North Carolina End-of-Grade ELA/Reading Tests: Third and Fourth 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 ELA/READING TESTS: THIRD AND FOURTH EDITION CONCORDANCES

This technical report describes the results and methods used by Pacific Metrics Corporation to create concordances between the Third and Fourth editions of North Carolina's End-of-Grade (EOG) ELA/Reading Comprehension Tests. 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 the Third and Fourth edition scales. There are six functions—one for each grade level. The functions are nearly collinear, with the grade 3 function showing a slightly greater slope than the other grades. The close proximity of the lines in figure 1 for the different grades and the ranges of scores within each grade suggest that the concordances generally conform to expectations and are consistent with the structure of the development scale. This result differs slightly from the Second to Third edition linking functions, in which the slopes generally increased as grades increased. These differences are likely due to the use of a different equating design for concordance table creation (a two-step, chained Stocking-Lord) and differences in the structure of the Third and Fourth editions of the developmental scale. Table 1 presents the final concordance tables between the Third Edition scale and the Fourth Edition scale for each EOG test in grades 3 through 8.

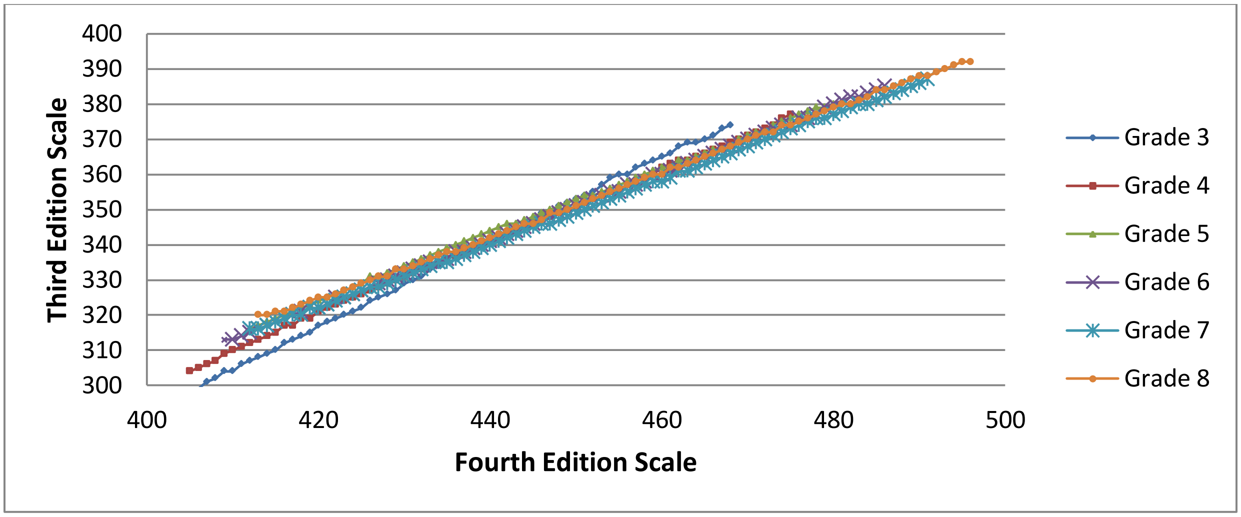

Figure 1. Linking Functions between the Third and Fourth Editions of the North Carolina EOG Tests of ELA/Reading Comprehension.

Table 1. Concordance Tables for Fourth Edition Scale Scores
to Third Edition Scale Scores

Fourth Edition Scale	Third Edition Scale					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
399	291					
400	292					
401	294					
402	295					
403	296					
404	297					
405	299	304				
406	299	305				
407	301	306				
408	302	307				
409	304	309		313		
410	304	310		313		
411	306	311		314		
412	307	312	316	315	316	
413	308	313	317	316	316	320
414	309	314	318	317	317	320
415	310	315	319	318	318	321
416	312	317	320	319	319	321
417	313	317	321	320	320	322
418	314	319	322	321	320	323
419	315	319	323	322	322	324
420	317	321	324	322	322	325
421	318	322	325	323	323	325
422	319	323	326	325	324	326
423	320	324	327	325	325	327
424	321	325	328	326	326	328
425	322	326	329	327	327	329
426	324	327	331	328	328	330
427	325	328	331	329	328	331
428	326	329	332	330	329	331
429	327	330	333	331	330	333
430	329	331	334	332	331	333
431	330	332	335	333	332	334
432	331	333	336	334	333	335
433	333	334	337	335	334	336
434	334	335	338	335	335	337

