Technical Report # 0915

Internal Consistency of General Outcome Measures in Grades 1-8

Daniel Anderson

Gerald Tindal

Julie Alonzo

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Note: Funds for this data set used to generate this report come from a federal grant awarded to the UO from the Institute of Education Sciences, U.S. Department of Education: *Assessment for Accountability (*PR/Award # R324A070188 funded from June 2008 – May 2011).

Abstract

We developed alternate forms of a math test for use in both screening students at risk of failure and monitoring their progress over time. In this technical report, we present results of the screener, used in the fall of 2009. The 48-item test was aligned to the National Council of Teachers of Mathematics (NCTM) Curriculum Focal Point Standards and was administered on a computer to all students from a single school district. The data were analyzed using Cronbach's alpha to reflect the internal consistency of the test forms. The results suggest sufficient consistency to use the scores in screening students within a district.

Internal consistency of general outcome measures in grades 1-8

Reliability is generally described in terms of score 'stability.' The *Standards for Educational and Psychological Testing* (1999) defines reliability as "the consistency of [such] measurements when the testing procedure is repeated on a population of individuals or groups" (p. 25). Reliability typically refers to the measurement error that is introduced into the "entire measurement process" (p. 27) and both limits the degree to which generalizations can be made beyond the specific testing event and quantifies the confidence that can be held in the value assigned to any performance. "Reliability data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores" (p. 31). Specifically, for the purposes of this technical report, we are concerned about the reliability (internal consistency) of behavior on items within each grade level test.

Reliability requires quantifying the measurement error associated with (a) observed behaviors, and (b) numeric scores assigned to our observations. We focus on internal consistency if we believe we have introduced error from our specific sample of items, tasks, or behaviors. In this technical report, we present results using Cronbach's alpha, which is based on the concepts of observed score variance, true score, and error score variance (Feldt & Brennan, 1989). We represent reliability as the ratio of true score variance to observed score variance (true score plus error variance). Ideally, we want to diminish error and maintain an observed score that is largely composed of true score. Generally, as error score variance diminishes, the correlation of observed and true scores approaches the maximum value '1'. Conventional reliability indices and estimates of standard error allow us to understand the stability (consistency) of the score within the distribution and further calculate confidence intervals around the true score.

Methods

Setting and Subjects

The following demographics are from the spring of 2009. The first grade sample was comprised of 1,314 students: 50.8% female, 73.1% White, and 11% receiving special education services. In grade two, the sample included 1,296 students, with 47.5% female, 74.5% White, and 13.3% receiving special education services. The third grade sample consisted of 1,280 students; 48% female, 25% historically low-achieving, 43% economically disadvantaged, and 16% receiving special education services. In fourth grade, the sample consisted of 1,334 students: 51% female, 25% historically low-achieving, 43% economically disadvantaged, and 17% receiving special education services**.** The fifth grade sample consisted of 1,211 students: 50% female, 23% historically low-achieving, 41% economically disadvantaged, and 18% receiving special education services. The sixth grade sample consisted of 1,115 students: 52% female, 25% historically low-achieving, 38% economically disadvantaged, and 16% receiving special education services. The seventh grade sample consisted of 1,306 students: 49% female, 25% historically low-achieving, 38% economically disadvantaged, and 15% receiving special education services. The eighth grade sample consisted of 1,359 students: 49% female, 24% historically low-achieving, 35% economically disadvantaged, and 14% receiving special education services.

Measurement/Instrument Development

We focused on developing three benchmark measures (fall, winter, and spring) that address three critical focal point standards and 10 forms for each focal point. We used a structured item writing process to ensure the tasks were developed systematically using principles of universal design; then we reviewed the items for bias and sensitivity. We addressed *reliability* by collecting procedural evidence as part of the training of teachers in the

administration of the test to ensure proper implementation statewide. During measurement development, we piloted the items and calculated IRT fit statistics for each item.

