Running Head: WHAT DOES AN AA-AAS MEASURE?

What Does an Alternate assessment of alternate achievement standards Measure?

A Multitrait-Multimethod Analysis across Six States

Ryan J. Kettler, Stephen N. Elliott, and Peter A. Beddow

Peabody College of Vanderbilt University

Elizabeth Compton

Idaho Department of Education

Dawn McGrath

Indiana Department of Education

Kris Kaase

Mississippi Department of Education

Charles Bruen

Arizona Department of Education

Lisa Ford

Nevada Department of Education

Kent Hinton

Hawaii Department of Education

Please direct all correspondence concerning this manuscript to Dr. Ryan J. Kettler, 410E Wyatt Center, Peabody #59, 230 Appleton Place, Nashville, TN 37203-5721, email: r.j.kettler@vanderbilt.edu, Telephone: (615)343-5702, fax: (615)322-4488.

Abstract

There exists a paucity of research on the construct validity of states' alternate assessments of alternate achievement standards (AA-AASs), used to measure and document the proficiency levels of students with significant cognitive disabilities who cannot be assessed using states' general assessments. This multitrait-multimethod (MTMM) study incorporated students with disabilities who would be eligible (SWD-Es) for such an assessment, along with students with disabilities who would not be eligible (SWD-NEs) as a comparison group, at both the elementary school and the middle school grade bands. The study replicates findings across six states, using general assessments and established measures of academic skills, academic enablers, and adaptive behavior as criterion measures. Findings included that reading and mathematics scores shared a high degree of variance, AA-AAS scores and general assessments did not share a high degree of variance, and AA-AAS scores overlapped with both academic skills and adaptive behavior. These findings with regard to AA-AAS scores and overlapping constructs, along with their implications for special educators, are discussed.

What Does an Alternate assessment of alternate achievement standards Measure?

A Multitrait-Multimethod Analysis across Six States

For more than a decade, alternate assessments of alternate achievement standards (AA-AASs) have been administered to students with significant disabilities who cannot meaningfully participate in their states ' general achievement tests. As a result of federal legislation starting with the Individuals With Disabilities Education Act of 1997 (IDEA), and reiterated in the No Child Left Behind Act of 2001 (NCLB) and the Individuals With Disabilities Education Improvement Act of 2004 (IDEIA), every state has an AA-AAS and is required to ensure its technical soundness.

The technical soundness of such alternate assessments, however, remains an area of concern. Basic questions about the constructs measured and their relationship to other measures of achievement remain largely unsubstantiated with rigorous research and validation studies. This assertion is supported by the paucity of published studies or document evidence for construct validity in states' AA-AAS technical manuals. The forthcoming National Study of Alternate Assessments (NSAA) report (SRI International, 2009) provides a comprehensive descriptive summary of key attributes of AA-AASs and resulting accountability data for each state in the country. The NSAA indicates that directors of AA-AASs in 41% of the states and one territory reported conducting a formal study to document that test and item scores are related to internal or external variables as intended. They also reported in 59% of the states a formal study had been conducted to document measurement of the construct relevance of their test. These formal studies apparently are reported in the various states ' technical manuals for their AA-AASs. The information, however, is not widely available.

This lack of availability is further attested to by Towles-Reeves, Kleinert, and Muhomba (2009). In a recent review of research on AA-AASs, these authors identified 23 empirical studies completed since 2003. Specifically, Towles-Reeves et al. lament "there is considerably less research that has examined the extent to which actual student scores were associated with empirically verified instructional or other outcome variables" (2009, P. 245). These authors called for "future research to investigate the relationship between AA-AAS (regardless of approach: portfolios, performance assessments, and checklists) and another accepted measure of student learning" (p. 246). They concluded, "there is no evidence to support the correlation of alternate assessments with other accepted measures of student learning" (p. 246). This is a serious claim and should cause all users of these assessments to be cautious when interpreting AA-AAS scores.

Previous Research on AA-AASs

Some published evidence for the validity of the constructs measured by AA-AASs does exit but was not reviewed by Towles-Reeves et al. (2009). For example, a validation study of the Idaho Alternate Assessment (IAA; Idaho Department of Education, 1999) scores focused on evidence about the underlying construct being measured (Elliott, Compton, & Roach, 2007). This study was an examination of the relationships between ratings on the IAA for students with significant disabilities and corresponding ratings on two norm-referenced teacher rating scales: the *Academic Competence Evaluation Scales* (ACES; DiPerna & Elliott, 2000) and the *Vineland Adaptive Behavior Scales* (VABS; Sparrow, Balla, & Cicchetti, 1985). The study examined IAA performance for a representative group of students with disabilities (N = 116), who according to their IEP teams were eligible and actually participated in the state's alternate assessment, and for another group of students who had disabilities (N = 54), but were not officially eligible for the

alternate assessment. Both groups of students were assessed with the IAA 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 (Idaho Standards Achievement Tests, ISAT; Idaho Department of Education, 2008) with accommodations. We examine this seminal alternate assessment study in detail because it provided the basis for the design of the present study.

The evidence of interest examined by Elliott et al. (2007) concerned relationships between the constructs measured by the IAA and two other types of variables: (1) established rating scale measures of academic competence (ACES) and adaptive behavior (VABS), and the ISAT. The correlational results highlighted that the IAA Reading, Language Arts, and Mathematics scales all shared more variance with the measures of adaptive behavior and academic enablers than with measures of academic skills. In comparison to the same IAA to ACES Academic Skills relationships for the eligible group of students, the correlations for the not eligible group tended to be about twice as large. A final set of correlations provided insight 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. In comparison to the same IAA to ISAT content areas, correlations were in the medium (reading, language arts) or large (mathematics) ranges, but correlations between different content areas also tended to be in these ranges (e.g., *r* between AA-AAS mathematics and ISAT reading = .67).

Elliott et al. concluded that the evidence to support the validity of the IAA was mixed, yet on balance promising. 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 medium at best.

When the correlations for the same score relationships were examined for the not eligible students, the magnitude of the correlations increased noticeably. Collectively, these findings provided 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. The evidence for both groups of students provides support for the construct validity of the IAA scores. Although the correlations between academic skills on the IAA and other measures indicated a meaningful amount of shared variance (i.e., 20% to 40%), there were cases, particularly at the elementary grade levels, where there was more shared variance with the academic enabling and adaptive behavior constructs.

