic Competence Evaluation Scales			
Academic Skills Total	61.50 (20.23)	68.00 (21.45)	57.43 (22.17)
Reading/Language Arts	21.86 (6.26)	23.00 (6.67)	20.80 (8.44)
Mathematics	16.08 (5.63)	15.27 (6.21)	14.47 (6.80)
Critical Thinking	24.52 (8.95)	29.09 (8.95)	24.14 (10.78)
Academic Enablers Total	124.43 (29.50)	137.09 (30.00)	128.00 (32.92)
Interpersonal Skills	39.79 (8.15)	38.91 (7.85)	38.71 (8.10)
Engagement Skills	25.57 (7.55)	28.36 (5.10)	25.80 (6.58)
Motivation	27.91 (9.30)	33.10 (9.33)	26.67 (10.38)
Study Skills	33.39 (9.67)	36.73 (10.29)	31.92 (11.31)
Idaho Alternate Assessment (Raw Achievement Ratings Total)			
Reading	27.39 (7.13)	29.91 (7.18)	26.36 (5.89)
Language Arts	13.68 (2.87)	14.64 (3.29)	12.93 (3.59)
Mathematics	38.12 (10.89)	42.91 (10.44)	39.87 (8.99)

**Table 8. Comparisons of IAA Achievement Ratings for Students Eligible and Not Eligible for an Alternate Assessment**

Content Area	Student Status	Elementary Students M (SD)	Middle School Students M (SD)	High School Students M (SD)
Reading	Eligible for AA	14.37 (6.95)	16.72 (9.31)	18.75 (7.69)
	Not Eligible for AA	27.39 (7.13)	29.91 (7.18)	26.36 (5.89)
Language Arts	Eligible for AA	5.00 (3.30)	6.90 (4.37)	7.00 (3.97)
	Not Eligible for AA	13.68 (2.87)	14.64 (3.29)	12.93 (3.59)
Mathematics	Eligible for AA	12.41 (9.12)	18.10 (12.98)	20.94 (9.50)
	Not Eligible for AA	38.12 (10.89)	42.91 (10.44)	39.87 (8.99)

Note: The Reading scale is composed of 12 items with a possible score range from 0 to 36 on the dimension of Achievement only; the Language Arts scale is composed of 6 items with a possible score range from 0 to 18; and the Mathematics scale is composed of 18 items with a possible score range from 0 to 54.

**Table 9. Comparisons of the Number and Percentage of Proficient Performances for Students Eligible and Not Eligible to Take the IAA**

Content Area	AA Eligibility Status (N)	IAA		IAA Proficiency % Proficient (Proficient + Advanced Levels)	IAA Estimated Cumulative Percentile Rank
		Mean	SD		
Reading	Eligible (91)	93.20	39.61	72.5%	64%tile
	Not Eligible (39)	157.62	31.04	97.4%	99%tile
Language Arts	Eligible (91)	43.47	21.60	62.7%	61%tile
	Not Eligible (40)	67.62	16.60	100%	84%tile
Math	Eligible (90)	105.91	56.45	68.9%	54%tile
	Not Eligible (38)	224.76	48.46	100%	98%tile

Note: All students in this sample were students with an IEP. Those students defined as Eligible meet the Participation Criteria for the IAA and were administered the IAA during the standard testing period. Those students defined as Not Eligible took the regular statewide assessment and then their teachers completed the IAA to assess their knowledge and skills.

each group who would meet the state's cut-score criterion for Proficiency, and the groups' relative standing (cumulative percentile rank) among all those who were administered the IAA during the 2003 and 2004 school years. As documented in this table, 100% of the *not eligible* students were functioning at the Proficient level for the alternate assessment in Language Arts and Mathematics, and 97.4% were Proficient in Reading. In comparison, 72.5%, 62.7%, and 68.9% of the *eligible* group were achieving at the Proficient level respectively in Reading, Language Arts, and Mathematics on the alternate assessment.