We used the procedures described by Ketterlin-Geller, Alonzo, Braun-Monegan, and Tindal (2007) with items formatted in simplified-language.

- Replace indirect sentences with direct sentences.
- Reduce the number of words.
- Rewrite conditional phrases.
- Replace long words with shorter synonyms.
- Organize the information into a logical sequence.
- Do not replace mathematics-specific vocabulary.

Where needed, the ELD Core Vocabulary list was used to maintain grade-level readability while maintaining the integrity of the targeted mathematics. To date, we have published four technical reports on the development of mathematics items in each of several grade levels.

Lai, C.F., Alonzo, J., Tindal, G. (2009). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 5* (Technical Report No. 0901). Eugene, OR: Behavioral Research and Teaching: University of Oregon.

Alonzo, J., Lai, C.F., Tindal, G. (2009). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 3* (Technical Report No. 0902). Eugene, OR: Behavioral Research and Teaching: University of Oregon.

Alonzo, J., Lai, C.F., Tindal, G. (2009). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 4* (Technical Report No. 0903). Eugene, OR: Behavioral Research and Teaching:

University of Oregon.

Lai, C.F., Alonzo, J., Tindal, G. (2009). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 8* (Technical Report No. 0904). Eugene, OR: Behavioral Research and Teaching: University of Oregon.

Additional technical reports documenting the development of the mathematics measures at the other grade levels are in press.

All items were equated using a Rasch 1PL model and are loaded onto a web-based system for districts to use. All items were aligned with grade level standards, as required by the 2% regulations, and a formal alignment of items to grade level content standards is planned for January 2010, using Tindal's (2005)[1] adaptation of Webb's process, focusing on categorical concurrence, range of knowledge, depth of knowledge, and balance of representation.

Design and Operational Procedures

The test is computer-based: individual items are presented on a screen with three options. Each option is presented in a large bracketed area that can be selected by clicking anywhere in the area. For this study, the tests were group-administered in computer labs (N.B. An algorithm is used to randomly rotate options for each problem to prevent students who are sitting close to each other from copying responses). The computer scores each response and provides an export of the data after the testing window has been closed by district administrators.

[1] Tindal, G. (2005) *Alignment of Alternate Assessments Using the Webb System*. Washington, DC: Council of Chief State School Officers.

Data Preparation and Analysis

After the normative period was done, all data were transferred to a data file in which individual items were depicted with three fields: (a) the option selected, (b) the correctness of the item (0=incorrect and 1=correct), and (c) the focal point domain. The following field codes were used to organize the data file. The column headers for each file were different to reflect the focal points for each grade. The following key maps grades to test types and test names.

Kindergarten	=>	'math_numop', 'math_geo', 'math_msmt'
Grade 1	=>	'math_numop', 'math_geo', 'math_numopalg'
Grade 2	=>	'math_numop', 'math_geo', 'math_numopalg'
Grade 3	=>	'math_numop', 'math_geo', 'math_numopalg'
Grade 4	=>	'math_numop', 'math_mda', 'math_numopalg'
Grade 5	=>	'math_numop', 'math_gma', 'math_numopalg'
Grade 6	=>	'math_numop', 'math_alg', 'math_numoprat'
Grade 7	=>	'math_noag', 'math_mga', 'math_numopalg'
Grade 8	=>	'math_alg', 'math_geomsmt', 'math_danoa'

Test names

'math_numop'	=>	'Math Numbers and Operations',
'math_geo'	=>	'Math Geometry',
'math_mda'	=>	'Math Measurement',
'math_gma'	=>	'Math Geometry Measurement and Algebra',
'math_noag'	=>	'Math Nums Ops Algebra and Geometry',
'math_mga'	=>	'Math Measurement Geometry and Algebra',
'math_numopalg'	=>	'Math Numbers Operations and Algebra',
'math_alg'	=>	'Math Algebra',
'math_numoprat'	=>	'Math Numbers Operations and Ratios',
'math_geomsmt'	=>	'Math Geometry and Measurement',
'math_msmt'	=>	'Math Measurement',
'math_danoa'	=>	'Math Data Analysis Nums Ops and Algebra'