Purpose of the Present Research: A Multi-State Replication Study

The IAA validity study conducted by Elliott, Compton, and Roach (2007) served as the model for the present multi-state investigation. The participating states in this study all used a comprehensive rating scale approach to alternate assessment for students with significant disabilities. To understand each of these states' alternate assessments, a multitrait-multimethod (MTMM) design was used to determine the relationship among the AA-AAS, the state's general achievement test , and two established teacher-based rating scales.

Validity evidence based on relationships with other variables is one of five main types mentioned in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Evidence based on relationships with other variables is the degree to which scores from an instrument converge with indicators of similar constructs (convergent validity) and diverge from indicators of dissimilar constructs (divergent validity), as

well as the degree to which the scores share no relationship with indicators of unrelated constructs (discriminant validity).

Campbell and Fiske (1959) suggested an approach by which scores from multiple methods and indicative of multiple traits could be used as evidence for the validity of a new measure. The MTMM approach allows for an integrative multivariate framework within which information about convergent and discriminant validity is systematically gathered in a single study. The MTMM approach is also useful for providing evidence about the construct being measured. In the current study, multiple traits that were considered included academic performance, academic skills, academic enablers, and adaptive behavior. The multiple methods that were considered included individually administered achievement tests and teacher completed rating scales.

The method of choice for characterizing evidence based on relationships with other variables, as well as for completing an MTMM matrix, is often the Pearson correlation. The magnitude can be considered an effect size, and when the correlation is squared, the product is the amount variance shared between the two sets of scores. In the social sciences, Cohen's (1992) guidelines for classifying effect sizes are generally used, with a medium effect size (r = .30 or r = -.30) intended to be an effect that can be witnessed by the naked eye. Cohen's suggestions for small (r = .10 or r = -.10) and large (r = .50 or r = -.50) effect sizes were to be noticeably and equally smaller and larger, respectively. The values that Cohen suggested are typically considered the inner boundaries of each range, such that small positive correlations would be between .10 and .30. Hopkins (2002) extended Cohen's framework to include nonexistent (r = .00), very large (r = .70 or r = -.70), and nearly perfect correlations (r = .90 or r

= -.90). The range for nearly perfect correlations is consistent with indicators of acceptable reliability for scores from a single measure.

Research Questions

Legislation and research related to AA-AASs have inspired two research questions about the constructs measured by these assessments. Using an MTMM replication design, we examined evidence to address the following questions across multiple states:

1. *Do the AA-AAS subscales measure distinct content areas, which correlate with the same content areas on each state's general assessment, when both are used on a common sample?*
2. *Which constructs (academic skills, academic enablers, adaptive skills) are being measured by AA-AASs when used with students with disabilities who would be eligible (SWD-Es) and students with disabilities who would not be eligible (SWD-NEs) across six states?*

Method

Participants

The sample included 402 elementary school (third grade through fifth grade) students and 317 middle school (sixth grade through eighth grade) students from six states, balanced across two groups: students with disabilities who were eligible for an AA-AAS (SWD-Es, n = 361) and students with disabilities who were not eligible (SWD-NEs, n = 358). The sample included a higher number of male students (n = 444) than female students (n =274). Participants were primarily European American (n = 433), African American (n = 83), and Latino American (n = 115) students. These demographics were very consistent across grade bands and across eligibility groups. The sample was comprised of students from the following six states: Indiana (n = 284),

Arizona (n = 131), Nevada (n = 104), Idaho (n = 81), Mississippi (n = 49), and Hawaii (n = 46). Table 1 depicts these demographics disaggregated by grade band and eligibility group.

Insert Table 1 about here

Table 4 contains disability category data for the current sample disaggregated by grade band and eligibility status. Participants represented 13 different disability categories. For the eligible group, students identified with mental retardation comprised the largest proportion of the sample across grades. For the non-eligible group, students identified with a specific learning disability comprised the largest proportion of the sample. Other highly-represented categories for the eligible group were autism and multiple disabilities. There were no major differences in disability category representation across grade bands.

Insert Table 2 about here

Materials

Measures used in the current study included AA-AASs from all six states, general achievement tests from the two states (Indiana and Idaho) that had the largest sample sizes across groups, and established measures of academic skills, academic enablers, and adaptive behavior.

Alternate assessments of alternate achievement standards. The Arizona Instrument to Measure Standards - Alternate (AIMS-A; Arizona Department of Education & Elliott, 2006) is an assessment based on rating scale, performance task, and multiple-choice components. Each student is assessed on all three content standards in reading and all five content standards in mathematics. The AIMS-A is administered by the student's special education teacher. Teacher judgment is incorporated into decisions around administration and interpreting and reporting of scores. Performance evidence is based on state required standardized scales, tasks, and items. The AIMS-A was developed by a team with representatives from the state department,

researchers, parents, and other stakeholders. Internal consistency was calculated as an estimate for reliability. Some evidence for validity based on relationships with other variables, internal structure, and consequences of the assessment is reported in the AIMS-A Technical Manual.

The Hawaii State Alternate Assessment (HSAA; Hawaii Department of Education, 2007) is a rating scale with a portfolio of evidence submitted for independent scoring. Each student is assessed on all three content standards in language arts and all five content standards in mathematics. The HSAA is administered by the student's special education teacher and by a certified educator who is not the student's teacher. Teacher judgment is incorporated into selection of materials, decisions around administration, and interpreting and reporting of scores. The ratings are based on a diverse set of evidence including student work samples, collected through both daily and on-demand approaches. The HSAA was developed by a team with representatives from the state department and from an assessment company, researchers, parents, and other stakeholders. Internal consistency and inter-rater scoring consistency were calculated as estimates of reliability. Evidence for validity based on relationships with demographic variables, internal structure, and consequences of the assessment is reported in a comprehensive HSAA Technical Manual.

