Using Linguistic Profiles of Test Items to Improve Stability and Efficiency of Pre-equating in English Language Proficiency Assessment

For Symposium Session: Pre-Equating in Large Scale Assessment

Huan Wang and Bin Wei
Data Recognition Corporation

The Annual Meeting of the National Council on Measurement in Education
April 30, 2013, San Francisco, CA

Introduction

Stability of pre-equated item parameters has been reportedly susceptible to factors such as position and response type of the item, size and characteristics of the sample, and dimensionality of the test (e.g., Eignor & Stocking, 1986; Kolen & Brennan, 2004; Kolen & Harris, 1990; Meyers, Murphy, Goodman, & Turban, 2012; Stocking & Eignor, 1986; Tong, Wu, & Xu, 2008). In a language proficiency assessment, linguistic expectations from the test items are directly related to the measured construct (Bachman & Palmer, 2010) and may have implications for test dimensionality. The present study investigates the possibility of using linguistic profiles of test items to improve stability of pre-equated item parameters in English language proficiency assessment. The study also explores the usefulness of systematically controlling linguistic profiles in reducing the demand for field-test sample size. A smaller sample size is particularly valuable for quickly replenishing item pools and reducing item exposure.

Method

Data

Real-life test data were used in the study that came from a large-scale K–12 English language proficiency assessment program. The assessment included four subtests (Listening, Speaking, Reading, and Writing) and covered five grade span levels (K–1, 2–3, 4–5, 6–8, and 9–12). The present study focused on the Speaking subset form for the grade span 6–8. A total of over 10,000 operational test records were obtained from the target test population for this form.

The Speaking subtest was unique in that all its items were constructed-response. The test form from the grade span 6–8 was chosen for its representativeness of the Speaking subtest across grade spans regarding coverage of item types and intended test standards. This subtest form has 20 items and 41 score points in total. All the items from the form were field tested and their item parameters estimated. The test form has been used in operational administrations since then, and raw score (RS) to scale score (SS) tables based on the established item parameters have been applied for score reporting.

Procedure

A three-phase analysis was conducted. In Phase 1, exploratory factor analysis (EFA) with a promax rotation was used to investigate data dimensionality, which was then linked to linguistic features of test items. The EFA analysis was conducted in SAS 9.1. Results from the EFA analysis were used to inform identification of key linguistic profiles that could have noticeable impact on dimensionality.

In Phase 2, hypothetical sets of operational items and pretest items were chosen from the Grade Span 6–8 Speaking subtest form controlling for their coverage of key linguistic profiles. In particular, a fixed set of operational items and three alternative sets of pretest items were chosen. These item sets were constructed so that the hypothetical operational item set included items that covered all three key linguistic profiles that had been identified. Regarding the three hypothetical alternative pretest sets, Set #1 included items covering one single linguistic profile; Set #2, two linguistic profiles; and Set #3, three. The three alternative pretest sets also shared two common items.

In Phase 3, calibration was conducted for all three possible pairs of the hypothetical operational and pretest item sets. Items were calibrated using the two-parameter partial credit (2PPC) model. The Stocking and Lord procedure was used in equating with the hypothetical set of operational items as anchor. Six random samples of three different sizes (N=2,000, 1,000, and 500, with two samples per size) were drawn from the test population and used as the calibration sample for each pair. All the calibration and equating analyses were conducted using PARDUX (Burket, 2002), a proprietary software that has been routinely used for industry-scale applications at CTB/McGraw-Hill Education.

Stability of the pretest item parameters was then examined by comparing item parameter estimates of the shared items across samples of same or different sizes for each pretest item set. Efficiency in pre-equating was investigated by comparing changes in item parameter estimates from the baseline across pretest item sets per sample size. Item parameter estimates from the full sample size calibration and subsequent equating with all items as anchor were used as the baseline in comparisons.

Results

Investigation of Data Dimensionality

The Cronbach's reliability estimate for the Grade Span 6–8 Speaking subtest form was over 0.90, suggesting reasonable data dimensionality assumption for applying the unidimensional item response theory (IRT) model. Results from the EFA using a promax rotation show that the eigenvalues for the first four factors were 10.41, 0.73, 0.60, and 0.26, respectively. The scree plot shows that the "elbow" falls on the second eigenvalue. It should be also noted that the third eigenvalue, although small, is relatively closer in value to the second eigenvalue than with the fourth. The first three factors, therefore, were retained for investigation of their relationship with item-level linguistic expectations.