Results

Table 1 reports the descriptive statistics for the sample by grade-level (1-8), and demographic variables collected from the district in the spring of 2009. Because the data reported in this manuscript were gathered in the fall of 2009, the demographic information should be viewed as an approximation of the demographics at the time of the study and not the exact characteristics of the students in our sample. The mathematics tests analyzed in this study

were administered during the fall of the 2009 school year in one mid-sized district in Oregon. The grand mean, range, minimum/maximum values, and variance are reported by grade-level and NCTM focal point standard in Table 2. Each grade-level test was comprised of 48 total items, 16 for each focal point. All items were coded dichotomously, with 0 representing an incorrect response and 1 representing a correct response. Table 3 reports the inter-item correlation mean, range, and Cronbach's alpha by grade-level and focal point standard. The overall Cronbach's alpha, standard deviation, and standard error of measurement are reported in Table 4.

Grade One

The NCTM focal points assessed on the grade one assessments are: (a) number and operations, (b) geometry, and (c) number and operations and algebra. The sample for this study ranged from 205-207 valid cases. Overall, the items had a mean of .57, with a minimum value of .18 and a maximum value of .97, producing a range of .79. The items had a mean variance of .2, with a minimum value of .03 and a maximum value of .25, producing a range of .22. The mean number of items correct was 27.22 out of the 48 total items, with a standard deviation of 6.83. The inter-item correlations had a mean of .08, with a minimum value of -.15 and a maximum of .41, producing a range of .56. A Cronbach's alpha of .82 indicated strong internal consistency.

Grade Two

The grade two mathematics tests measure: (a) number and operations, (b) geometry, and (c) number and operations and algebra. The sample for this study ranged from 16-80 valid cases. Overall, the items had a mean of .52, with a minimum value of .07 and a maximum value of .92, producing a range of .85. The mean variance of the items was .23, with a minimum value of .07 and a maximum value of .27, producing a range of .20. The mean number of items correct was 25.08 out of the 48 total items, with a standard deviation of 8.53. The inter-item correlations had

a mean of .12, with a minimum value of -.72 and a maximum of .86, producing a range of 1.58. A Cronbach's alpha of .86 indicated strong internal consistency.

Grade Three

The NCTM focal points assessed on our grade three tests include: (a) geometry, (b) number and operations, and (c) number and operations and algebra. The sample size ranged from 1,222-1,231 valid cases. Overall, the items had a mean of .65, with a minimum value of .14 and a maximum value of .99, producing a range of .86. The mean variance of the items was .17, with a minimum value of .01 and a maximum value of .25, producing a range of .24. The mean number of items correct was 31.38 out of the 48 total items, with a standard deviation of 6.30. The inter-item correlations had a mean of .08, with a minimum value of -.10 and a maximum of .69, producing a range of .79. A Cronbach's alpha of .80 indicated strong internal consistency.

Grade Four

The NCTM focal points assessed on the grade four test include: (a) number and operations, (b) measurement, and (c) number and operations and algebra. The sample size ranged from 1,205-1,211 valid cases. Overall, the items had a mean of .71, with a minimum value of .13 and a maximum value of .99, producing a range of .87. The mean variance of the items was .17, with a minimum value of .00 and a maximum value of .25, producing a range of .25. The mean number of items correct was 33.87 out of the 48 total items, with a standard deviation of 7.13. The inter-item correlations had a mean of .11, with a minimum value of -.08 and a maximum of .70, producing a range of .78. A Cronbach's alpha of .86 indicated strong internal consistency.