The Idaho Alternate Assessments (IAA; Idaho Department of Education, 2005) are evidence-based rating scales delivered online and approved by the USDE in 2003. Each student is assessed on all five content standards in reading/language arts and all seven content standards in mathematics. The IAA is administered by each student's special education teacher and by a certified educator who is not the student's teacher. Teacher judgment is incorporated into selection of materials, decisions around administration, and interpreting and reporting of scores. The ratings are based on a diverse set of evidence including student work samples, collected

through both daily and on-demand approaches. The IAA was developed by a team with representatives from the state department, researchers, parents, and other stakeholders. The reliability of the IAA rating scales was established using both coefficient alphas at the scale level and inter-rater reliability at the item level. The coefficient alphas for the scales ranged from .84 to .94. The mean inter-rater agreement approached 85%. In Elliott et al. (2007), the mean inter-rater agreement across each scale was 93%. The item content for the IAA was determined to be well aligned with Idaho's general education content standards (Roach, 2003). Evidence for validity based on relationships with other variables, internal structure, and consequences of the assessment is also provided in the IAA Technical Manual.

The Indiana Standards Tool for Alternate Reporting (ISTAR; Indiana Department of Education, 2009a) is a rating scale designed as a measure of academic and functional progress from birth to employment, and it was approved by the USDE in 2006. Each student demonstrates on all seven content standards in language arts and all seven content standards in mathematics. The items on the ISTAR are derivatives from the Indiana State Academic Standards and the extensions or Foundations to the standards in the areas of Language Arts and Mathematics. The range of items at each level depends on the content of the state standards. The ISTAR is administered by the student's special education teacher and by a certified educator who is not the student's teacher. Teacher judgment is incorporated into selection of content and materials, decisions around administration, and interpreting and reporting of scores. The ratings are based on a diverse set of evidence including student work samples and teacher recollection of student performance. The ISTAR was developed by a team with representatives from the state department, researchers, parents, and other stakeholders. The reliability of the ISTAR was established using both coefficient alphas at the scale level and inter-rater reliability. A range of

.48 to .94 was found to be the estimate of internal consistency through the Cronbach alpha statistic. Inter-rater reliability was reported through application of the Kappa statistic and percent agreement. The intraclass correlation of scores ranged from .85 to .99 with a mean of .95. With regard to validity evidence based on relationships with other variables, five statistical procedures were conducted to examine the variability of groups assessed with the ISTAR instrument, with the consistent finding of discrimination between groups of students at differing grade bands and between students with and without disabilities. Evidence for validity based on internal structure and consequences of the assessment is also provided in the ISTAR Technical Manual.

The Mississippi Alternate Assessment of Extended Curriculum Frameworks (MAAECF; Mississippi Department of Education, 2007) are evidence-based rating scales with portfolios submitted for independent scoring. Each student demonstrates on all four content standards in language arts and various (by grade) content standards in mathematics. The MAAECF is administered by each student's special education teacher and scored by both the student's teacher and a school- or district-based educator. Teacher judgment is incorporated into selection of materials, decisions around administration, and interpreting and reporting of scores. The ratings are based on a diverse set of evidence including student work samples, and are collected through daily approaches, guided by state instructions regarding type and amount of evidence. The MAAECF was developed by a team with representatives from the state department, researchers, and other stakeholders. Internal consistency and inter-rater scoring consistency were calculated as estimates of reliability. Evidence for validity based on relationships with demographic variables, alignment to content standards, internal structure, and consequences of the assessment is provided in the MAAECF Technical Manual.

The Nevada Alternate Scales of Academic Achievement (NASAA; Nevada Department of Education, 2009) are performance based assessments in which each student demonstrates on two different benchmark skills for three language arts standards and three mathematics standards. Thus, there are six total performance measures for each student for language arts and six for mathematics. The NASAA is administered by the student's teacher after instruction and is video-taped. Teacher judgment is incorporated into selection of content and materials, decisions around administration, and interpreting and reporting of scores. A second, independent team of raters scores the video-taped performances as well. The ratings are based on a diverse set of evidence including student work samples, collected through both daily and on-demand approaches, based on state-provided instructions. The NASAA was developed by a team with representatives from the state department, an assessment company, researchers, parents, and other stakeholders. Inter-rater scoring consistency was calculated as an estimate of reliability. Evidence for validity based on a content analysis, relationships with demographic variables, and consequences of the assessment is also provided in the NASAA Technical Manual.

General Assessments for Idaho and Indiana. The Idaho Standards Achievement Tests (ISAT; Idaho Department of Education, 2008) are multiple-choice assessments of student achievement in reading, language arts, and mathematics. The reading assessment consists of two components: reading process and comprehension/interpretation. The writing assessment consists of two components: writing process and writing components. The mathematics assessment consists of five components: numbers and operation; concepts and principles of measurement; concepts and language of algebra and functions; principles of geometry; and data analysis, probability, and statistics. Item content was determined by state standards across these areas. Internal consistency was calculated as an estimate of reliability. Coefficient alphas ranged from

.80 to .91 for the grade bands included in the current study, and were slightly higher in reading and mathematics than in language arts. Evidence for validity based on content and based on internal structure was also collected.

The Indiana Statewide Testing for Educational Progress-Plus (ISTEP+; Indiana Department of Education, 2009b) is an assessment including multiple-choice, short answer, and essay questions to measure student achievement in language arts and mathematics. Assessment in both language arts and mathematics consist of sections on both basic and applied skills. Item content was determined by state standards across these areas. Internal consistency was calculated as an estimate of reliability. Coefficient alphas ranged from .90 to .94 for the grade bands included in the current study. Evidence for validity based on content and based on relationships with other variables was also collected.

Established Norm-Referenced Measures of Academic and Adaptive Behaviors. The Academic Competence Evaluation Scales (ACES; DiPerna & Elliott, 2000) were designed to measure students' skills, attitudes, and behaviors that contribute to academic competence (DiPerna & Elliott, 2000). The teacher version of the ACES is an 81-item questionnaire with two separate scales (Academic Skills and Academic Enablers). The Academic Skills scale includes three subscales (Reading/Language Arts, Mathematics, and Critical Thinking), and the Academic Enablers scale includes four subscales (Interpersonal Skills, Motivation, Study Skills, and Engagement). Teachers rate items in terms of the level of the students' academic skills compared with grade band expectations from "1 (*Far Below*)" to "5 (*Far Above*)." Teachers rate the existence/ frequency of academically enabling skills from "1 (*Never*)" to "5 (*Almost Always*)." Sum scores at the scale and subscale levels are categorized into Developing (weaknesses), Competent, and Advanced (strengths) ranges. Coefficient alpha for ACES has a mean of .99 on

the Academic Skills and Academic Enablers scales across grade bands. The test-retest reliability of ACES over a two to three weeks interval ranges from .88 - .97. The reported standard error of measurement for the Academic Skills scale ranges 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 scales, Academic Skills and Academic Enablers, were derived from factor analysis. In relation to other standardized achievement (i.e., Iowa Test of Basic Skills; University of Iowa College of Education, 2009) and behavior measures (i.e., Social Skills Rating System; Gresham & Elliott, 1990), ACES also demonstrated solid evidence for the convergent, discriminant, and criterion-related validity.

