An Empirical Investigation of Item-Pool and Year-to-Year Equating Plans: Using Large-Scale Assessment Data

Xinya Liang
University of Arkansas, Fayetteville

Jin Koo
The Enrollment Management Association

Hülya Yürekli
Turkish Ministry of National Education

Insu Paek
Florida State University

Betsy Jane Becker
Florida State University

Salih Binici
Florida Department of Education

Hirotaka Fukuhara
Pearson

In large-scale assessments, constructing multiple test forms can increase test security, allow for tests to be implemented on different dates, and meet other practical needs (Kolen & Brennan, 2004; Kolen & Whitney, 1982; van der Linden & Adema, 1998). Operationally, multiple test forms are constructed similar to be in content and statistical specifications. Differences in test difficulty levels are adjusted through an equating process. Equating aims to adjust for differences in test-form difficulty so that scores obtained from multiple forms can be interpreted interchangeably (Kolen & Brennan, 2004). The equated scores are considered to be comparable across different test forms.

At present, a variety of data-collection designs and equating methods have been developed to obtain comparable scores across test forms (Holland, 2007). One prevalent data collection design is the common-item nonequivalent groups design (Kolen & Brennan, 2004). This design uses a set of common items (also called anchor items) to equate different test forms. Anchor items are administered as part of the tests, and can be internal, if counted toward the total score, or external, if not contributing to the total score. An advantage of this design is to separate group ability differences from test-form differences. Score equating based on item response theory (IRT; Hambleton & Swaminathan, 1985) has been widely applied in large-scale assessments. IRT equating places item-parameter estimates from different test forms or administrations on the same scale, then equates scale scores so they can be interpreted interchangeably across test forms (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004; Petersen, Cook, & Stocking, 1983). Based on IRT equating, different equating linkage plans have been developed

to equate multiple test forms (Kolen & Brennan, 2004; see Section 8.2.2). Equating plans involve the choice of a base score scale to which the new form is to be equated. Common ways of establishing the base scale are to use a calibrated item pool or use a score scale from an old test form (Battauz, 2013; Kolen & Brennan, 2004). Additional complexity is added to the equating plan if multiple forms are equated through multiple links and chains.

In current large-scale assessments (e.g., Educational Testing Service, 2009; Florida Department of Education, 2006), two popular equating plans based on IRT methods are the item-pool equating plan and the year-to-year equating plan. Each equating plan has its unique benefits and issues in educational practice. Our study aimed to compare the performance of these two equating plans using empirical data from large-scale assessments. Specifically, this study had two major purposes that were explored. The first purpose was to compare the year-to-year and item-pool equating plans in terms of IRT-based scaling coefficients, equated scores, and conditional standard errors of measurement. The second purpose was to evaluate the impact of equating linkage plans on scale scores and achievement-level classifications relevant to score reporting. In the next section, we first briefly introduce each equating plan and its practical influences in large-scale assessments. We then summarize findings from the current literature regarding different linkage plans and their practical concerns.

Item-Pool Equating Plan

In the IRT-based item-pool equating plan (also called common-item equating to a calibrated pool; Kolen & Brennan, 2004; see Section 6.9), common items are equated to a calibrated item pool to establish score equivalence over time. A calibrated item pool (Lord, 1980; Vale, 1986) contains a group of operational items from all previously administered tests. Item parameter estimates from the item pool are placed on a common scale. When a new form is constructed, some common items from the pool and some newly developed items are included. After the test is administered, parameter estimates from the new form are transformed to the score scale through the common items that was established for the item pool. IRT equating methods can then be applied to complete the equating process. With calibrated item parameters, new items can be added to the calibrated item pool after the equating is completed. The item pool is regularly maintained and can be continuously expanded by updating item parameters, adding new test items, and removing outdated items.

Item-pool equating is similar to the common-item nonequivalent groups design. They both use common items to conduct equating, but common items in the item-pool equating plan are drawn from an item pool rather than from a reference form. This provides greater flexibility for anchor-item selection and test-form construction. Anchor items can be selected from a pool consisting of all calibrated items ready for use, and the selection is not restricted to items from a specific previous test (Kolen & Brennan, 2004; Tate, 2003). In addition, item-pool equating minimizes the number of links from the new form to the base scale (Kolen & Brennan, 2004). The one-step equating to a pool may reduce equating errors accumulated over multiple years and thus improve accuracy of equating (Tate, 2003).

