Developmental Scale for North Carolina End-of-Grade/End-of-Course Mathematics/Math I Tests, Fourth Edition

Alan Nicewander, Ph.D.
Tia Sukin, Ed.D.
Josh Goodman, Ph.D.
Huey Dodson, B.S.
Matthew Schulz, Ph.D.
Susan Lottridge, Ph.D.
Phoebe Winter, Ph.D.

Submitted to the
North Carolina Department of Education

December 9, 2013

Pacific Metrics Corporation
1 Lower Ragsdale Drive
Building 1, Suite 150
Monterey, CA 93940

Developmental Scale for North Carolina End-of-Grade/End-of-Course Mathematics/Math I Tests, Fourth Edition

This technical report describes the methods used and results found by Pacific Metrics Corporation in deriving a developmental scale for the North Carolina End-of-Grade/End-of-Course Mathematics/Math I Tests, Fourth Edition. Each time a new edition of a test series is introduced, the North Carolina Department of Public Instruction (NCDPI) analyzes the data for the purpose of creating a developmental, or vertical, scale for that edition. The developmental scale is linked to the scale used for previous editions of the tests so that consistent interpretations of the data can be made across time.

For the Fourth Edition, the tests included in the Mathematics developmental scale were the End-of-Grade tests in grades 3 through 8 and the End-of-Course test in Math I. The addition of Math I to the scale is new with this edition; prior editions included grades 3 through 8 only and also included Pre 3. To create the vertical scale, Pacific Metrics used methods previously employed by NCDPI as described in the *North Carolina Mathematics Tests Technical Report*, chapter 4 (Bazemore, Kramer, Yelton, & Brown, 2006). For the Mathematics/Math I scale, Pacific Metrics used Appendix C (Thissen, Sathy, Edwards, & Flora, 2006) of that report. The article by Williams, Pommerich, and Thissen (1998) was also used as a reference.

After reviewing the results of the scaling effort, Pacific Metrics and NCDPI determined that the data from school year 2012-2013 did not support the use of a developmental scale, and NCDPI did not adopt it. In 2013, the tests covered a number of new content standards and changed the grade levels when some content is expected to be taught. One plausible reason for the behavior of the data is that curricular and instructional practice was still adjusting to the new Mathematics standards so that they were not yet taught in the same vertical manner as they were tested. Pacific Metrics recommended that NCDPI revisit the incorporation of a developmental scale for Mathematics after the standards have been in place long enough for instruction and learning to be structured accordingly.

Developmental Scale Derived for the Fourth Edition (Not Adopted by NCDPI)

To derive the developmental scale, the same items (called a linking set) were administered to students in adjacent grades. Below-grade links were used for the Mathematics/Math I scale. Thus, the linking-set items were operational (i.e., items contributed to student scores) in the lower grade but not in the upper grade. Each grade pair for grades 3 through 8 had nine below-grade linking sets; the grade-pair 8–Math I had sixteen below-grade linking sets. The linking sets contained either eight or ten items. Grade 5 was the base grade for the developmental scale, using a mean of 450 and standard deviation of 10.

Table 1 presents the resulting, but not adopted, Fourth Edition developmental scale for the population for Mathematics/Math I. The table shows an unexpected growth pattern in that mean score increases very little as grade level increases, except for the 4-point scale score gain between grade 3 and grade 4. The smallest growth occurred between grade 8 and Math I, where negative gain (–0.53) is observed.

Table 1. Developmental Scale Means and Standard Deviations
Derived from Spring 2013 Item Calibration for
North Carolina End-of-Grade/End-of-Course Tests of
Mathematics/Math I, Fourth Edition

	Population	
Grade	Mean	Standard Deviation
3	443.71	10.65
4	448.03	9.43
5	450.00	10.00
6	450.44	10.95
7	452.51	11.45
8	452.81	9.80
Math I	452.28	9.65

The values for the developmental scale are based upon item response theory (IRT) estimates of differences between adjacent-grade mean thetas (θ) and ratios of adjacent-grade standard deviations of θ. The three-parameter logistic model was used to estimate item and person parameters. flexMIRT^{TM} version 1.88 (Cai, 2012) was used and calibrations were obtained for each of the grade-pair linking sets. In flexMIRT^{TM}, the below grade was considered the reference group; its population mean and standard deviation were set to 0 and 1, respectively. The above-grade mean and standard deviation were estimated using the scored data and the IRT parameter estimates. These parameters were provided in the flexMIRT^{TM} output and did not require independent calculation.