Identification of Key Linguistic Profiles

Standardized regression coefficients by test item were obtained for each of the three retained factors (see Table 1). Any factor loading larger than 0.30 was bolded and highlighted in yellow. There were 11 marked items in total under Factor 1, 7 under Factor 2, and 4 under Factor 3. Two items were marked for both Factor 1 and Factor 2.

Results from mapping those marked items under each factor to their linguistic expectations suggest that the three factors can be explained by a combination of the following linguistic features: 1) length of the expected response, 2) type of function that language is intended to serve, and 3) topical characteristics of the elicited language.

Table 1.

Standardized Regression Coefficients of the Three Retained Factors

Item ID	**Factor 1**	**Factor 2**	**Factor 3**
1	**0.78210**	0.14723	-0.06452
2	**0.67240**	0.18019	-0.04689
3	**0.58595**	0.28100	-0.04574
4	0.25949	**0.43202**	0.10110
5	0.03914	**0.50646**	0.21118
6	0.22485	**0.61526**	0.02327
7	0.26632	**0.57291**	0.01331
8	-0.15000	**0.73081**	0.12840
9	**0.31360**	**0.58251**	-0.07460
10	**0.39223**	**0.50879**	-0.02907
11	**0.62457**	0.19104	0.06706
12	**0.60674**	0.18867	0.19009
13	**0.74516**	-0.02104	0.04707
14	**0.59261**	-0.02735	0.25087
15	**0.87814**	-0.09620	-0.00589
16	**0.75967**	0.03143	0.14154
17	0.05797	0.13494	**0.51903**
18	0.03888	0.13752	**0.54885**
19	-0.03851	-0.01484	**0.54930**
20	0.19795	0.04573	**0.61948**

Items under Factor 1 generally elicit a single utterance in oral communication that performs ideational functions (e.g., naming or describing objects). Factor 2 is similar to Factor 1 regarding the expected length of discourse and type of function, but shows more demand on disciplinary-specific language use that is typical in academic settings.

Different from the first two factors, items from Factor 3 tend to elicit two or more utterances that require using language to perform manipulative (such as making a request) or heuristic (such as explaining a process) functions. These items cover language use from both social and academic settings.

Selection of Hypothetical Operational and Pretest Item Sets

Based on the analysis above, three key linguistic profiles were identified. They correspond to the three factors that were obtained from the EFA analysis. Using these key linguistic profiles, a hypothetical operational item set and three hypothetical alternative pretest item sets were selected, as shown in Table 2.

It can be seen that nine items are included in the selected hypothetical operational item set. These items cover all three key linguistic profiles that were identified. The hypothetical Pretest Set #1 has five items that cover only one key linguistic profile. Pretest Set #2 has six items and cover two key linguistic profiles. Pretest Set #3 spans across three linguistic profiles with five items. There are two common items (Items #1 and #11) that are shared across the three pretest item sets. Both of them are from the first linguistic profile.

Table 2.

Hypothetical Operational (OP) and Pretest Item Sets and Relation to Linguistic Profiles

Item ID	**Three Key Linguistic Profiles**			**Hypothetical OP Item Set**	**Three Hypothetical Alternative Sets of Pretest Items**		
	Profile # 1	**Profile #2**	**Profile #3**		**Set #1**	**Set #2**	**Set #3**
1	X				X	X	X
2	X				X		
3	X			X			
4		X				X	X
5		X				X	
6		X				X	
7		X				X	
8		X		X			
9	X	X		X			
10	X	X		X			
11	X				X	X	X
12	X				X		
13	X				X		
14	X			X			
15	X			X			
16	X			X			
17			X				X
18			X				X
19			X	X			
20			X	X			

Variation in Item Parameter Estimates across Samples

Figure Sets 1 and 2 below illustrate how resultant item parameter estimates of the two shared items vary across samples for each pretest item set. Item parameter estimates obtained from the full sample size (N>10,000) calibration and subsequent equating with all items from the subtest form as anchor were provided as well and used as the baseline.

Figure Set 1. Estimated item parameters across samples (N=2,000, 1,000, and 500): Item ID #1.