A series of one-way ANOVAs were conducted to test the significance of the differences in the IAA achievement scores for these two groups of students. The F-ratios based on a total sample of 130 students for Reading (81.53), Language Arts (39.68), and

Mathematics (128.34) were all significant at the .0001 level of probability. In addition, a series of chi-square analyses were conducted to examine the frequency distribution for the two groups of students with regard to proficiency criterion. Specifically, the chi-square results were: Reading  $\chi^2(1, N = 130) = 10.59, p = .001$ , Language Arts  $\chi^2(1, N = 131) = 20.18, p = .0001$ , and Mathematics  $\chi^2(1, N = 128) = 15.13, p = .0001$ . These results support the prediction that the number of student *not eligible* for the IAA who met the criterion for Proficient performance would be significantly greater than for the *eligible* group of students. In summary and as expected, the *ACES* and *IAA* rating data clearly established that the students in the *not eligible* group were a statistically significant higher performing group of students than their peers who were *eligible* for the IAA.

A logistic regression analysis for each grade level cluster (i.e., elementary, middle, high school) of students was conducted to determine how well the various IAA scales differentiate membership in the *eligible* and *not eligible* groups. (Although the purpose of the IAA is not to predict or assign membership to a group, the content is designed to be most appropriate for students functioning substantially below grade level. Thus, it follows that higher functioning students as determined by IEP teams review who are not eligible for the IAA should out-perform students who are eligible.) The results of each of the three analyses indicated that the IAA Mathematics scale was the best predictor and that the Reading and Language Arts scales contributed minimally to the discriminant validity for the IAA. Specifically, the resulting regression model accurately classified 93.4% of the students. When student race and

**Table 10. Correlation Matrix for the IAA-Reading & ACES for Students Not Officially Eligible for the IAA**

	IAA-Reading		
	Elementary Students (N = 20)	Middle School Students (N = 10)	High School Students (N = 14)
ACES			
Academic Skills Total	<b>.59**</b>	<b>.62</b>	<b>.30</b>
Reading/Language Arts	.56**	.65*	.40
Mathematics	.35	.32	.38
Critical Thinking	.55*	.62*	.31
Academic Enablers Total	<b>.58**</b>	<b>.53</b>	<b>.56</b>
Interpersonal Skills	.38	.44	.03
Engagement Skills	.54**	.76**	.01
Motivation	.64**	.40	.36
Study Skills	.57**	.49	.37

\* $p < .05$ .  
\*\* $p < .01$ .

gender was added to a subsequent logistic regression with the *IAA* scores, they did not improve the classification accuracy beyond the 93.4% level.

An examination of the relationship between ratings on the *IAA* and the *ACES* for the *not eligible* group of students provided additional relevant evidence about the validity of the *IAA*. Specifically, Tables 10–12 provide the correlations for reading, language arts, and mathematics. A review of the data across these tables reveals that all the correlations between the *IAA* scales and the *ACES* Academic Skills scales are moderate to very high ( $r$  range .30 to .84). By comparison to the same *IAA-ACES* Academic Skills relationships for the *eligible* group of students, the correlations for the *not eligible* group are more than twice as robust. The pattern of correlations between the *IAA* scales and the *ACES* Academic Enablers scales for the *not eligible* group ranged ( $r$  range .15 to .80) from low to high in strength. These correlations were very similar to those found (in Tables 3–5) for the *eligible* group of students.

A final set of correlations provides some insights into the degree of shared variance between the *IAA* and *ISAT* scores for *not eligible* students in the content areas of reading, language arts, and mathematics. Table 13 displays these correlations, as well as those between the *IAA* and *ACES*. As can be seen, the *IAA* Reading score correlated highest with the *ISAT* reading score

( $r = .46$ ), the *IAA* Language Arts score correlated highest with the *ISAT* Reading scale ( $r = .59$ ) but also moderately with Language Arts scale ( $r = .49$ ), and finally the *IAA* Mathematics score correlated highest with the *ISAT* Reading scale ( $r = .67$ ) but moderately high with the Mathematics score ( $r = .56$ ) as well. The correlations between content-similar scales on the *IAA* and *ISAT* were all moderately high in strength, but slightly lower than those between the *IAA* and the *ACES*. It should be noted,

however, that these latter two measures share a common rater source, whereas the *IAA* and *ISAT* were generated by different sources (i.e., the teacher rating versus an individual student assessment).

The collection of evidence for the validity of the *IAA* scores from this series of analyses is synthesized and discussed in the next section within the context of validity expectations and standards.