Under the assumption of equivalent groups, the form results were averaged within grade pairs to produce one set of values per adjacent grade. Outlying values were dropped if they were greater than two standard deviations from the mean. Three sets of values were dropped as outliers—one each from the 3–4, 5–6, and 8–Math I grade pairs. Table 2 displays the average difference in adjacent-grade means and standard deviation ratios for Mathematics/Math I. Note that the values in table 2 are on the theta scale, while the –0.53 difference noted above is on the reporting scale. Table 3 presents the mean difference and standard deviation ratio for each adjacent-grade link.

Table 2. Average Mean Difference in Standard Deviation Units of
Lower Grade and Average Standard Deviation Ratios Derived from
Spring 2013 Item Calibrations for North Carolina
End-of-Grade/End-of-Course Tests of Mathematics/Math I, Fourth Edition

Grades	Average Mean Difference	Average Standard Deviation Ratio	Number of Grade-Pair Forms
3–4*	0.406	0.886	8
4–5	0.209	1.061	9
5–6*	0.044	1.095	8
6–7	0.189	1.046	9
7–8	0.027	0.856	9

8–Math I*	–0.055	0.985	15

Note: An asterisk (*) denotes that one outlier was removed from the average for this grade pair.

Table 3. Values for Adjacent-grade Means in Standard Deviation (SD) Units of Lower Grade and Standard Deviation Ratios, Derived from Spring 2013 Item Calibrations for North Carolina End-of-Grade/End-of-Course Tests of Mathematics/Math I, Fourth Edition

Grades 3–4		Grades 4–5		Grades 5–6		Grades 6–7		Grades 7–8		Grade 8–Math I	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.428	0.844	0.271	0.977	0.145	1.000	0.171	1.221	0.022	0.953	–0.107	1.096
0.429	0.957	0.226	1.043	0.017	1.260	0.295	0.947	0.062	0.899	–0.045	1.043
0.273	1.075	0.182	1.089	–0.073	1.230	0.229	0.946	–0.042	0.896	–0.181	1.088
0.371	0.801	0.126	0.980	0.021	1.006	0.128	1.208	0.099	0.996	–0.233	0.983
0.409	0.942	0.159	1.120	–0.319	1.195	0.303	0.887	–0.120	0.953	–0.012	1.016
0.398	0.798	0.213	0.970	–0.025	1.152	0.168	0.996	0.117	0.722	0.013	1.035
0.411	0.931	0.322	1.173	0.055	0.924	0.086	0.971	–0.021	0.843	0.068	0.862
0.406	0.800	0.023	1.198	0.108	1.064	0.193	1.076	0.169	0.683	–0.091	0.871
0.397	1.011	0.356	0.996	0.108	1.119	0.127	1.161	–0.046	0.759	–0.087	1.011
										–0.032	0.876
										–0.022	1.022
										–0.102	1.132
										0.072	0.912
										–0.041	0.896
										–0.152	1.017
										–0.101	0.894

Note: Means and standard deviations in shaded cells were dropped from analyses as outliers.

Comparison of Fourth Edition Developmental Scales to First through Third Edition Scales

Table 4 presents the mean scale scores by grade for the First, Second, Third, and Fourth editions for Mathematics/Math I. To facilitate comparison of the growth between grades among the First, Second, Third, and Fourth editions, figure 1 presents the mean scores plotted together for Mathematics/Math I. To place the First, Second, Third, and Fourth edition scores on similar scales, a value of 300 was added to the First Edition scores, a value of 200 was added to the Second Edition scores, and a value of 100 was added to the Third Edition scores.

For Mathematics, greater average growth between grades 3–8 occurred in the First, Second, and Third editions (31.8, 18.82, and 15.98, respectively) than in the Fourth Edition (9.10). As shown in figure 1, while growth of the mean scores of the Second and Third editions are similar, First and Fourth edition mean scores are quite different. The most notable difference between the Fourth Edition and previous editions is the lower growth rate between grades 3–8.

Based on the data analyzed, the lack of increasing means across grades calls into question the validity of a developmental scale for the Fourth Edition. Discussions with North Carolina psychometric and content staff revealed that in transition to testing Common Core standards, it is entirely possible that teachers had not yet incorporated tested standards into classroom instruction, thus making such test items appear more difficult than they might otherwise be. Reports also revealed that the grades at which certain concepts in mathematics were taught were changed significantly enough to disrupt proper scaling of a developmental scale. It was recommended that horizontal scaling within each grade be maintained and a developmental scale be considered in the future once curriculum and assessments are properly aligned.