The Vineland Adaptive Behavior Scales – 2nd Edition (VABS-II; Sparrow, Cicchetti, & Balla, 2006) were 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. Sum scores for composites and domains are categorized into Low (at least 2 standard deviations below the mean), Moderately Low (1 to 2 standard deviations below the mean), Adequate (within a standard deviation of the mean), Moderately High (1 to 2 standard deviations above the mean), and High (at least 2 standard deviations above the mean) adaptive levels. The VABS-II 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-II is adequate for the 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 was lower and ranged from .62 to .78. The standard error of measurement for the various scales range (depending on age) from 3.4 to 6.6.

Procedure

Participant recruitment and data collection took place across all six states during the spring of 2007. All participants in the study were evaluated using the state's AA-AAS, the ACES, and the VABS-II. The ACES and the VABS-II were administered during the week before or just after the completion of each state's AA-AAS, and were scored by the members of the research team who represented each state's department of education. Only the SWD-NE group participated in each state's general assessment, as part of their regular statewide achievement testing.

Data Analyses

Given the nature of the research questions, the data analyses were correlational and exploratory, intended to provide quantitative indices of the relationships among various academic skills and related behaviors. Calculations of the means and standard deviations on the ACES and VABS-II facilitated comparison of the SWD-E and SWD-NE groups. The validity evidence based on relationships with other variables was characterized through Pearson correlations between each state's AA-AAS and the established, norm-referenced assessments. Because of variability in recruiting success, not all states had large numbers of participants available for each analysis. Cells that would be based on less than 10 participants were excluded from these exploratory analyses. The correlations between the same scales were examined for similarity across grade clusters, as well as between eligibility groups, providing a form of within-study replication. In addition to the magnitude of the correlations, we provide information about statistical significance to facilitate comparisons and clarify the probability that such correlations

could occur by chance. Significance tests were made at $\alpha = .05$ and were one-tailed, based on the general prediction that all of the relationships would be positive, regardless of magnitude.

Results

Descriptive analyses of established measures of academic skills, academic enablers, and adaptive behavior supported that SWD-Es and SWD-NEs were distinct groups in the current sample. Among SWD-Es, the mean scores for ACES Academic Skills at both grade bands were in the Developing range. The standard deviation among scores for SWD-Es was about half of the standard deviation among scores for SWD-NEs, reflecting a restriction of range, as almost all of the students in the SWD-E group scored at the very low end of the ACES. The mean Academic Skills scores for SWD-NEs were also in the Developing range at both grade bands, but were higher than were the mean scores for SWD-Es. Using the pooled standard deviation at each grade band, SWD-NEs in elementary school scored 1.39 standard deviations higher on Academic Skills, and SWD-NEs in middle school scored 1.03 standard deviations higher. Table 3 depicts means and standard deviations on the established measures of academic skills, as well as academic enablers and adaptive behavior, across grade bands and eligibility groups.

Insert Table 3 about here

The pattern between groups on ACES Academic Enablers was similar. Among SWD-Es, the mean scores for Academic Enablers at both grade bands were in the Developing range. These scores were higher in the Developing range than were the scores for Academic Skills, and the standard deviations were consistent across groups, indicating that range restriction was not an issue when considering Academic Enablers. The mean Academic Enablers scores for SWD-NEs were also in the Competent range, and were higher at both grade bands than were the mean scores for SWD-Es. Using the pooled standard deviation at each grade band, SWD-NEs in

elementary school scored .95 standard deviations higher on Academic Skills, and SWD-NEs in middle school scored .49 standard deviations higher.

The pattern between SWD-Es and SWD-NEs on the VABS-II also indicated non-overlapping groups. Among SWD-Es, the mean scores on the Adaptive Behavior Composite at both grade bands were at the Low level, nearly 3 standard deviations below the normative mean of 100. The mean Academic Enablers scores for SWD-NEs were at the Adequate level, and were higher at both grade bands than were the mean scores for SWD-Es. Using the normative standard deviation of 15, SWD-NEs in elementary school scored 2.18 standard deviations higher on Academic Skills, and SWD-NEs in middle school scored 1.51 standard deviations higher.

Relationships among AA-AAS Subscales and with General Achievement Tests

The correlations between language arts or reading and mathematics tended to be in the very large range or higher (11 of 14 coefficients) across grade bands and states. The only exceptions to this trend were observed among Nevada students at both grade bands (medium range) and Idaho students at the middle school grade band (large range). Table 4 provides a detailed account of the correlations between the reading and mathematics subscales within each state's alternate assessment at the various grade bands. In Idaho and Indiana, where we had large enough samples to meaningfully disaggregate by group, this trend was prevalent for both SWD-Es (4 of 6 coefficients in the very large range or higher) and for SWD-NEs (5 of 6 coefficients in the very large range or higher). Idaho was the only state that had scores for both reading and language arts; correlations were in the very large range or higher across groups and grade bands, except for the middle school SWD-E sample, in which this correlation was in the medium range.

Insert Table 4 about here

Idaho and Indiana had large enough samples of SWD-NEs for correlations between alternate assessment scores and general assessment scores to be interpreted. The strength of these correlations varied, and the coefficients between content areas that were supposed to represent the same construct were not systematically different from the coefficients that were to represent distinct constructs. Among 10 correlations between tests that were designed to measure the same construct, 6 coefficients were in the medium range and 4 coefficients were in the small range. Among 16 correlations between tests that were designed to measure distinct constructs, 8 were in the medium range or higher, 5 were in the small range, and 3 were in the nonexistent range. Table 5 depicts correlations between AA-AASs and general assessments in Idaho and Indiana.