## Discussion

This investigation focused on providing more information about the constructs measured by the *IAA*, a USDOE approved alternate assessment, and the ability of its resulting scores to differentiate between groups of students known to be eligible or not eligible to participate in the statewide alternate assessment. The investigation was motivated by best practices in test development as described in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999) and assessment requirements of the No Child Left Behind Act of 2001 as articulated in the USDE *Standards and Assessments Peer Review Guidance* (April, 2004) document. Consistent with the conceptualization of validity evidence in these documents,

**Table 11. Correlation Matrix for the IAA-Language Arts & ACES for Students Not Officially Eligible for the IAA**

	IAA-Language Arts		
	Elementary Students (N = 20)	Middle School Students (N = 10)	High School Students (N = 14)
ACES			
Academic Skills Total	<b>.59**</b>	<b>.68**</b>	<b>.50</b>
Reading/Language Arts	.59**	.69	.36
Mathematics	.46*	.49	.46
Critical Thinking	.57*	.57	.52
Academic Enablers Total	<b>.28</b>	<b>.62*</b>	<b>.75**</b>
Interpersonal Skills	.26	.60	.29
Engagement Skills	.24	.59	.39
Motivation	.32	.52	.62**
Study Skills	.14	.59	.60**

\* $p < .05$ .  
\*\* $p < .01$ .

**Table 12. Correlation Matrix for the IAA-Mathematics & ACES for Students Not Officially Eligible for the IAA**

	IAA-Mathematics		
	Elementary Students (N = 20)	Middle School Students (N = 10)	High School Students (N = 14)
ACES			
Academic Skills Total	.50*	.80**	.84**
Reading/Language Arts	.48	.80**	.70**
Mathematics	.50*	.55	.78**
Critical Thinking	.49	.78**	.80**
Academic Enablers Total	.15	.76**	.81**
Interpersonal Skills	.31	.77**	.40
Engagement Skills	.09	.80**	.47
Motivation	.17	.63*	.76**
Study Skills	.04	.68*	.73**

\* $p < .05$ .

\*\* $p < .01$ .

we utilized a multitrait (i.e., academic skills, academic enablers, adaptive behavior, and demographic characteristics), multimethod (i.e., teacher completed rating scales, demographic survey, and for a subset of students an individualized achievement test) correlational design to provide information about the relative strength of relationships between the *IAA* Reading, Language Arts, and Mathematics scales and more established measures of similar (i.e., *ACES*-Academic Skills scale, *ISAT*) and different (i.e., *ACES*-Academic Enabler scale, *VABS*) constructs. In addition to this design approach, we used two samples of students that differed with regard to their perceived severity of disability and ultimately their eligibility to participate in the statewide alternate assessment. This use of known groups is a novel design element in the psychometric research on alternate assessments and provided useful evidence relevant for examining the concurrent, discriminant, and consequential validity of the *IAA*.

Before discussing the major findings, it is important to note that the results of this study are based on the assessment of a representative sample of students from across the state of Idaho. In most cases, a student's primary teacher completed the rating scales with required knowledge about the student's background. Assessment instruments like the *IAA*, *ACES*, and *VABS* all assume that teachers can objectively summarize several weeks of direct interactions

with and observations of a student. Of course, there is a potential for bias in such ratings, but given that all the *IAA* scale total scores met the state's criteria for reliability (i.e., inter-rater agreement) and the *ACES* and *VABS* have been rigorously studied and found to consistently yield reliable scores, the results from this study are considered to be reliable and useful.

#### Major Findings

Evidence to support the validity of the *IAA* was mixed, yet on balance, promising. Specifically, we found the *IAA* correlated minimally with key stu-

dent demographic variables. This was an expected and desired outcome for a rating scale assessment designed to measure academic content.

The relationship between the reading, language arts, and mathematics achievement level ratings on the *IAA* and the concurrent scores on the *ACES*-Academic Skills scales for the *eligible* students varied across grade clusters, but in general were moderate at best. When the correlations for the same score relationships were examined for the *not eligible* students, the magnitude of the correlations increased noticeably. This finding was expected given the score range on both instruments

**Table 13. Correlation Matrix for the IAA, ACES, and ISAT for Students with Disabilities Who Did Not Qualify to Participate in the IAA**

	IAA		
	Reading Proficiency Score	Language Arts Proficiency Score	Mathematics Proficiency Score
ACES			
Academic Skills Total	.59**	.48*	.71**
Academic Enablers Total	.68**	.60**	.74**
ISAT			
Reading	.46*	.59**	.67**
Language Arts	.30	.49*	.48*
Mathematics	.21	.55*	.56*

\* $p < .05$ .