Grade Five

The NCTM focal points assessed on the grade five test include: (a) number and operations, (b) geometry, measurement, and algebra, and (c) number and operations and algebra. The sample size ranged from 1,269-1,270 valid cases. Overall, the items had a mean of .71, with a minimum value of .24 and a maximum value of .97, producing a range of .74. The mean variance of the items was .17, with a minimum value of .03 and a maximum value of .25, producing a range of .22. The mean number of items correct was 34.16 out of the 48 total items, with a standard deviation of 7.04. The inter-item correlations had a mean of .11, with a minimum value of -.07 and a maximum of .66, producing a range of .73. A Cronbach's alpha of .85 indicated strong internal consistency.

Grade Six

The NCTM focal points assessed on the grade six test include: (a) number and operations, (b) algebra, and (c) number and operations and algebra. The sample size ranged from 1,238-1,249 valid cases. Overall, the items had a mean of .69, with a minimum value of .34 and a maximum value of .99, producing a range of .65. The mean variance of the items was .17, with a minimum value of .01 and a maximum value of .25, producing a range of .24. The mean number of items correct was 33.29 out of the 48 total items, with a standard deviation of 7.34. The inter-item correlations had a mean of .12, with a minimum value of -.07 and a maximum of .50, producing a range of .56. A Cronbach's alpha of .87 indicated strong internal consistency.

Grade Seven

The NCTM focal points assessed on the grade seven test include: (a) number and operations, algebra, and geometry, (b) measurement, geometry, and algebra, and (c) number and operations and algebra. The sample size ranged from 707-720 valid cases. Overall, the items had a mean of .58, with a minimum value of .15 and a maximum value of .93, producing a range of

.78. The mean variance of the items was .21, with a minimum value of .07 and a maximum value of .25, producing a range of .19. The mean number of items correct was 27.98 out of the 48 total items, with a standard deviation of 7.88. The inter-item correlations had a mean of .13, with a minimum value of -.12 and a maximum of .52, producing a range of .64. A Cronbach's alpha of .86 indicated strong internal consistency.

Grade Eight

The NCTM focal points assessed on the grade eight test include: (a) algebra, (b) geometry and measurement, and (c) data analysis, number and operations, and algebra. The sample size ranged from 715-723 valid cases. Overall, the items had a mean of .57, with a minimum value of .24 and a maximum value of .96, producing a range of .71. The mean variance of the items was .21, with a minimum value of .04 and a maximum value of .25, producing a range of .21. The mean number of items correct was 27.37 out of the 48 total items, with a standard deviation of 7.40. The inter-item correlations had a mean of .09, with a minimum value of -.09 and a maximum of .37, producing a range of .45. A Cronbach's alpha of .83 indicated strong internal consistency.

Discussion

The internal consistency of the mathematics screener appears to be adequate when all 48 items are used to reflect a total score. With individual subtests, however, the levels of reliability are consistently lower, not a surprising finding. In great part, reliability is a function of the number of items, and the subtests are considerably shorter than the screener when all three subtests are included. Therefore, caution is warranted when reporting subtest performance when using the mathematics screener.

We developed the screener with reference to specific and consistent sampling from these (subtest) domains to ensure adequate alignment with the focal points. We also developed

alternate forms for progress monitoring for each of these subtest domains so that teachers could track growth over time in a domain-specific manner. When used in this manner, short comings in any single subtest performance value can be adjudicated by collecting data on progress, with multiple measures over time.

Table 1
Descriptive Statistics from Spring of 2009.