Insert Table 5 about here

Relationships among AA-AAS Scores and Scores from Established Measures of Other Constructs

Correlations between AA-AAS scores and ACES Academic Skills scores tended to be in the large range or higher (17 of 25 coefficients. This trend held for reading and language arts (9 of 14 coefficients in the large range or higher) and for mathematics (8 of 11 in the large range or higher). This trend also held at the elementary school grade band, as 10 of 13 correlations were in the large range or higher. The correlations were spread more evenly at the middle school grade band, with 4 of 12 coefficients in the medium range, 3 in the large range, and 4 in the very large range. Exceptions to these trends included Nevada in both content areas at the elementary school grade band, one of two Arizona reading assessments at the middle school grade band, and Indiana language arts at the elementary school grade band. These exceptions were in the small or nonexistent ranges. Table 6 depicts the correlations between AA-AAS subscales and established measures. Sample sizes were large enough to disaggregate these correlation coefficients by eligibility status in Idaho and Indiana, but no consistent differences between SWD-Es and SWD-

NEs from these states were observable. For SWD-Es these correlations were spread, with 4 of these 10 coefficients in the medium range, and 3 in the small range. For SWD-NEs, 5 of 10 coefficients were in the small range, with the other coefficients spread across ranges.

A great deal of variation was observed for correlations between AA-AAS scores and ACES Academic Enabler scores. In language arts and reading, 5 of 14 coefficients were in the medium range, 3 were in the large range, and 4 were in the very large range. In mathematics, correlations tended to be in the medium or large ranges (3 of 11 coefficients). Variation was great within grade bands. At the elementary school grade-band, 5 of 13 coefficients were in the very large range, but another 4 coefficients were in the medium range. Correlations at the middle school grade band tended to be in the medium or large ranges (9 of 12 coefficients). Separate trend analyses were done by state when considering the relationship between AA-AAS scores and ACES Academic Enablers scores by eligibility group. In Idaho, correlations at the elementary school grade band were in the large range (3 of 3 coefficients) for SWD-Es, but were in the medium range (3 of 3 coefficients) for SWD-NEs. All six correlation coefficients at the middle school grade band across eligibility groups and content areas were in the small or medium ranges. In Indiana, correlations were in the small range (5 of 8 coefficients) across eligibility groups, with some correlations in the medium (2 of 8 coefficients) or nonexistent (1 of 8) ranges, but without any clear trend between eligibility groups.

Correlations between AA-AAS scores and VABS-II Adaptive Behavior composite scores tended to be in the very large range or higher (15 of 25 coefficients). This trend held at the elementary school grade band (8 of 13 coefficients in the very large range or higher) and at the middle school grade band (7 of 12 coefficients in the very large range and the remaining 5 in the large range). This trend also held for reading and language arts (9 of 14 coefficients in the very

large range) and for mathematics (6 of 11 coefficients in the very large range or higher). Exceptions to this trend were all at the elementary school grade band: Indiana language arts (medium range), Nevada language arts (small range), and Nevada mathematics (medium range). In Idaho and Indiana, correlations between AA-AAS scores and the VABS-II Adaptive Behavior composite were higher for SWD-Es than for SWD-NEs in 11 out of 12 comparisons. These correlations for SWD-Es were similar to those for the sample aggregated across groups, with 11 of 12 coefficients in the large range or higher. The correlation coefficients among SWD-NEs were primarily in the medium range (10 of 12 coefficients).

Correlations between the VABS-II Adaptive Behavior Composite and the ACES Academic Enablers scores were in the very strong range for both SWD-Es (r = .70) and SWD-NEs (r = .67). The correlation between the Adaptive Behavior Composite and ACES Academic Skills scores was in the large range for SWD-Es (r = .50) and was in the medium range (r = .44) for SWD-NEs. The correlations between ACES Academic Enablers and ACES Academic Skills scores were in the medium range for both groups (r = .40 for SWD-Es and r = .37 for SWD-NEs). The correlations between Academic Enablers scores and Academic Skills subscale scores, disaggregated by group, were all in the medium range.

Discussion

Very little research is published that examines the constructs measured by AA-AASs. This is due in part to the challenges of assessing the student population for whom alternate assessments are intended, and in some states, the lack of adequate sample sizes to conduct MTMM studies. The current study addresses these limitations and represents a response to the call by Towles-Reeves et al. (2009) for empirical research that examines the relations among student scores on AA-AASs and other established measures of student achievement. This study

generally replicated the Elliott et al. (2007) study and extended its findings to alternate assessments in five additional states. The study used a known-groups sample to examine the relationships between these states' alternate assessments and grade-level tests of general content standards, along with an established measure of achievement, the ACES, that yields scores based on a large national normative sample. These results go beyond the call by Towles-Reeves et al., by providing an indication of the magnitude of relations among the constructs measured by AA-AASs and the construct of adaptive behavior as measured by the VABS-II. To the extent the current accountability legislation demands assessments that clearly measure the core academic domains, validity studies of AA-AASs should result in the refinement of assessment instruments, with the ultimate intent to measure constructs that are strongly correlated with measures of academic skills and less strongly correlated with measures of adaptive behavior. As a means of reviewing and summarizing the key findings of this multi-state validity study, we re-visit and provided data-based answers to our two motivating questions.

Do the AA-AAS subscales measure distinct content areas, which correlate with the same content areas on each state's general assessment, when both are used on a common sample?

In most states the relationship among content areas (typically the correlation between reading and mathematics) within the AA-AAS are in the range that would be acceptable reliability coefficients for a single, unitary construct. This finding could indicate that some degree of success in one of the content areas (likely reading) is a pre-requisite for success in the other content area (mathematics), and that for students in this sample, variation in the pre-requisite skill explains all of the variation in both sets of scores. This explanation is supported by the pattern of correlations between reading and mathematics scores being lower at the middle school band than at the elementary school band. A second possibility is that both reading and

mathematics are measuring a third construct, such as adaptive behavior. This explanation is supported by high correlations between reading and mathematics scores not only for SWD-Es, but also for SWD-NEs, a group of students who exhibited academic skills, academic enablers, and adaptive behavior much closer to the normative mean on validated measures. It is important to note, however, that regardless of the magnitude of a correlation, it is only one of multiple pieces of evidence required to show that a single construct is being measured.