N=2,000 (Parameters 01 and 02)

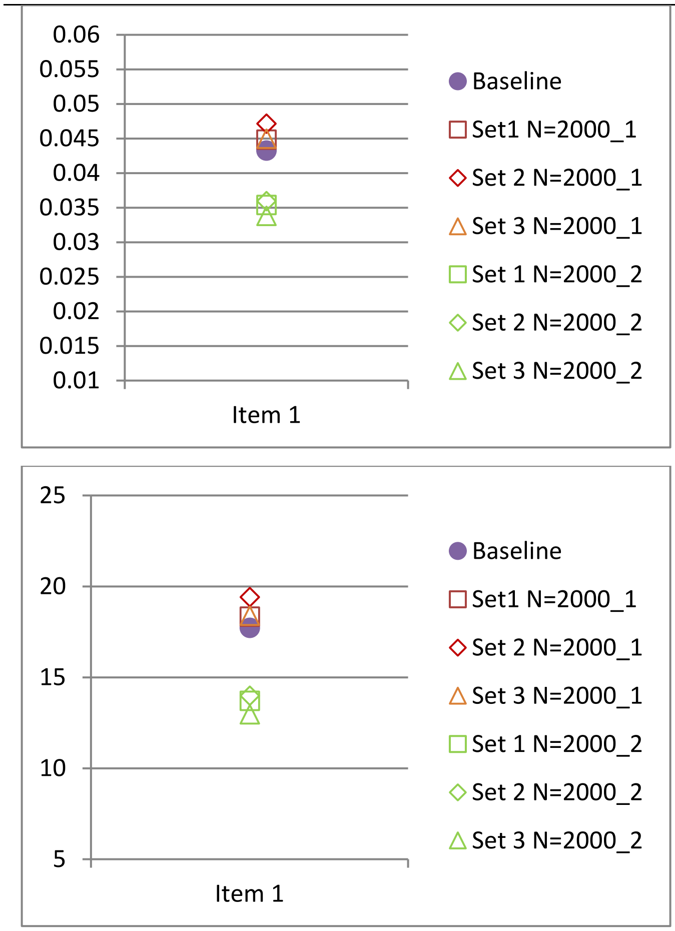
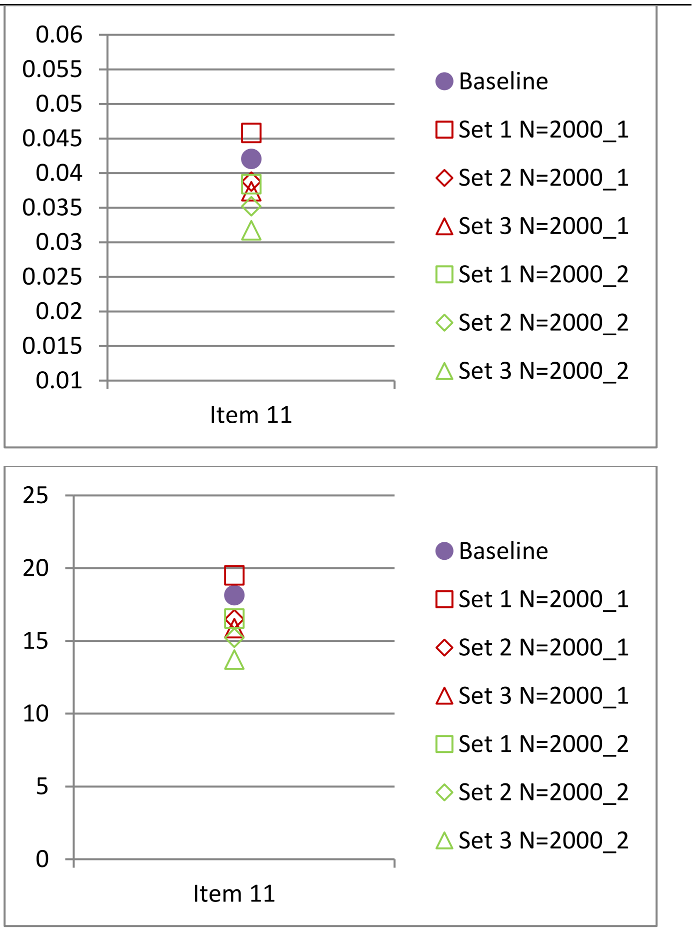
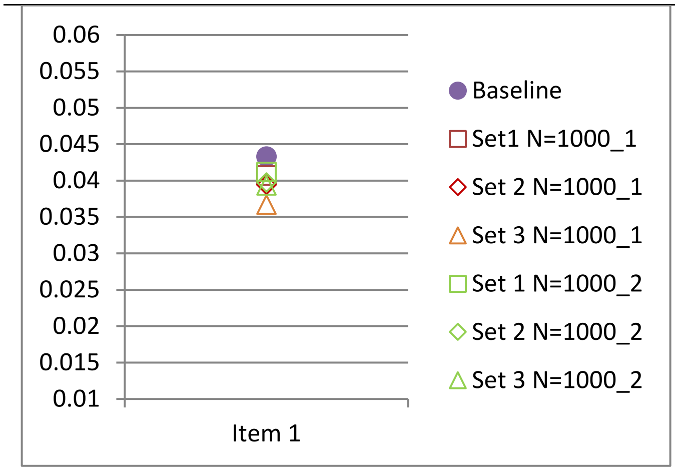