Despite its sound features, many practical issues need to be considered when using the item-pool equating plan. First, selecting items from all previous tests in the pool may result in position shift of common items. Shifts in item positions can affect student performance and decrease accuracy of IRT parameter estimates for item calibration and equating. Therefore, it is important to place common items at similar positions to where they appeared in previous test forms (Kingston & Dorans, 1984; Kolen & Brennan, 2004; Meyers, Miller, & Way, 2008). Second, the IRT procedure may be the only option for item-pool development and equating (Kolen & Brennan, 2004). If IRT assumptions are violated (Hambleton & Swaminathan, 1985), for example the test form is multidimensional, item-parameter estimates may be affected when using a unidimensional IRT model, leading to a distorted base scale and inadequate equating. To satisfy the unidimensionality assumption, common items should be selected with the same content specifications and represent well the total test content proportionally. This requirement for IRT equating may be relaxed if the IRT model fits the population and a reasonable number of anchor items is selected.

Third, maintaining an item pool takes time and effort, and is not always cost effective. To keep the item pool usable, test content should be aligned with recent curriculum standards, and the exposure of items should be minimized to preserve test security (Battauz, 2013). To meet these requisites, psychometric properties of items need updating after each equating (Arai & Mayekawa, 2011). Items can be added or deleted periodically to warrant the quality of the item pool (Kolen & Brennan, 2004). The constant updates could lead to differential change in parameter values over time, also called item parameter drift (Bock, Muraki, & Pfeiffenberger, 1988; Goldstein, 1983). Consequently, the base scale established from an item pool can be slightly changed after each adjustment in the pool. New forms administered after the item-pool adjustment may be equated to a slightly different score scale.

Year-to-Year Equating Plan

In the year-to-year equating plan, the base score scale is established on a single form of the test. A new test form is equated to the base scale through an equating chain. For example, considering a test that is administered and equated once a year, the base scale is determined using a single test form in one specific year. In the second year, the new form is equated to the base scale through parameter estimates of items common to the new and old forms. In the third year, the new form is equated to the form administered one year prior (i.e., the second-year form), but it will also be placed on the base scale through the second year's equating. Regardless of the number of equating links, the test form will be eventually linked back to the base scale. This plan adjusts scores on the new form based on common items from an old form. Therefore, it does not require the construction of a calibrated item pool. Once the base score scale is established, it will remain the same in the following equating activities. Compared to item-pool equating, this plan is easy to implement, and time and cost effective.

Nevertheless, some practical problems are associated with year-to-year equating. The goal of equating is to produce comparable scores among forms over a long period. For new forms given in successive years, as the year of linking increases, the number of links from the new form to the initial form also increases. This may cause equating errors to accumulate over time that lead to systematic scale drift (Guo, 2010; Haberman & Dorans, 2009). Linking through an equating chain can result in considerable amounts of errors. Scale scores obtained in later years may indicate different levels of performance than those in earlier years of equating (Kolen & Brennan, 2004). The scale drift in the year-to-year equating results from errors in a series of equating, whereas the scale drift in the item-pool equating results from updates to the item pool. In addition, several other problems exist when an old form is used as a link form in the equating chain. First, the use of common items from the link form administered in one previous year may cause test security concerns more seriously than selecting common items from an item pool. Second, in the same way that violations of unidimensionality are problematic in item-pool equating, when a test shows a multidimensional structure, parameter estimates may not be accurate and appropriate for item calibration in year-to-year equating. This may affect equating results in the given year as well as in the years thereafter. Third, if any link form is not well constructed to align with the curriculum, equating errors could be greater in that year and may also seriously impact score equating in successive years (Kolen & Brennan, 2004; Tate, 2003). Maintaining score stability is an important issue for any testing program. When any of these problems occurs, the equating chain becomes unreliable.

The Current Study

Previous studies have investigated various impact of different equating linkage plans on equating results. One research interest was to identify factors affecting the reliability of equating results (e.g., Puhan, 2008; Von Davier, 2011; Wang, Qian, & Lee, 2013). It was found that equating biases slightly increased for shorter anchor test length and/or reduced sample sizes, whereas the biases were too small to raise practical concerns. To improve estimation accuracy of equating coefficients in linkage plans involving chains or multiple links, Battauz (2013) proposed an efficient weighting method for averaging coefficients and derived standard errors for equating coefficients. With respect to development of an item pool, a simulation study by Arai and Mayekawa (2011) examined three linking plans for developing an IRT calibrated item pool, using four item-calibration methods. They found that the item characteristic curve method performed the best in general. In addition to the methodological challenges, educational testing programs often encounter many practical issues (Kolen & Brennan, 2004), such as item-parameter drift and scale drift (Goldstein, 1983). These issues have been extensively investigated for the item-pool equating plan (Bock et al., 1988) and the year-to-year equating plan (Guo, 2010; Guo, Liu, Dorans, & Feigenbaum, 2011), respectively.