\*\* $p < .01$ .

was less restricted for the *not eligible* group of students. Collectively, these findings provide evidence that the *IAA* scales measure skills indicative of the academic content characterized in the state's content standards. This point was further reinforced by the moderate to high correlations between the *IAA* and *ISAT* for the *not eligible* students. Thus, the evidence for both groups of students provides support for the construct validity of the *IAA* scores.

Although the scores between academic skills on the *IAA* and other measures indicated a meaningful amount of shared variance (i.e., 20 to 40%), there are cases, particularly at the elementary grade levels, where there was more shared variance with the academic enabling and adaptive behavior constructs. Neither of these constructs is featured in the state's academic content standards, yet both are clearly part of successful achievement for students. If replicated, however, these findings could cause concern because they suggest that the *IAA* may be a broader measure of student functioning than called for by the state's content standards. In other words, the *IAA* appears to measure an array of academic skills in the basic content areas of reading, writing, and mathematics along with a number of skills and behaviors typically characterized as functional and part of the process of learning.

Important evidence concerning the valid use of the *IAA* was provided by the findings that the resulting scores, especially for the Mathematics scale, do an excellent job of differentiating students who were *eligible* from those *not eligible* to participate in an alternate assessment. As detailed in the results, the *IAA* accurately classified a very high percentage of students at each grade cluster into their known group. Although the *IAA* is not intended as a classification tool, these data suggested that the *IAA* is sensitive to ability differences portrayed by students with disabilities. In addition, the resulting scores from the *IAA* for the two groups were almost non-overlapping with students in the *eligible* group functioning at the 59th percentile (on average) and the students in the *not eligible* group functioning at the 94th percentile (on average) when compared to all students who took the *IAA* in the previous two years. Thus, the evidence supported our expectations about the measurement content and sensitivity of the *IAA*.

### *Limitations and Implications*

This validity study was undertaken with care and the willing participation of many special educators with experience in assessing the classroom performance of students with significant disabilities. The findings are based upon a large and representative sample of students from those that actually participated in the statewide assessment, yet one must be cautious in making strong conclusions without further replications and related studies that can serve to cross-validate the observed trends.

As noted earlier, we primarily used a rating scale approach that depended on teachers' knowledge of students' skills and behaviors to provide assessment results. Well-constructed rating scales have been shown to correlate highly with actual student performances and work samples (DiPerna & Elliott, 2000), but the MTMM approach to examining relationships between newly developed and more established measures calls for multiple methods of assessment to minimize inflated correlations due to shared variance in method or source of information. This use of a single method of assessment is not a unique limitation to this study or to the state's alternate assessment. In fact, virtually all states are forced to utilize an assessment approach that features teachers' or other educators' judgments when it comes to assessing students with significant cognitive disabilities because these students cannot be administered typical achievement tests. By taking steps to ensure high inter-rater reliability and the scoring accuracy of the assessments, we have minimized bias and threats to internal validity. The addition of the *ISAT*, an individualized achievement measure, as a concurrent measure for a group of students strengthened the information and provided a means for making stronger inferences about relationships between measures based on teacher judgments.

Another limitation of the study is that the results generalize only to students with significant disabilities who use alternate assessment rating scales comparable to the *IAA*. While the results may be limited in utility to others conducting alternate assessments, the validation design features of the study are not. In fact, given the validity evidence for alternate assessments outlined in the U. S. Department of Education's

*Standards and Assessments* document, we believe the use of both an MTMM and known groups method facilitates understanding of a number of aspects critical to the evolving efforts to establish the validity of alternate assessment results.