When taken by SWD-NEs, the general assessment scores and AA-AAS scores in the current sample do not share much variance. In only one case does any combination of the scores from Idaho or Indiana share more than 25% of the variance, and that is the relationship between language arts on the general assessment and mathematics on the AA-AAS. There is no pattern of scores from like content areas (e.g., general assessment reading and AA-AAS reading) sharing any stronger relationships than do combinations of non-like constructs. This finding is consistent with the high correlations observed between AA-AAS reading and mathematics scores. While relationships between general assessment scores and alternate assessment scores in the Elliott et al. (2007) study were stronger (8 of 9 coefficients in the medium range or higher), there likewise existed no pattern of same content areas sharing stronger relationships. These results indicate that when taken by SWD-NEs, the general assessment score reflects a very different construct than is reflected by the AA-AAS score. It is important to remember that the alternate assessments are not designed to be used with SWD-NEs, and these correlations may be limited by a restriction of range, as students in this group generally have high scores on the AA-AASs.

Which constructs (academic skills, academic enablers, adaptive skills) are being measured by AA-AASS when used with students with disabilities who would be eligible (SWD-Es) and students with disabilities who would not be eligible (SWD-NEs) across six states?

Scores from the AA-AASs and the ACES Academic Skills scale shared a degree of variance typically in the range representing related but distinct constructs. This relationship is somewhat stronger at the elementary school band, and does not vary by eligibility group. This latter finding is in contrast with results from Elliott et al. (2007), in which the correlations between scores from AA-AASs and the ACES Academic Skills scale were much larger for SWD-NEs than for SWD-Es. In the current study, the AA-AASs appear to reflect academic skills for SWD-Es to the same degree that they would reflect these skills if used with SWD-NEs.

Relationships between AA-AAS scores and academic enablers vary greatly in terms of strength, indicating that states' AA-AASs reflect enabling behaviors to different degrees. Disaggregated by eligibility group, grade, and state within subsamples from Idaho and Indiana, these constructs share relatively little variance, indicating that they measure two constructs that are not highly related. The one exception to this rule is the Idaho fourth grade SWD-E group, in which the two scores share substantial variance, and may represent a common construct. In the Elliott et al. (2007) study, it was the SWD-NEs whose scores on the alternate assessment shared a great deal of variance with academic enablers, however, both groups yielded correlations between these scores indicative of related constructs. It is likely that the IAA measures academic enablers to a greater degree than the most states included in the study, largely because its scoring rubric incorporates progress during the year along with performance.

Scores for AA-AASs in both reading and mathematics share a great deal of variance with adaptive behaviors across grades, states, and eligibility groups. These two types of measures appeared to reflect very similar constructs across all states, with the exception of Nevada. Some educators would call these academic and work place survival skills, or functional skills. Many students with pre-symbolic and concrete symbolic communication skills need to develop very

basic reading and mathematics skills. Such skills are part of some objectives in the extended content standards with which AA-AASs are aligned. Correlations between academic skills as measured by the IAA and adaptive behaviors were found to be in this same range by Elliott et al. (2007), although in the corresponding analysis SWD-NEs were not evaluated with the VABS-II

The final piece to this discussion on shared variance and overlapping constructs is to consider the degree that academic skills, academic enablers, and adaptive behavior are overlapping constructs, even when measured by established, norm-referenced tools. In the current sample of students with disabilities, the strongest relationship is between adaptive behavior and academic enablers, which appear to be highly related constructs. Adaptive behavior and academic skills were related but distinct constructs, as were academic skills and academic enablers. These relationships likely explain some of the shared variance between adaptive behavior and AA-AASs in this sample. Given the conceptual relationships among these constructs, it seems unlikely that any good measure of academic achievement in this population will be entirely independent of adaptive behavior. Scores from all three measures represent constructs that are distinct but related. A depiction of the relationship among these constructs, based on correlational findings from across grades, states, and groups, is provided as Figure 1.

Insert Figure 1 about here

Practical Implications

Historically, developing appropriate accountability standards for the unique population of students who are eligible for alternate assessments has been a challenging endeavor from several perspectives. First, the students, although a very small proportion of the entire student population, represent an extraordinarily broad range of abilities and needs. Thus, teachers who work with them often must develop curricula and individualized supports to provide appropriate

instruction in their classrooms. While legislative efforts focus on accountability for student learning in the core academic subjects, many teachers opt to maintain a difficult balance in their classroom instruction between academic skills and non-academic skills for these students, in a good-faith effort to provide what they deem to be essential tools for these students to live successful lives outside of school. For some teachers, the ideal alternate assessment is a test that reflects this balance; other stakeholders believe the AA-AASs should measure only the academic skills contained in the content standards. The incompatibility of the two views in the collaborative development of these tests may result in assessments that do not measure what they purport to measure with the same level of precision and focus that has come to be expected of large-scale assessments. Indeed, the results of this study indicate that alternate assessments often measure a number of constructs. While all of these constructs are arguably represented in the totality of an appropriate educational program for this unique group of students, they may not be reflected in the inferences that are ultimately made from the assessment's scores.

Limitations

Research on the academic achievement of students with significant disabilities is challenging. Researchers generally must deal with relatively small, but heterogeneous, samples of students. In the present study, such students were assessed with relatively new evidence-based rating scales completed by teachers and scored by trained scorers. These assessments were designed to be aligned with state content standards and expected classroom instruction. Even when excellent instruction is provided, many students with significant disabilities and a limited number of years of exposure to academic skills still perform at the lowest end of the assessments.

In the current assessments there were a disproportionate number of students near the lower end of the score distribution. This distribution results in a restriction of range and can

deflate correlations. A second challenge with students like those featured in this study is the impracticality of using direct assessments. Thus, to collect information on concurrent measures of related constructs we had to use a number of indirect measures (i.e., teacher rating scales). The ACES and VABS-II are psychometrically sound measures of academic competence and adaptive behavior respectively, but they are both completed by the same teacher, who also provides ratings on an AA-AAS. This limitation may have resulted in increased similarities among the various ratings on different instruments, and thus lead to inflated correlations across constructs.

Conclusions

The current data are the result of an MTMM study that incorporated two eligibility groups, two grade bands, six states, and four criterion measures. The following five main trends are apparent: (1) Reading and mathematics scores from AA-AASs may reflect a unitary construct, (2) When these measures are used with SWD-NEs, the scores yielded are only moderately related to scores from states' general achievement tests, (3) The AA-AAS scores reflect a construct that is related to, but also distinct from, academic skills, (4) The AA-AAS scores reflect a construct that is highly related to adaptive behavior, and (5) Even when measured by established tools, academic skills, academic enablers, and adaptive behavior are related constructs. Collectively these results indicate that the constructs measured by AA-AASs share common ground with related constructs, such as adaptive behavior, academic skills, and academic enablers. While it is a positive sign that none of these relationships are strong enough that the AA-AASs could be a measure any of these three constructs alone, the results of this study indicate that continued efforts must be taken to ensure the constructs measured by alternate assessments are clear and distinct from other constructs, particularly including adaptive behavior.