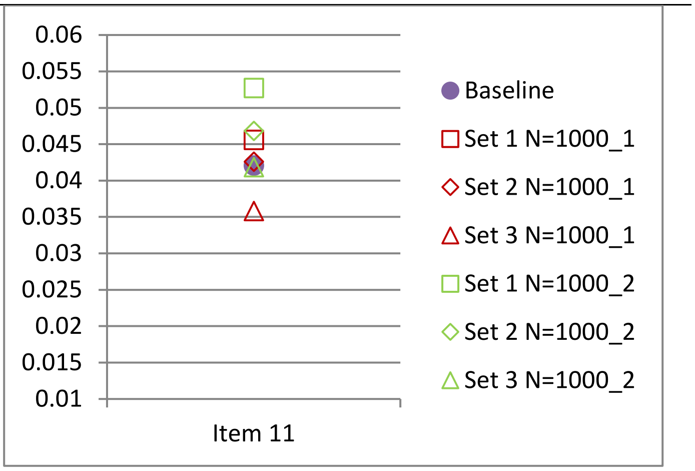
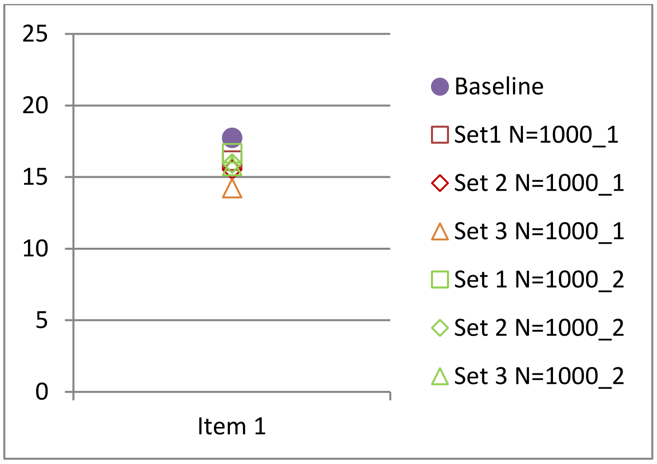
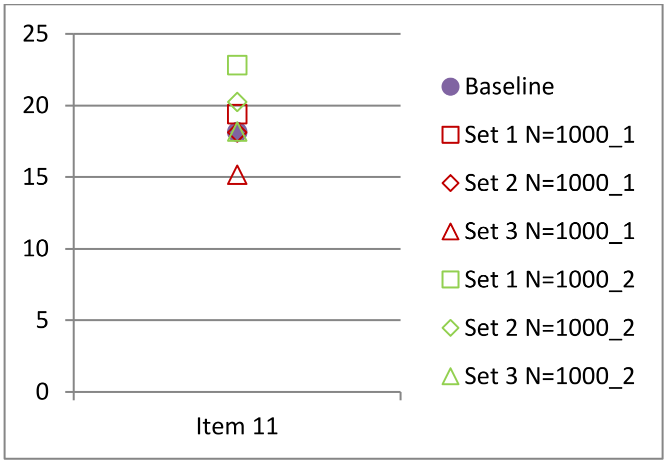