				Gender		Ethnicity						
SD 1	Count	Sped	Econ. dis.	Male	Female	AmerInd/ AK-Nat.	Asian/Pac -Islander	Black	Latino	White	Multi-ethnic	Decline
Gr1	1314	145 (11%)	.	647 (51%)	667 (51%)	32 (2%)	85 (6%)	40 (3%)	147 (11%)	961 (73%)	.	49 (4%)
Gr2	1296	173 (13%)	.	681 (53%)	615 (48%)	31 (2%)	75 (6%)	49 (4%)	139 (11%)	971 (75%)	.	31 (2%)
Gr3	1280	200 (16%)	554 (43%)	632 (49%)	611 (48%)	20 (2%)	52 (4%)	28 (2%)	109 (9%)	892 (70 %)	110 (9%)	32 (3%)
Gr4	1334	224 (17%)	559 (42%)	659 (49%)	675 (51%)	21 (2%)	69 (5%)	32 (2%)	103 (8%)	956 (72%)	105 (8%)	48 (4%)
Gr5	1211	217 (18%)	495 (41%)	607 (50%)	604 (50%)	35 (3%)	53 (4%)	34 (3%)	79 (7%)	867 (72%)	72 (6%)	71 (6%)
Gr6	1115	175 (16%)	420 (38%)	532 (48%)	583 (52%)	14 (1%)	56 (5%)	32 (3%)	88 (8%)	793 (71%)	85 (8%)	47 (4%)
Gr7	1306	197 (15%)	495 (38%)	661 (51%)	645 (49%)	20 (2%)	60 (5%)	37 (3%)	114 (9%)	894 (69%)	92 (7%)	89 (7%)
Gr8	1359	186 (14%)	479 (35%)	698 (51%)	661 (49%)	22 (2%)	72 (5%)	34 (3%)	86 (6%)	973 (72%)	106 (8%)	66 (5%)

Note. Data not available for grades 1 and 2 on students of economic disadvantage or students of multi-ethnicity. Raw numbers reported; percentages in parentheses are rounded to the nearest whole percent, meaning some demographics sum to more than 100%. Further, because some students failed to respond, not all gender percentages sum to 100%.
GR = Grade level
Sped = Special education placement
Econ. Dis = Economically disadvantaged – students eligible for free or reduced lunch
Amer-Ind/AK-Native = American Indian or Alaskan Native
Asian/Pac-Islander = Asian or Pacific Islander

Table 2
Item Descriptive Statistics by Grade-Level.

Grade 1	Count	Grand mean	Min	Max	Variance
Number & operations					
Item means	205	.58	.18	.84	.03
Item variances	205	.21	.13	.25	.00
Geometry					
Item means	207	.71	.28	.97	.05
Item variances	207	.16	.03	.25	.01
Number & operations and algebra					
Item means	206	.41	.21	.79	.02
Item variances	206	.22	.17	.25	.00
Grade 2					
Number & operations					
Item means	70	.55	.26	.90	.04
Item variances	70	.22	.09	.25	.00
Geometry					
Item means	80	.62	.21	.85	.04
Item variances	80	.21	.13	.25	.00
Number & operations and algebra					
Item means	16	.53	.13	.94	.04
Item variances	16	.22	.06	.27	.00
Grade 3					
Number & operations					
Item means	1222	.59	.14	.97	.06
Item variances	1222	.19	.03	.25	.00
Geometry					
Item means	1231	.78	.35	.99	.05
Item variances	1231	.13	.01	.25	.01
Number & operations and algebra					
Item means	1224	.60	.36	.96	.03
Item variances	1224	.21	.04	.25	.00
Grade 4					
Number & operations					
Item means	1206	.66	.25	.96	.03
Item variances	1206	.19	.03	.25	.01
Measurement					
Item means	1211	.78	.13	.99	.06
Item variances	1211	.12	.00	.25	.01
Number & operations and algebra					
Item means	1205	.68	.33	.90	.03
Item variances	1205	.19	.09	.25	.00
Grade 5					
Number & operations					
Item means	1269	.69	.28	.95	.05
Item variances	1269	.17	.05	.25	.01
Geometry, measurement, & algebra					
Item means	1270	.74	.24	.97	.04
Item variances	1270	.16	.03	.25	.01
Number & operations and algebra					
Item means	1270	.72	.55	.90	.01
Item variances	1270	.19	.09	.25	.00
Grade 6					
Number & operations					
Item means	1245	.58	.38	.80	.02
Item variances	1245	.23	.16	.25	.00
Algebra					
Item means	1249	.74	.34	.98	.05
Item variances	1249	.15	.02	.25	.01
Number & operations and algebra					
Item means	1238	.76	.54	.99	.03
Item variances	1238	.16	.01	.25	.01
Grade 7					
Number & operations, algebra, & geometry					
Item means	712	.71	.49	.88	.02
Item variances	712	.19	.11	.25	.00
Measurement, geometry, & algebra					
Item means	720	.46	.15	.79	.04
Item variances	720	.22	.13	.25	.00
Number & operations and algebra					
Item means	707	.57	.20	.93	.04
Item variances	707	.21	.07	.25	.00
Grade 8					
Algebra					
Item means	723	.48	.24	.84	.03
Item variances	723	.22	.13	.25	.00
Geometry & measurement					
Item means	715	.56	.34	.92	.03
Item variances	715	.22	.08	.25	.00
Data analysis, number & operations, & algebra					
Item means	717	.67	.42	.96	.03
Item variances	717	.19	.04	.25	.01