References

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

Arizona Department of Education & Elliott, S.N. (2006). *Arizona Instrument to Measure Standards – Alternate: Technical report*. Retrieved February 22, 2009 from the Arizona Department of Education Web site: http://www.azed.gov/ess/SpecialProjects/aims-a/

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cohen, J. (1992). *A Power Primer, 112* (1), 155-159.

DiPerna, J.C., & Elliott, S.N. (2000). *Academic Competence Evaluation Scales*. San Antonio, TX: Psychological Corporation.

Elliott, S.N., Compton, E., & Roach, A.T. (2007). Building validity evidence for scores on a state-wide alternate assessment: A contrasting groups, multi-method approach. *Educational Measurement: Issues & Practice, 26* (2), 30-43.

Gresham, F.M., & Elliott, S.N. (1990). *Social Skills Rating System*. Circle Pines, MN: American Guidance Service.

Hawaii Department of Education (2007). *Frequently asked questions about the HSAA and alternate assessments*. Retrieved February 22, 2009 from the Hawaii Department of Education Web site: http://doe.k12.hi.us/nclb/

Hopkins, W.G. (2001). *A scale of magnitudes for effect statistics.* Retrieved February 22, 2009 from the New View of Statistics Web site: http://doe.k12.hi.us/nclb/

Idaho Department of Education (1999). *Idaho Alternate Assessment*. Boise, ID: Author.

Idaho Department of Education (2005). *Idaho Alternate Assessments*. Retrieved February 22, 2009 from the Idaho Department of Education Web site: http://www.sde.state.id.us/SpecialEd/AltAssessment/iaamanual.pdf

Idaho Department of Education (2008). *Idaho Standards Achievement Tests*. Retrieved February 22, 2009 from the Idaho Department of Education Web site: http://www.boardofed.idaho.gov/saa/Technical-Reports.asp

Individuals with Disabilities Education Act, 20 U. S. C. §1400 et seq. (1997).

Individuals with Disabilities Education Improvement Act of 2004. Pub. L. No. 108-446, § 118 Stat. 2647 (2004).

Indiana Department of Education (2009a). *Indiana Standards Tool for Alternate Reporting*. Retrieved February 22, 2009 from the Indiana Department of Education Web site: https://ican.doe.state.in.us/beta/istarinfo.htm

Indiana Department of Education (2009b). *ISTEP+ program manual*. Retrieved February 22, 2009 from the Indiana Department of Education Web site: http://doe.state.in.us/istep/ProgramManual.html

Mississippi Department of Education (2009). *Technical Manual for the Mississippi alternate assessment of the extended curriculum frameworks for students with significant cognitive disabilities*. Retrieved February 22, 2009 from the Mississippi Department of Education Web site: http://www.mde.k12.ms.us/maaecf/

Nevada Department of Education (2009). *Nevada Alternate Scales of Academic Achievement*. Retrieved February 22, 2009 from the Nevada Department of Education Web site: http://nde.doe.nv.gov/Assessment_NASAA.htm

No Child Left Behind Act, 20 U. S. C. § 16301 et seq. (2001).

Sparrow, S.S., Balla, D.A., & Cicchetti, D.V. (1985). *Vineland Adaptive Behavior Scales: Classroom Edition.* Circle Pines, MN: American Guidance Service.

Sparrow, S.S., Cicchetti, D.V. & Balla, D.A. (2005). *Vineland-II teacher rating form manual*. Pearson Assessment. Minnesota, USA.

SRI International (2009). *National study of alternate assessments*. Retrieved February 22, 2009 from the SRI International Web site: http://policyweb.sri.com/cehs/projects/displayProject.jsp?Nick=nsaa

Towles-Reeves, E., Kleinert, H., & Muhomba, M. (2009). Alternate assessment: Have we learned anything new. *Exceptional Children, 75* (2), 233-252.

University of Iowa College of Education (2009) *Iowa Tests of Basic Skills*. Retrieved February 22, 2009 from the Iowa Testing Programs Web site: http://www.education.uiowa.edu/itp/itbs/

Footnote

The study reported in this manuscript was part of a U.S. Department of Education, Office of Elementary and Secondary Schools' Enhanced Assessment Grant (#S368A060012) to the Idaho Department of Education. The ideas and viewpoints represented in this manuscript are those of the authors only.

Table 1.

Demographics by Grade Band and Group

		Elementary School *n* (%)		Middle School *n* (%)	
		SWD-Es	SWD-NEs	SWD-Es	SWD-NEs
Gender	Female	79 (41%)	73 (35%)	66 (40%)	56 (38%)
	Male	114 (59%)	136(65%)	101 (60%)	93 (62%)
Ethnicity	African American	16 (8%)	29 (14%)	12 (7%)	26 (17%)
	Asian American	13 (7%)	5 (2%)	8 (5%)	1 (1%)
	European American	99 (51%)	141 (67%)	95 (57%)	98 (66%)
	Latino American	42 (22%)	24 (11%)	36 (21%)	13 (9%)
	Native American	1 (1%)	1 (0%)	6 (4%)	3 (2%)
	Native American	1 (1%)	1 (0%)	6 (4%)	3 (2%)
	Other / Not identified	13 (7%)	7 (3%)	5 (3%)	7 (5%)
State	Arizona	46 (24%)	23 (11%)	49 (31%)	13 (9%)
	Hawaii	26 (14%)	6 (3%)	12 (8%)	2 (1%)
	Idaho	20 (10%)	21 (10%)	20 (13%)	20 (14%)
	Indiana	40 (21%)	115 (57%)	43 (27%)	86 (60%)
	Mississippi	16 (8%)	11 (5%	11 (7%)	11 (8%)
	Nevada	43 (23%)	26 (13%)	24 (15%)	11 (8%)
Total		193 (48%)	209 (52%)	168 (53%)	149 (47%)

Table 2.