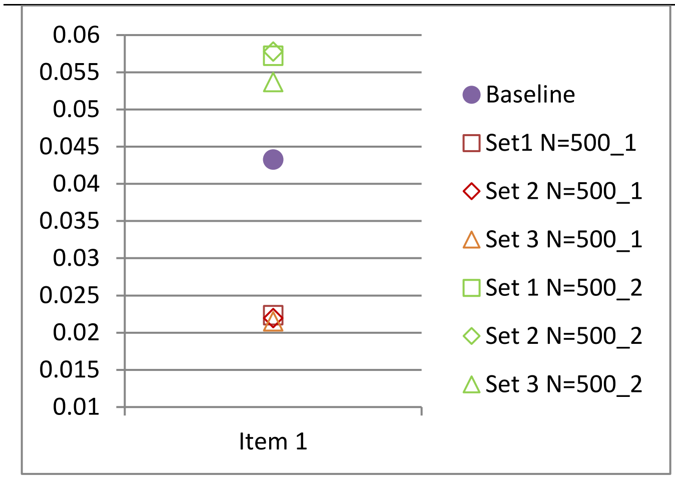
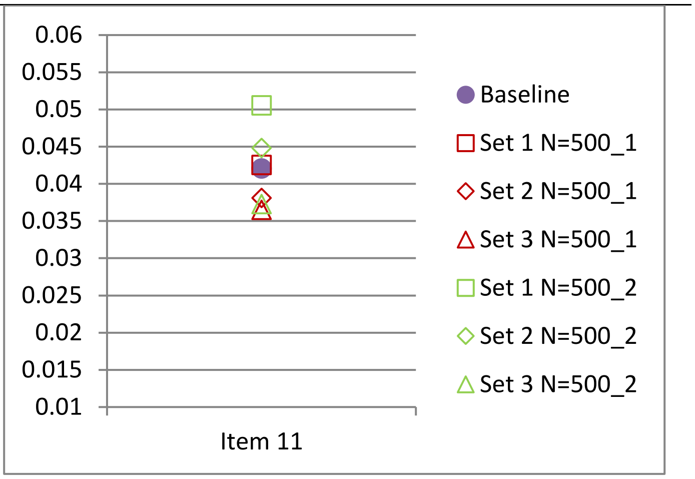
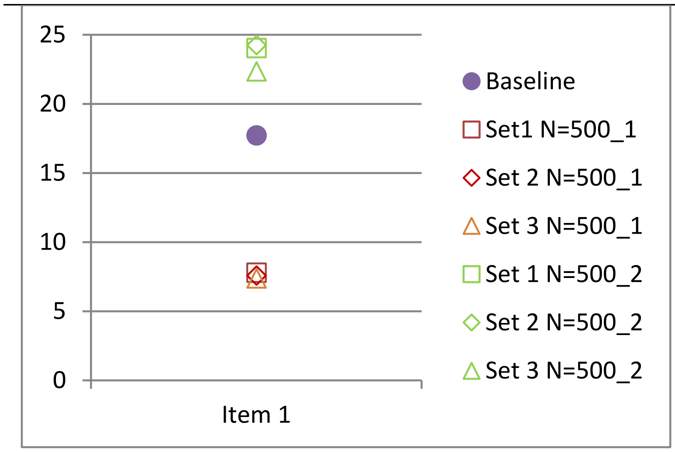
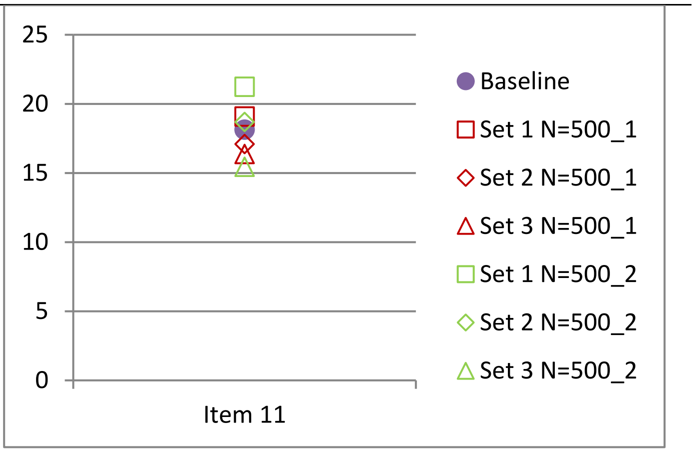