Fourth Edition Scale	Third Edition Scale					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
435	335	336	339	336	335	338
436	336	337	340	338	336	338
437	337	338	341	338	337	339
438	339	339	342	339	338	340
439	340	340	343	340	339	341
440	341	341	344	341	340	342
441	342	342	345	342	341	343
442	343	343	346	343	342	344
443	345	344	346	344	343	345
444	346	345	347	345	344	346
445	347	346	348	346	345	346
446	349	347	349	347	346	347
447	349	348	350	348	346	349
448	351	349	351	349	347	349
449	352	350	352	350	348	350
450	353	351	353	351	349	351
451	354	352	354	352	350	352
452	355	353	354	353	351	353
453	357	354	355	354	352	354
454	359	355	356	355	353	355
455	360	356	357	355	354	356
456	360	357	358	357	355	357
457	362	358	359	357	356	358
458	363	359	360	358	357	359
459	364	360	361	360	358	360
460	365	362	362	360	358	360
461	366	363	362	361	359	362
462	368	364	364	362	361	362
463	369	364	364	363	361	363
464	369	365	365	364	362	364
465	370	366	366	365	363	365
466	371	367	367	366	364	366
467	373	368	367	367	365	367
468	374	369	368	368	366	368
469		370	370	369	367	369
470		371	371	370	368	370
471		372	372	371	369	371
472		373	372	372	370	372

Fourth Edition Scale	Third Edition Scale					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
473		374	374	373	371	372
474		376	375	374	372	374
475		377	376	374	373	374
476			377	376	374	375
477			378	376	375	376
478			379	377	376	377
479			379	379	376	378
480			380	380	377	379
481				381	378	380
482				382	379	380
483				382	380	381
484				383	380	382
485				384	381	384
486				385	382	384
487					383	385
488					384	386
489					385	387
490					386	388
491					387	388
492						389
493						390
494						391
495						392
496						392

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. Each EOG test contained three paper-based forms (A, B, and C).

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 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 Fourth Edition operational item parameters for each form were rescaled to the Third Edition bank scale by applying the appropriate set of form-by-form Stocking-Lord scaling constants.

Table 2. Stocking-Lord Scaling Constants

Grade	Form A		Form B		Form C		All Forms	
	A	B	A	B	A	B	A	B
3	1.001	0.456	1.084	0.141	1.064	0.026	1.088	0.249
4	0.800	0.294	1.006	0.138	1.040	0.166	0.942	0.202
5	0.758	0.421	0.929	0.307	0.943	0.127	0.877	0.281
6	0.949	0.101	1.035	0.033	1.133	0.100	1.022	0.073
7	0.622	0.700	1.028	–0.072	1.117	–0.061	1.056	–0.068
8	0.991	0.140	0.800	0.370	1.328	–0.193	1.124	0.125

Notes: The constants in shaded cells (grade 7, form A) were dropped as outliers in the analyses. The "All Forms" constants are based only on linking items from forms B and C.

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

Grade	Form A		Form B		Form C		All Forms	
	Total	Dropped	Total	Dropped	Total	Dropped	Total	Dropped
3	16	1	13	0	14	0	43	1
4	16	1	15	0	15	0	46	1
5	15	2	15	0	15	1	45	3
6	15	0	14	0	13	3	42	3
7	15	3	15	0	15	1	44	4
8	10	0	13	2	14	0	37	2

Note: The values in shaded cells (grade 7, form A) were flagged as unstable. However, as noted in table 2 above, no items from form A were used in the analyses.

Using the Fourth Edition developmental scale means and standard deviation for each grade (see Nicewander et al., 2013) and the Fourth Edition operational item parameters, an expected *a posteriori* (EAP) score and corresponding Fourth Edition scale score were created for each possible sum-score. The same process was repeated using the Fourth Edition item parameters rescaled to the Third Edition scale (using the constants in table 2) and the Third Edition developmental scale means and variances for each grade level. The concordance tables were created by merging the two sets of scale scores, thinning the table such that each Fourth Edition scale score appeared only once, and using linear interpolation to ensure that the entire range of Fourth Edition scale score values was represented. The cut scores defining the boundaries of the four achievement level categories on the Third Edition tests were applied to the Fourth Edition scores using the concordance tables (table 1). These ranges appear in table 4.

Table 4. Cut Scores for Third and Fourth Editions of the North Carolina EOG Tests of ELA/Reading Comprehension

	Level	Third Edition	Fourth Edition
3	I	≤330	≤431
	II	331–337	432–437
	III	338–349	438–447
	IV	≥350	≥448
4	I	≤334	≤433
	II	335–342	434–441
	III	343–353	442–452
	IV	≥354	≥453
5	I	≤340	≤436
	II	341–348	437–445
	III	349–360	446–458
	IV	≥361	≥459
6	I	≤344	≤443
	II	345–350	444–449
	III	351–361	450–461
	IV	≥362	≥462
7	I	≤347	≤448
	II	348–355	449–456
	III	356–362	447–464
	IV	≥363	≥465
8	I	≤349	≤448
	II	350–357	449–456
	III	358–369	457–469
	IV	≥370	≥470

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.

Nicewander, A., Sukin, T., Goodman, J., Dodson, H., Schulz, M., Lottridge, S., & Winter, P. (2013). *Developmental Scale for North Carolina End-of-Grade/End-of-Course ELA/Reading and English II, Fourth Edition*. Monterey, CA: Pacific Metrics Corporation.

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.