North Carolina End-of-Grade/End-of-Course Science Tests: Edition Concordances

Alan Nicewander, Ph.D.
Josh Goodman, Ph.D.
Tia Sukin, Ed.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

November 27, 2013

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

NORTH CAROLINA END-OF-GRADE/END-OF-COURSE SCIENCE TESTS: EDITION CONCORDANCES

This technical report describes the results and methods used by Pacific Metrics Corporation to create concordances between the First and Second editions of North Carolina's End-of-Grade (EOG) Science Tests for grades 5 and 8 and the End-of-Course (EOC) Biology test. Concordance tables for each test were generated using the Stocking-Lord (Stocking & Lord, 1983) scaling and item response theory true-score equating methods (Kolen & Brennan, 2006). Strictly speaking, the term equating should only be used when the two tests that are to be linked are parallel in content (Mislevy, 1992). Presumably, the newer tests assess slightly different constructs due to curriculum changes implemented by the state. While equating methods were employed in completing these analyses, this report will refer to results as "linking" or "concordances" to underscore that the relationships established between editions do not meet the criteria to be considered equating.

CONCORDANCES BETWEEN EDITIONS

Figure 1 displays the linking functions between new and old edition scale for all three Science tests (grades 5 and 8 EOGs; Biology EOC). All three functions are collinear. Given the lack of a developmental scale for either edition, these results conform to expectations. Table 1 presents the final concordance tables for all three Science assessments.

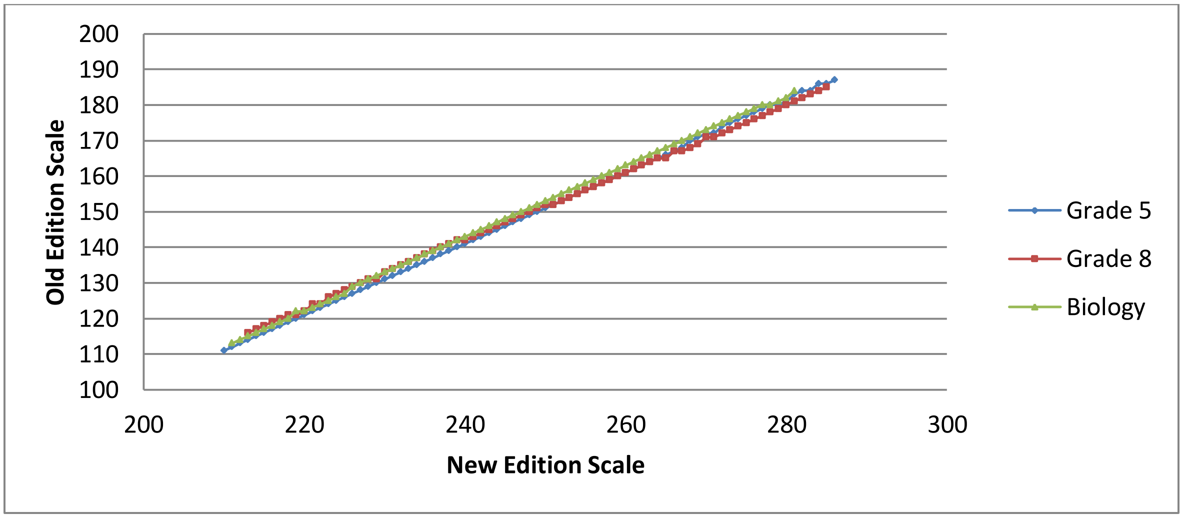

Figure 1. Linking Functions between the New and Old Editions of the North Carolina Grades 5 and 8 EOG Tests and Biology EOC Test.