### **Summary and Conclusions**

This investigation of the Idaho Alternate Assessment has advanced understanding of what the *IAA* measures and its ability to differentiate levels of functioning for students with significantly different degrees of impairment. There is sound evidence that the *IAA* measures the reading, language arts, and mathematics skills of students and also measures a class of behaviors (i.e., social skills, engagement/motivation, and study skills) that enable academic function and daily living. This measurement of a broader array of skills than called for by the state's academic content standards is not surprising given the fact that the classroom-based evidence used to substantiate achievement ratings often comes from a curriculum that emphasizes functional living skills, fundamental behaviors that foster engagement, and basic interpersonal communication. Although students with significant disabilities have a right to access the general reading, language arts, and mathematics curriculum, they also still have the need for instruction that focuses on basic living and communication skills. Thus, it may be appropriate that the *IAA* is capturing a broader array of skills than expected by the state's academic standards. From a construct validity perspective, there is a meaningful amount of construct-irrelevant variance in the *IAA* scores, yet the scores are functioning rather well in differentiating performances by known groups of students. It is hypothesized that this irrelevant variance will be reduced over time as (a) students' IEPs become directed more at core academic skills and (b) teachers continue to receive professional development that improves their collection of standards-based evidence for *IAA* ratings.

The evidence we have established for the validity of the *IAA* scores suggests it is a promising assessment of basic academic skills of students with significant disabilities. Our evidence about the validity of scores from this rating scale approach to alternate assessment

is not meant to argue against other approaches to alternate assessment. Research on the IAA's use and technical properties has the potential to advance inclusive assessment practices in Idaho, as well in many other states where a rating scale approach to alternate assessment is implemented.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Browder, D. M., Fallin, K., Davis, S., & Karvonen, M. (2003). Consideration of what may influence student outcomes on alternate assessment. *Education and Training in Developmental Disabilities, 38*, 255–270.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.
- DiPerna, J. C., & Elliott, S. N. (2000). *Academic competence evaluation scales*. San Antonio, TX: Psychological Corporation.
- Elliott, S. N., & Roach, R. T. (in press). Alternate assessments of students with significant disabilities: Alternative approaches, common technical challenges. *Applied Measurement in Education*.
- Idaho Department of Education (1999). *Idaho Alternate Assessment*. Boise, ID: Author.
- Johnson, E., & Arnold, N. (2004). Validating an alternate assessment. *Remedial and Special Education, 25*(5), 266–275.
- Kampfer, S. H., Horvath, L. S., Kleinert, H. L., & Kearns, J. F. (2001). Teachers' perceptions of one state's alternate assessment: Implications for practice and preparation. *Exceptional Children, 67*, 361–374.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.
- Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont portfolio assessment program* (CSE Technical Report 355). Los Angeles: NCREST.
- Meyer, L., Eichinger, J., & Downing, J. (1992). *The program quality indicators (PQI): A checklist of most promising practices in educational programs for students with severe disabilities*. Syracuse, NY: Syracuse University Division of Special Education and Rehabilitation.
- Roach, A. T. (November 2003). *Alignment of Idaho Academic Standards with the Idaho Alternate Assessment (IAA)*. Technical Report Completed for the Idaho Department of Education, Boise, ID.
- Roach, A. T., Elliott, S. N., & Webb, N. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin Alternate Assessment. *Journal of Special Education, 38*(4), 218–231.
- Shinn, M. R. (1995). Curriculum-based measurement and its use in a problem-solving model. In A. Thomas and J. Grimes (Eds.), *Best practices in school psychology III* (pp. 547–567). Washington, DC: National Association of School Psychologists.
- Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1985). *Vineland adaptive behavior scales: Classroom edition*. Circle Pines, MN: American Guidance Service.
- Tindal, G., McDonald, M., Tedesco, M., Glasgow, A., Almond, P., Crawford, L., & Hollenbeck, K. (2003). Alternate assessments in reading and math: Development and validation for students with significant disabilities. *Exceptional Children, 69*, 481–494.
- Turner, M. D., Baldwin, L., Kleinert, H. L., & Kearns, J. F. (2000). The relation of a statewide alternate assessment for students with severe disabilities to other measures of instructional effectiveness. *Journal of Special Education, 34*, 69–76.
- U.S. Department of Education (April, 2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC.: Author.
- Witt, J. C., Elliott, S. N., Daly, E. J., III, Gresham, F. M., & Kramer, J. J. (1998). *Assessment of at-risk and special needs children* (2nd ed.). Boston: McGraw-Hill.