

VI. Technical Adequacy

Research on the extended assessments used as part of the Alaska Alternate Assessment has been published in several outlets over the past few years. We have published three papers on reliability and validity in refereed journals.

Tindal, G., McDonald, 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(4), 481-494.

Alternate assessments in reading and math: Development and validation for students with significant disabilities. *Exceptional Children*, 69(4), 481-494.

This study determined that the extended measures in reading and mathematics functioned well when administered to hundreds of students in various disability groups comprising 1% of grade level population. The results reported three types of criterion-related evidence. Task variation was generally developmental (counting numbers was higher than adding or subtracting numbers). Differences were found as a function of disability groups that were consistent with opportunity to learn (e.g., students with significant mental retardation scored lower than students with severe learning disabilities). Finally, judges were able to sort students into groups of proficiency that were consistent with performance on the tasks. See Appendix 14a.

Yovanoff, P., & Tindal, G. (2006). Scaling early reading alternate assessments with statewide measures. *Exceptional Children*, 73(2), 184-201.

This research on 800 grade 2 and 3 students reports on three outcomes from several reading subtests used in the extended assessments (reading words, blending sounds, reading names, reading sentences, oral reading of easy passages, and oral reading of difficult passages): (a) the unidimensionality of the extended production measures with selected statewide items, (b) a Rasch model for scoring items dichotomously that would place reading extended measures as much less difficult than items from a statewide test, and (c) differences between special and general education students that would not reflect DIF. All three hypothesized outcomes were confirmed. See Appendix 14b.

Crawford, L., Tindal, G., & Carpenterr, III, D. M. (2006). Exploring the validity of the Oregon Extended Writing Assessment. *Journal of Special Education*, 40(1), 16-27.

Using state test data from approximately 1,200 students with significant disabilities, this study determined that the writing tasks had adequate reliability. Average interrater agreement for each task (across a random selection of 80 protocols) ranged from .89 to .98. Overall agreement across all writing tasks was .93. (p. 22). For the first four tasks, alpha coefficients of .97, .97, .94, and .77). Finally, a four-factor model reflected adequate internal representation that mapped closely with the factor structure the test was designed to reflect: Factor 1 was comprised of *story writing*, Factor 2 contains *copying letters and words*, Factor 3 consists of *dictating words and sentences*, and Factor 4 encompasses *writing sentences*. See Appendix 14c.

We also have conducted considerable research on the assessment of students with the most significant disabilities as part an Enhanced Assessment Instrument grant (Project DAATA: Developing Alternate Assessment Technical Adequacy) with three studies on reliability, generalizability, and criterion-related evidence using assessment systems that are extremely comparable to (and use subtests from) the extended assessment used in Alaska.

Tindal, G., Almond, P., & Yovanoff, P. (2006). *Reliability of alternate assessments*. Unpublished research report from Project DAATA (retrieved from <http://daata.org> on March 23, 2007).

Various types of reliability are described in this paper, using tasks from the Alaska Alternate Assessment. Using item level data in which approximately 500 students in grade 5 took the *Reading Words* subtest, three reliability coefficients are described: (a) split half (which was .83 when corrected for length), (b) Cronbach's Alpha (which was .67 and .77 for the two parts of 4 items each), and (c) parallel form (which was .84). See Appendix 14d.

Tindal, G., Almond, P., Geller, J., & Yovanoff, P. (2006). *Generalizability theory applied to reading assessments for students with significant cognitive disabilities*. Unpublished research report from Project DAATA (retrieved from <http://daata.org> on March 23, 2007).

This study of 81 5th grade students with significant cognitive disabilities that was conducted over 7 states included six different reading tasks that are essentially from the Alaska Alternate Assessment: letter signs and symbols, letter naming, word reading, sentence reading, passage reading, and comprehension. Using generalizability theory, several facets were investigated, including person, task, item, administration format (expressive versus receptive), form, and rater. Consistent with much of the earlier empirical research in this area, an interaction was found between person and task with several caveats: (a) receptive tasks were easier, different tasks were not of the same difficulty when administered receptively or expressively, some tasks interacted with format, and some tasks showed considerable variation. The D-study revealed that receptive tasks interacted more with tasks, requiring more items to achieve comparable levels of reliability. Rater variance was negligible. See Appendix 14e.

Tindal, G., Almond, P., Geller, J., & Yovanoff, P. (2006). *Criterion-related evidence for alternate assessment in reading and mathematics*. Unpublished research report from Project DAATA (retrieved from <http://daata.org> on March 23, 2007).

This report uses Campbell and Fiske's¹ multi-trait, multi-method matrix to interpret data from an alternate assessment that used tasks from the Alaska Alternate Mathematics Assessment (money, problem-solving, measurement, numbers, time, tables and graphs, probabilities, and subtraction) and Reading Assessment (signs and symbols, letter naming, word reading, sentence reading, passage reading, and passage comprehension). In addition, an Assistive Technology (AT) survey was administered. Correlations were computed across tasks within each subject area and between reading and mathematics.

¹ Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait, multi-method matrix. In W. A. Mehrens & R. L. Ebel (Eds.), *Principles of educational and psychological measurement: A book of selected readings* (pp 273-302). Chicago, IL: Rand McNally & Company.

The correlations were high in both instances. Using the AT survey to group students into various language proficiency levels, patterns were found between various language–communication skills and performance on reading and mathematics tasks. See Appendix 14f.