Table 1. Concordance Tables for
North Carolina EOG/EOC Science Tests

New Edition	Old Edition		
	Grade 5	Grade 8	Biology
210	111	.	.
211	112	.	113
212	113	.	114
213	114	116	115
214	115	117	116
215	116	118	117
216	117	119	118
217	118	120	119
218	119	121	120
219	120	121	122
220	121	122	122
221	122	124	123
222	123	124	124
223	124	126	125
224	125	127	126
225	126	128	127
226	127	129	129
227	128	130	130
228	129	131	131
229	130	131	132
230	131	133	133
231	132	134	134
232	133	135	135
233	134	136	136
234	135	137	137
235	136	138	138
236	137	139	139
237	138	140	140
238	139	141	141
239	140	142	142
240	141	142	143
241	142	143	144
242	143	144	145
243	144	145	146
244	145	146	147
245	146	147	148
246	147	148	149
247	148	149	150

New Edition	Old Edition		
	Grade 5	Grade 8	Biology
248	149	150	151
249	150	151	152
250	151	152	153
251	152	152	154
252	153	153	155
253	154	154	156
254	155	155	157
255	156	156	158
256	157	157	159
257	158	158	160
258	159	159	161
259	160	160	162
260	161	161	163
261	162	162	164
262	163	163	165
263	164	164	166
264	165	165	167
265	166	165	168
266	167	167	169
267	168	167	170
268	170	168	171
269	171	169	172
270	172	171	173
271	172	171	174
272	174	172	175
273	175	173	176
274	176	174	177
275	177	175	178
276	178	176	179
277	179	177	180
278	180	178	180
279	180	179	181
280	181	180	182
281	183	181	184
282	184	182	.
283	184	183	.
284	186	184	.
285	186	185	.
286	187	.	.

Psychometrics Underlying the Linking Process

The linking process employed a common item, non-equivalent groups equating design. In this design, a set of items from the previous edition was embedded within each new edition form. After the new edition forms were calibrated, the common items had item parameter values on both the new and old edition scales. In each of the Science tests, the three new edition forms were administered in both paper and online formats, with three operational forms (A, B, and C for paper; M, N, and O for online) associated with each mode of delivery. The operational forms could be considered paired (A/M, B/N, and C/O) across modes where each form-pair contained the same operational and concordance linking items. However, the form-pairs were calibrated in a manner that allowed corresponding items that performed differently across modes to have unique parameter values. As a result of the number of and magnitude of the differences in item parameters across item-pairs, Pacific Metrics completed a separate concordance table for each mode. For all three tests, the end results of the two concordances were similar and the two concordances were aggregated to form a single concordance.

All item parameters used in the linking process were provided by North Carolina Department of Public Instruction (NCDPI). Using the linking-item parameters calibrated to each edition's scale, Stocking-Lord scaling constants were estimated with a program developed in the R statistical programming language (R Development Core Team, 2012). Scaling constants were estimated in two ways: 1) for each separate form within each grade level, and 2) for the entire set of linking items across all forms. Given that there were enough linking items, the form-by-form method of scaling was preferred as it dispensed with the assumption that each form was administered to an equivalent group. However, the scaling constants that were produced from using the entire set of linking items aided in quality assurance and, more importantly, provided an alternative scaling method should a large number of linking items be dropped from a single form or should a single form display a problematic scaling relationship. Table 2 presents the scaling constants for each test. The new edition operational item parameters for each form were rescaled to the old edition bank scale by applying the appropriate set of form-by-form Stocking-Lord scaling constants.

Table 2. Stocking-Lord Scaling Constants

Test	Form A/M		Form B/N		Form C/O		All Forms	
	A	B	A	B	A	B	A	B
5 (Paper)	0.992	0.290	1.070	0.121	0.945	0.022	1.003	0.141
5 (Online)	1.011	0.325	1.070	0.121	0.952	–0.014	1.010	0.138
8 (Paper)	0.941	0.186	0.888	0.188	0.953	0.137	0.926	0.169
8 (Online)	0.903	0.243	0.881	0.184	0.942	0.153	0.908	0.192
Biology (Paper)	0.986	0.251	0.960	0.578	1.014	0.230	0.998	0.353
Biology (Online)	0.985	0.214	0.976	0.566	1.012	0.211	1.002	0.323