Table 4. Comparison of Population Means and Standard Deviations for First through Fourth Editions of North Carolina End-of-Grade/End-of-Course Tests of Mathematics/Math I

	First Edition (1992)		Second Edition (2002)		Third Edition (2008)		Fourth Edition (2013)	
Grade	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Pre 3	131.60	7.80	234.35	9.66	326.98	12.69		
3	143.50	11.10	248.27	9.86	339.44	10.97	443.71	10.65
4	152.90	10.10	252.90	10.65	345.26	10.24	448.03	9.43
5	159.50	10.10	255.99	12.78	350.00	10.00	450.00	10.00
6	165.10	11.20	259.95	11.75	351.45	10.41	450.44	10.95
7	171.00	11.50	263.36	12.46	353.66	10.15	452.51	11.45
8	175.30	11.90	267.09	12.83	355.42	9.99	452.81	9.80
Math I							452.28	9.65

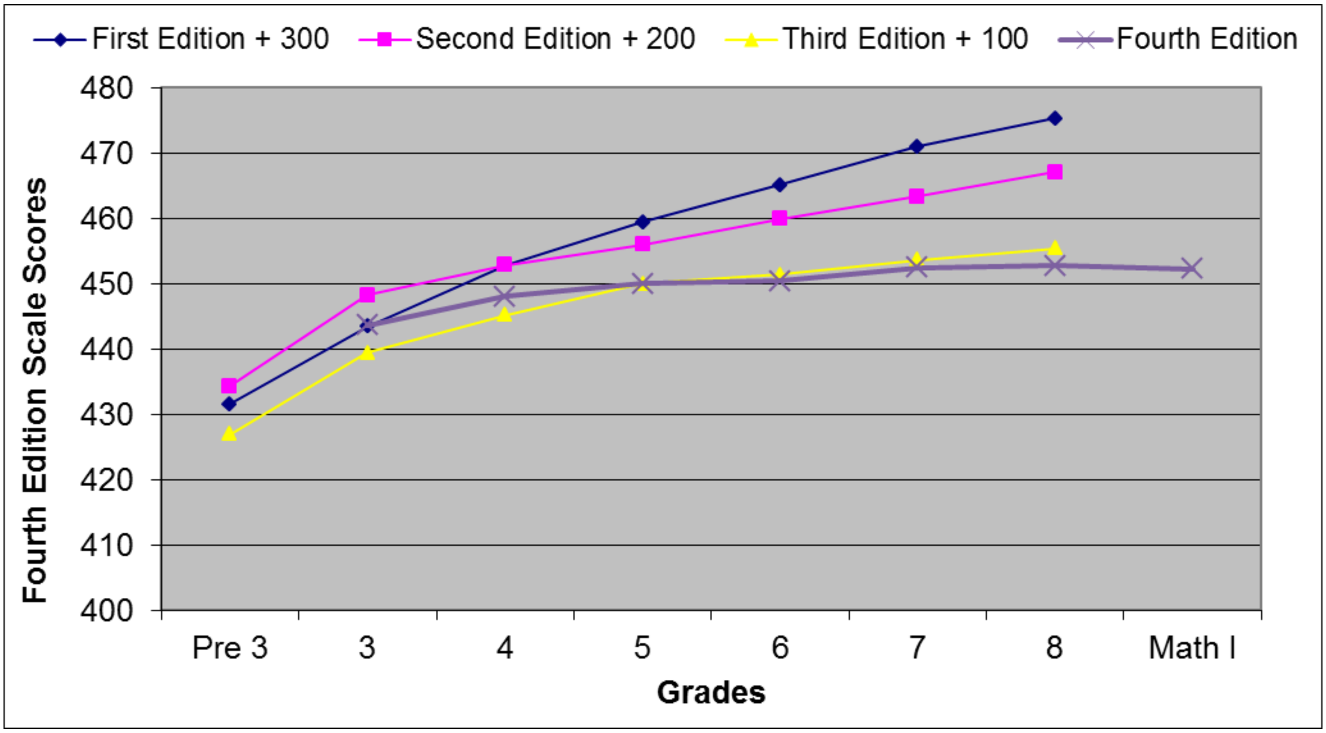

Figure 1. Comparison of Growth Curves between First, Second, Third, and Fourth Editions of North Carolina End-of-Grade/End-of-Course Tests of Mathematics/Math I.