Table 3
Inter-Item Correlations.

Grade 1	Count	Mean	Min	Max	Cronbach's alpha
Number & operations	16	.12	-.11	.37	.69
Geometry	16	.01	-.12	.38	.64
Number & operations and algebra	16	.10	-.06	.41	.64
Total	48	.08	-.15	.41	.82
Grade 2					
Number & operations	16	.08	-.17	.41	.58
Geometry	16	.11	-.13	.51	.67
Number & operations and algebra	16	.09	-.52	.71	.61
Total	48	.12	-.72	.86	.86
Grade 3					
Number & operations	16	.09	-.06	.33	.60
Geometry	16	.07	-.05	.35	.56
Number & operations and algebra	16	.12	-.02	.68	.70
Total	48	.08	-.10	.69	.80
Grade 4					
Number & operations	16	.14	-.09	.62	.72
Measurement	16	.09	-.06	.70	.61
Number & operations and algebra	16	.13	-.02	.36	.70
Total	48	.11	-.08	.70	.86
Grade 5					
Number & operations	16	.14	-.03	.35	.72
Geometry, measurement, & algebra	16	.07	-.08	.65	.55
Number & operations and algebra	16	.17	-.02	.38	.76
Total	48	.11	-.07	.66	.85
Grade 6					
Number & operations	16	.11	-.03	.27	.66
Algebra	16	.17	-.01	.50	.77
Number & operations and algebra	16	.14	.01	.43	.71
Total	48	.12	-.07	.50	.87
Grade 7					
Number & operations	16	.22	.08	.42	.82
Geometry	16	.07	-.08	.24	.54
Number & operations and algebra	16	.13	-.12	.52	.70
Total	48	.11	-.12	.52	.86
Grade 8					
Number & operations	16	.07	-.07	.34	.56
Geometry	16	.10	-.03	.27	.65
Number & operations and algebra	16	.14	-.01	.37	.73
Total	48	.09	-.09	.37	.83

Note. Cronbach's alpha scores based on standardized item.

Table 4
Overall Statistics.

Grade	Cronbach's Alpha	SD	SEM
1	0.82	6.83	2.90
2	0.86	8.53	3.19
3	0.80	6.29	2.81
4	0.86	7.13	2.67
5	0.85	7.04	2.73
6	0.87	7.34	2.65
7	0.86	7.88	2.95
8	0.83	7.40	3.05

References

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

Feldt, L. S, & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: American Council on Education/Macmillan.

Ketterlin-Geller, L.R., Alonzo, J., Braun-Monegan, J., & Tindal, G. (2007). Recommendations for accommodations: Implications of (in)consistency. *Remedial and Special Education, 28*(4), 194-206.