Finally, the Alaska Alternate Assessment has been implemented statewide in Oregon from 1998 to 2005 as part of their alternate assessment. The following data reflect the last year of its implementation that consists of cousin items used in the Alaska Alternate Assessment. See Appendix 14g for tables in reading, writing, mathematics, and science. In each area, the following five issues are addressed: (a) item intercorrelations within tasks and classical reliability, (b) item response theory (IRT) analysis of item and person fit as well as reliability, (c) individual task intercorrelations, (d) differential item functioning (DIF) among age-grade bands of elementary, middle, and high, and (e) task performance levels for each age-grade band, including the mean, standard deviation, and count.

Reading

A total of 13 tasks are present in reading. Intercorrelations among the items within tasks are presented in Table 1a. As expected, the average correlation is moderate. Task reliability is quite high for all but one task (n=12) and is near or above .80; the exception (Task 1 is above .50).

Using Item Response Theory (1 parameter model), item fit statistics and person fit statistics are presented for each task in Table 1b. The average fit is near 1 (with the only exception being Task 5). All reliability coefficients are high (near and above .90) when this statistic can be calculated.

Table 1c presents the intercorrelations of all tasks. They are moderate, reflecting unique information from combing the tasks into a single test. The average correlation is near .35.

Differential item functioning (DIF) analyses are presented in Table 1d. Five tasks (1, 2, 4, 8, and 9) reflect age-grade differences (DIF) in multiple items between elementary-middle, elementary-high, or middle-high grade bands when controlling for ability. Two tasks reflect no DIF on any items; six tasks reflect DIF on only one or two items.

Descriptive statistics on approximately 3,000 students are presented in Table 1e. Most tasks reflect a range of performance with an appropriate relation between the mean and the standard deviation as well as consistency across age-grade bands (elementary, junior high, and high school). Task 2 and Tasks 8-11 show consistent improvement across the age-grade bands.

Writing

A total of 10 tasks are present in writing. Intercorrelations among the items within tasks are presented in Table 2a. With the exception of task 7, the average correlation is moderate to high. Task reliability is quite high with all tasks in the range of .75 to .90.

Using Item Response Theory (1 parameter model), item fit statistics and person fit statistics are presented for each task in Table 2b for the four tasks that are appropriately scaled. The average fit is near 1 for all tasks. All reliability coefficients are high (near or above .90).

Table 2c presents the intercorrelations of all tasks. They are moderate (except task 4), reflecting somewhat common information from combining the tasks into a single test. The average correlation is near .40.

Differential item functioning (DIF) analyses for four tasks are presented in Table 2d. One task (Task 1) reflects age-grade differences (DIF) in multiple items between elementary-middle, elementary-high, or middle-high grade bands when controlling for ability. Three tasks reflect DIF on only one or two items.

Descriptive statistics on approximately 1,200 students are presented in Table 2e. Most tasks reflect a range of performance with an appropriate relation between the mean and the standard deviation as well as consistency across age-grade bands (elementary, junior high, and high school). Tasks 6, 9, and 10 show consistent improvement across the age-grade bands.

Mathematics

A total of 22 tasks are present in mathematics. Intercorrelations among the items within tasks are presented in Table 3a. As expected, the average correlation is moderate. Task reliability is quite high across most (n=13) tasks (near and above .70); the remaining coefficients are near or above .50.

Using Item Response Theory (1 parameter model), item fit statistics and person fit statistics are presented for each task in Table 3b. The average fit is near 1 (with the only exception being Task 16). Reliability coefficients are all quite high (near and above .90) with the exception of task 16 and 19.

Table 3c presents the intercorrelations of all tasks. They range from quite low to moderate, reflecting unique information from combining the tasks into a single test. The average correlation is near .35.

Differential item functioning (DIF) analyses are presented in Table 3d. Only two tasks (4 and 7) reflect age-grade differences (DIF) in multiple items between elementary-middle, elementary-high, or middle-high grade bands when controlling for ability. Nine tasks reflect no DIF on any items; eight tasks reflect DIF on only one or two items.

Descriptive statistics on approximately 2,000 students are presented in Table 3e. Most tasks reflect a range of performance with an appropriate relation between the mean and the standard deviation as well as consistency across age-grade bands (elementary, junior high, and high school). Only tasks 4, 7, and 10 show consistent improvement across the age-grade bands.

Science

A total of 11 tasks are present in science. Intercorrelations among the items within tasks are presented in Table 4a. As expected, the average correlation is moderate. Task reliability is moderate with five tasks in the range of .50-.62 and six tasks in the .65 to .83.

Using Item Response Theory (1 parameter model), item fit statistics and person fit statistics are presented for each task in Table 4b. The average fit is near 1 for all tasks. All reliability coefficients are high (near or above .90).

Table 4c presents the intercorrelations of all tasks. They are moderately high, reflecting somewhat common information from combining the tasks into a single test. The average correlation is near .45.

Differential item functioning (DIF) analyses are presented in Table 4d. No tasks reflect age-grade differences (DIF) in multiple items between elementary-middle, elementary-high, or middle-high grade bands when controlling for ability. Five tasks reflect no DIF on any items; six tasks reflect DIF on only one or two items.

Descriptive statistics on approximately 800 students are presented in Table 4e. Most tasks reflect a range of performance with an appropriate relation between the mean and the standard deviation as well as consistency across age-grade bands (elementary, junior high, and high school). No tasks show consistent improvement across the age-grade bands.