Before estimating scaling constants, the linking items were screened for stability using a Delta plot (Holland & Thayer, 1985) method. This process assumed that the difficulty of the linking items, if they were stable, would be ordered the same across the two editions despite being administered to two different populations. Thus, instability was defined as significant differences in the relative difficulty of any linking item across editions. Item difficulties were transformed to the Delta scale and plotted. Items falling more than two standard errors away from the plotted principal axis were flagged as unstable. The entire set of linking items was screened in a single application of the Delta method. A count of items dropped due to instability is presented in table 3.

Table 3. Number of Linking Items and Number of Items Flagged as Unstable

Test	Form A/M		Form B/N		Form C/O		All Forms	
	Total	Dropped	Total	Dropped	Total	Dropped	Total	Dropped
5 (Paper)	15	1	15	0	15	0	45	1
5 (Online)	15	1	15	0	15	0	45	1
8 (Paper)	30	4	30	1	30	0	90	5
8 (Online)	30	3	30	0	30	0	90	3
Biology (Paper)	29	1	30	2	30	0	89	3
Biology (Online)	29	1	30	4	30	0	89	5

Using the scale means and standard deviation for each test (μ=250 and σ=10) and the new edition operational item parameters, an expected *a posteriori* (EAP) score and corresponding new edition scale score were created for each possible sum-score. The same process was repeated using the new edition item parameters rescaled to the old edition scale (using the constants in table 2) and the old edition scale means and variances for each test (μ=150 and σ=10). The concordance tables were created by merging the two sets of scale scores, thinning the table such that each new edition scale score appeared only once, and using linear interpolation to ensure that the entire range of new edition scale score values was represented. The cut scores defining the boundaries of the four achievement level categories on the old edition tests were applied to the new edition scores using the concordance tables (table 1). These ranges appear in table 4.

Table 4. Cut Scores for New and Old Editions of the
North Carolina EOG/EOC Tests of Science and Biology

Test	Level	First Edition	Second Edition
5	I	≤145	≤244
	II	146–152	245–251
	III	153–160	252–259
	IV	≥161	≥260
8	I	≤348	≤240
	II	349–356	241–247
	III	357–367	248–256
	IV	≥368	≥257
	Level	Second Edition	Third Edition
Biology	I	≤137	≤234
	II	138–146	235–243
	III	147–158	244–255
	IV	≥159	≥256

QUALITY ASSURANCE PROCEDURES

In the construction of the concordance tables, Pacific Metrics applied a variety of analyses and procedures to ensure reasonable and accurate results. At each step in the linking procedure where item parameters were used, the values used as inputs were checked against the values supplied by NCDPI. Stocking-Lord scaling constants were computed using two different methods. All of the scaling constants resulting from the two different methods were expected to be consistent; this consistency served as a check on the reasonableness of the estimated constants and enabled any aberrant values to be removed prior to rescaling. Additionally, Test Characteristic Curves (TCCs) for the new and old edition linking items were compared for similarity after rescaling. A successful scaling results in TCCs that overlap significantly. For all tests, scaling was deemed reasonable and accurate.

In the production of the final concordance tables, it was essential to create EAP and scale score estimates in the same manner as the operational scoring tables created by NCDPI. To ensure that the methods used by Pacific Metrics were congruent with NCDPI's process, the operational scoring tables for each form were recreated and compared to the scoring tables of record created by NCDPI. In all cases, the two sets of scoring tables matched.

For each test, the final concordance was compared to the separate concordances based on each of the forms. The final concordance between editions, which was based on all operational items, was expected to be similar to concordances constructed using the operational items from a single form. At each grade level, the concordance functions were similar, suggesting that the final results were reasonable.

References

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Rep. No. 85–43). Princeton, NJ: Educational Testing Service.

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.

R Development Core Team. (2012). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Stocking, M. L., & Lord, F. M. (1983). Development of a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.