Figure Set 2. Estimated item parameters across samples (N=2,000, 1,000, and 500): Item ID #11.

N=2,000 (Parameters 01 and 02)



N=1,000 (Parameters 01 and 02)	N=1,000 (Parameters 01 and 02)
	
	

N=500 (Parameters 01 and 02)	N=500 (Parameters 01 and 02)
	
	

It was found that no particular trend in item parameter changes across samples held for each pretest item set. The pattern varies across items, across samples of the same size, and across samples of different sizes. It is noteworthy that similar trends in item parameter changes were observed for the first and second parameters for any given set of samples. These observations suggest that sample size and quality may have greater impact than linguistic profiles of the pretest items on the stability of pre-equated item parameters for the two items (Items #1 and #11) under study.

Change in Item Parameter Estimates across Pretest Item Sets per Sample Size

Table 3 summarizes the average absolute changes in item parameter estimates from the baseline across pretest item sets per sample size. Note that only the first two parameters were included in the summary.

Table 3.

Average Absolute Change in Item Parameter Estimates from the Baseline across Pretest Item Sets per Sample Size

Pretest Item Set	**Sample Size**	**Minimum Absolute Change**		**Maximum Absolute Change**		**Average Absolute Change**	
		Parm 01	**Parm 02**	**Parm 01**	**Parm 02**	**Parm 01**	**Parm 02**
Set #1	N=2,000	0.0006	0.0818	0.0079	4.0231	0.0029	1.2983
Set #2	N=2,000	0.0002	0.0575	0.0073	3.7344	0.0039	1.8724
Set #3	N=2,000	0.0001	0.0938	0.0103	4.7524	0.0038	1.7818
Set #1	N=1,000	0.0007	0.0263	0.0107	4.6668	0.0033	1.5285
Set #2	N=1,000	0.0005	0.037	0.0148	6.9964	0.0033	1.5904
Set #3	N=1,000	0.0002	0.0311	0.0066	3.5079	0.0039	1.9124
Set #1	N=500	0.0005	0.3690	0.0209	9.9333	0.0070	3.1491
Set #2	N=500	0.0002	0.0127	0.0213	10.1422	0.0069	3.1847
Set #3	N=500	0.0005	0.1658	0.0217	10.3478	0.0053	2.4890

Set #3 was observed to have the largest maximum absolute change in both the first and second parameters except when the sample size was 1,000. This exception may be explained by better overall quality of the two random samples for N=1,000 that had been drawn. Despite the slightly larger maximum absolute change, Set #3 shows the lowest average absolute change in item parameter estimates when the sample size is relatively small (N=500)---given use of the 2PPC model and multiple score levels associated with each item.

Discussion

Results from the study provide supporting evidence that covering key linguistic profiles that relate to data dimensionality in pretest items of a language proficiency assessment may be particularly valuable in scenarios where the sample size for calibration and equating is relatively small. Given the increased chance of sampling error for small samples, however, it is also important to improve quality of the small samples to achieve more stable and reliable item parameter estimates from pre-equating, as suggested by earlier discussion from comparing item parameters across samples for Items #1 and #11.

It should be noted that linguistic profiles may not be equivalent to item types in the sense of multiple-choice versus constructed-response items. As shown in the present study, even when all items were constructed-response items, the test may still show data dimensionality that relate to multiple linguistic aspects of the assessment tasks.

The study provides preliminary results that call for more attention to the relationship between item-level linguistic expectations, data dimensionality, and sample size and quality in pre-equating for language proficiency assessment. Future work is recommended to validate and model such relationship on a larger scale for a more comprehensive and systematic understanding of the interactions of various factors in improving stability and efficiency of pre-equating.

References

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice* (2nd ed.). Oxford: Oxford University Press.

Burket, G.R. (2002). *PARDUX* [Computer program]. Monterey, CA: CTB/McGraw-Hill.

Eignor, D. R. & Stocking, M. L. (1986). An investigation of possible causes for the inadequacy of IRT true-score pre-equating (Research Report 86-14). Princeton, NJ: Educational Testing Service.

Kolen, M.J., and Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). USA: Springer Science+Business Media, Inc.

Kolen, M.J., & Harris, D.J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27*, 27-39.

Meyers, J.L., Murphy, S., Goodman, J., & Turban, A. (2012, April). *The impact of item position change on item parameters and common equating results under the 3PL model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, B.C.

Stocking, M. L. & Eignor, D. R. (1986). *The impact of different ability distribution on IRT preequating*. (RR 86-49). Princeton, NJ: Educational Testing Service.

Tong, Y, Wu, S-S, & Xu, M. (2008). *A comparison of Pre-Equating and Post-equating using large-scale assessment data*. Paper presented at the American Educational and Research Association Annual Meeting in New York City.