Quality Assurance Procedures

Pacific Metrics applied a variety of analyses and procedures to ensure that the scaling and linking studies were performed without error and that all key results are accurate and not artifacts of methodology. For the developmental scale, the mean difference and standard deviation ratios for the grades and forms were compared to the classical test theory *p*-values of the linking items. The data provided evidence that the mean difference and standard deviation ratios were accurate in both direction and magnitude (see table 5). Also, previous work using the described statistical method to create the vertical scale was applied to the Second Edition data to ensure that it reproduced the scale correctly.

Table 5. Average Mean Difference in Standard Deviation Units of Lower Grade and Standard Deviation Ratios, and Average Difference in *p*-values (Higher Minus Lower Grade) of Linking Sets, for North Carolina End-of-Grade/End-of-Course Tests of Mathematics/Math I, Fourth Edition

Grade Pair	Average Mean Difference	Mean *p*-value Difference for Linking Items
3–4*	0.406	0.084
4–5	0.209	0.054
5–6*	0.044	0.014
6–7	0.189	0.045
7–8	0.027	0.010
8–Math I*	–0.055	–0.008

Note: An asterisk (*) denotes that one grade-pair link was dropped from analyses as an outlier.

Additionally, IRT parameters provided separately by the North Carolina Department of Education were correlated with the flexMIRT^{TM} calibrated item parameters within grade pairs and averaged across grades. For Mathematics, the average correlation for discrimination parameters was 0.97 with a standard deviation of 0.02 across grade and form pairs. The average correlation for difficulty was 0.97 with a standard deviation of 0.02. The average correlation for guessing parameters was 0.97 with a standard deviation of 0.01.

Psychometrics Underlying the Developmental Scale

The procedure for creating the developmental scale is based upon that described in Williams, Pommerich, and Thissen (1998). The procedure is divided into four steps, described below.

Step 1. flexMIRT^{TM} was used to calibrate the End-of-Grade and End-of-Course Mathematics tests' item and population parameters for adjacent grades. This process was described in the section entitled "Developmental Scale Derived for the Fourth Edition (Not Adopted by NCDPI)" of this report and resulted in average mean difference and average standard deviation ratios (m_n and s_n) for each grade n (see table 2).

Step 2. A (0,1) growth scale anchored at grade 3 was constructed to yield the following means (M_n) and standard deviations (S_n):

$M_n = M_{n-1} + m_n S_{n-1}$, mean for grade *n* on (0,1) growth scale anchored at the lowest grade (with grade 3 indexed as *n*=3),

$S_n = s_n S_{n-1}$, standard deviation for grade *n* on (0,1) growth scale anchored at the lowest grade (with grade 3 indexed as *n*=3),

where $M_2 \equiv 0$, and $S_2 \equiv 1$. This (0,1) growth scale was generated recursively upwards to the End-of-Course (Math I).

Step 3. The scale was re-centered (re-anchored) at grade 5, yielding

$$M_n^* = \frac{(M_n - M_5)}{S_5}$$

$$S_n^* = \frac{S_n}{S_5}$$

as the means ($M^*{}_n$) and standard deviations ($S^*{}_n$).

Step 4. The final step in constructing the growth scale was the application of a linear transformation in order to produce a growth scale with the grade 5 mean and standard deviations equal to 450 and 10, respectively, *viz.*,

$$\mu_n = 450 + 10 M_n^*$$

$$\sigma_n = 10 S_n^* \ ,$$

where μ_n is the mean of the final growth scale in grade *n* and σ_n is the standard deviation for the growth scale in grade *n*.

References

Bazemore, M., & Van Dyke, P. (2004). *North Carolina Reading Comprehension Tests Technical Report.* Raleigh, NC: North Carolina Department of Public Instruction.

Cai, L. (2012). flexMIRT^{TM} version 1.88: A numerical engine for multilevel item factor analysis and test scoring. [Computer software]. Seattle, WA: Vector Psychometric Group.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating methods and practices.* New York: Springer.

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Stocking, M.L., & Lord, F.M. (1983). Development of a common metric in Item Response Theory. *Applied Psychological Measurement, 7*(2), 201-210.

Thissen, D., Edwards, M., Coon, C. & Woods, C. (2002). *North Carolina Reading Comprehension Tests Technical Report, Appendix C.* Raleigh, NC: North Carolina Department of Public Instruction.

Williams, V.S.L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based upon Thurstone methods and item response theory. *Journal of Educational Measurement, 35,* 93-107.