Disability Status Frequencies by Grade Band and Group

Disability Category	Elementary School *n* (%)		Middle School *n* (%)	
	SWD-Es	SWD-NEs	SWD-Es	SWD-NEs
Autism	42 (22%)	15 (7%)	26 (15%)	10 (7%)
Deaf-Blindness	4 (2%)	0 (0%)	2 (1%)	0 (0%)
Deafness or Hearing Impairment	2 (1%)	1 (0%)	2 (1%)	0 (0%)
Developmental Delay	1 (1%)	1 (0%)	0 (0%)	0 (0%)
Emotional Disturbance	1 (1%)	12 (6%)	1 (1%)	19 (13%)
Mental Retardation	91 (47%)	23 (11%)	97 (58%)	31 (21%)
Orthopedic Impairment	1 (1%)	1 (0%)	2 (1%)	0 (0%)
Other Health Impairment	9 (5%)	17 (8%)	7 (4%)	12 (8%)
Specific Learning Disability	2 (1%)	115 (55%)	4 (2%)	71 (48%)
Speech or Language Impairment	0 (0%)	18 (9%)	1 (1%)	4 (3%)
Traumatic Brain Injury	0 (0%)	1 (0%)	1 (1%)	0 (0%)
Visual Impairment	2 (1%)	1 (0%)	1 (1%)	1 (1%)
Multiple Disabilities	34 (18%)	2 (1%)	22 (13%)	0 (0%)
Not known	4 (2%)	2 (1%)	2 (1%)	1 (1%)

Table 3.

Descriptive Statistics for Academic Skills, Academic Behavior, and Adaptive Behavior by Grade Band and Group

	Elementary School Mean (SD)		Middle School Mean (SD)	
	SWD-Es	SWD-NEs	SWD-Es	SWD-NEs
Academic Skills	38.81 (9.40)	64.86 (24.33)	40.16 (15.74)	63.36 (27.62)
Reading	13.32 (3.83)	21.91 (7.72)	13.38 (12.12)	20.98 (13.78)
Mathematics	9.37 (2.50)	15.96 (6.46)	9.35 (11.75)	17.49 (13.28)
Critical Thinking	16.18 (3.83)	27.24 (9.71)	17.03 (12.40)	28.25 (15.02)
Academic Enablers	93.21 (33.25)	129.00 (41.11)	105.92 (38.77)	123.28 (31.78)
Interpersonal Skills	32.19 (9.88)	39.16 (9.25)	34.74 (15.66)	38.15 (14.57)
Engagement	17.90 (8.74)	27.30 (7.52)	19.79 (14.02)	24.65 (13.01)
Motivation	21.70 (8.35)	30.86 (10.38)	24.18 (14.43)	29.52 (14.83)
Study Skills	23.41 (9.56)	34.35 (12.24)	27.85 (15.55)	32.11 (15.30)
Adaptive Behavior	60.44 (17.66)	93.12 (13.04)	64.03 (20.46)	86.65 (16.00)
Communication	61.03 (17.07)	93.38 (12.61)	63.97 (19.97)	89.03 (15.32)
Daily Living Skills	60.86 (18.98)	94.89 (12.49)	65.29 (21.05)	86.58 (15.66)
Socialization	69.41 (13.06)	92.83 (15.96)	71.37 (18.84)	87.41 (18.50)
Motor Skills	71.73 (24.59)	115.17 (17.56)	81.47 (30.76)	114.90 (19.44)

Table 4.

Correlations between Alternate Assessment Subscales by Grade Band

State	Comparison	Elementary School	High School
Arizona	Reading / Mathematics	.97*[1]	.93*[1]
Hawaii	Reading / Mathematics	.92	.87
Idaho	Reading / Mathematics	.94*	.83*
	Language / Mathematics	.93*	.89*
	Reading / Language	.97*[2]	.83*[2]
Indiana	Language / Mathematics	.61*	.93*
Mississippi	Reading / Mathematics	.87*	.95*
Nevada	Reading / Mathematics	.43	.34

* $p < .05$ (one-tailed).

[1]Correlations reported for AZ are as follows: Reading 1 and Mathematics 1 for grade 4 and Reading 2 and Mathematics 2 for grade 8. No other students in the AZ sample took more than one subscale of Arizona's assessment.

Table 5.

Correlations of General Assessment and Alternate Assessment Scores for Non-Eligible Students

		General Assessment		
	AA-AAS Subscale	Reading[2]	Language Arts[2]	Mathematics
		Elementary School		
Idaho	Reading	.31	.31	.12
	Language Arts	.24	.22	.07
	Mathematics	.31	.40	.26
Indiana	Language Arts	--	.13	.19
	Mathematics	--	.13	.18
		Middle School		
Idaho	Reading	.37	.43	.07
	Language Arts	.33	.43	.07
	Mathematics	.40	.68*	.42
Indiana	Language Arts	--	.30*	.37*
	Mathematics	--	.28*	.48*

* $p < .05$ (one-tailed).

Table 6.

Correlations of AA-AAS Subscales with Established Measures by Grade Band.

State		Elementary School			Middle School		
		Academic Skills	Academic Enablers	Adaptive Behavior	Academic Skills	Academic Enablers	Adaptive Behavior
Arizona	Reading 1	.63*	.66*	.86*	.18	.80*	.88*
	Reading 2	--	--	--	.82*	.18	.70*
	Mathematics	.69*	.69*	.86*	.82*	.21	.67*
Hawaii	Language Arts	.61*	.86*	.78*	.55	.65*	.85*
	Mathematics	.61*	.78*	.87*	.47	.58	.83*
Idaho	Reading	.68*	.72*	.86*	.67*	.36*	.59*
	Language Arts	.61*	.72*	.84*	.72*	.52*	.72*
	Mathematics	.74*	.73*	.90*	.79*	.42*	.70*
Indiana	Language Arts	.29*	.31*	.47*	.46*	.45*	.72*
	Mathematics	.52*	.48*	.75*	.45*	.41*	.65*
Mississippi	Language Arts	.57*	.32	.57*	.42	.40	.53*
	Mathematics	.66*	.17	.57*	.59	.61	.58
Nevada	Language Arts	-.07	.18	.17	--	--	--
	Mathematics	.06	.33*	.47*	--	--	--

* $p < .05$ (one-tailed).

Note. Correlations were not reported when the sample size was less than 10.

Figure Caption

Figure 1. *Relationships among achievement, as measured on states' AA-AASs, and established measures of academic skills, academic enablers, and adaptive behavior.*