It has been common practice in large-scale assessment programs to use the item-pool and the year-to-year equating plans. We reviewed the most recent public assessment technical reports for the 50 states in the U.S. between 2009 and 2016. It shows that approximately 42% of the states have used an item-pool equating plan and 32% of the states have used a year-to-year plan on their standard state assessment programs. Other states (about 26%) either did not specify the equating plan adopted, or did not have technical reports accessible to the public. Although a large number of states have used these equating plans, few empirical comparisons have been made between the item-pool and year-to-year equating plans using data from large-scale assessments. Simulation studies with limited number of conditions may not fully address practical concerns in large-scale assessments. How and to what extent the two equating linkage plans behave differently has not been discussed in the literature. Moreover, little is known about how student scale scores and achievement level classifications are impacted empirically by different linkage plans. An empirical study is uniquely beneficial for investigating these issues in different equating plans, given the complexity in real testing situations.

Accordingly, the purpose of this study was to empirically examine the item-pool and year-to-year equating plan using empirical data from a large-scale assessment. In the next section, we present the data, analysis methods, and evaluation criteria, followed by the study results. We then discuss benefits and limitations of the two equating plans and conclude the article with implications for current educational practice.

Methods

Data Source

This study used large-scale mathematics assessment data from a southern state. The mathematics assessment is designed to measure students' mathematics achievement in grades 3 to 10, and is administered to all students in the state once every year. The tests contain multiple-choice (MC) and gridded-response (GR) items. Both were dichotomously scored as either correct of incorrect. The population for data analysis was all 3rd, 6th, 7th, and 9th grade students tested in the years 2007, 2008, 2009, and 2010. Approximately 200,000 students are tested in each grade each year. Students receive reports with both scale scores and performance levels. The integer scale scores range from 100 to 500. For reporting, all students are classified into five achievement levels (ALs), scored 1 (lowest) to 5 (highest), on the basis of their

scale scores. Achievement levels of 3 or higher are considered to be on-grade. Four cut points were determined through a standard-setting process, and approved by the state's Board of Education. The ALs are used for retention decisions in grade 3 and are a part of graduation requirements using the grade-10 ALs.

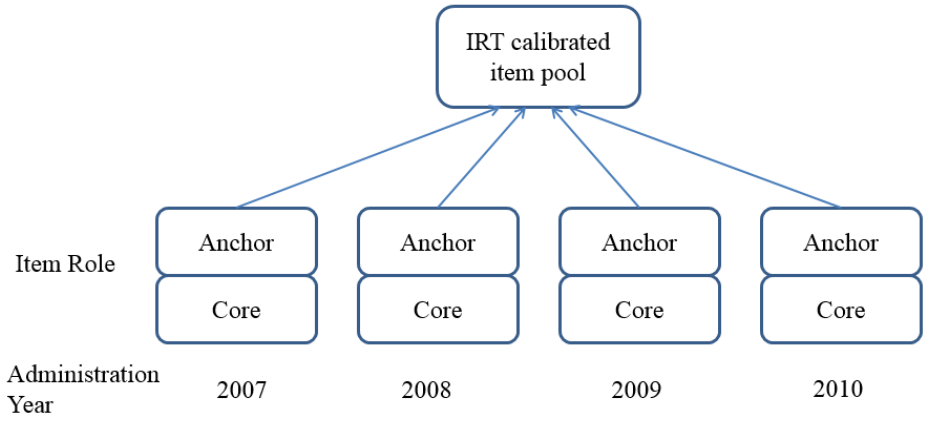
Equating Linkage Plans

Data collected for IRT equating in this study are based on common-item equating to a calibrated item-pool design (Kolen & Brennan, 2004). The external anchor-test design is used to statistically adjust for differences in group abilities. Specifically, the assessment consists of anchor items, core items, and field-test items. Anchor items are not included in calculation of the total scores but are used for purposes of item calibration and score equating. Core items are the only ones used for scoring students. Field-test items are also not included in scoring; their item statistics are evaluated to determine whether they can be used in future test administrations. In the current study, field-test items were excluded from the data analysis. The total test length for each grade was 60-70 items, consisting of 15-32 anchor items and 40-44 core items. The same data were used for investigating both item-pool and year-to-year equating plans.

The conceptual models for the item-pool and year-to-year equating plans in this study are shown in Figure 1. In the item-pool plan, anchor items were selected based on the test blue print and the number of anchor items that has been used operationally. Anchor items for a given administration were selected from an item pool established on all past calibrations of items. For instance, in 2007, a total of 26 anchor items were used operationally and the same number of anchor items were selected based on the test blue print in the item-pool plan in this study. For each of the four years from 2007 to 2010, item parameter estimates were calibrated to an item pool; scores were also equated to the scale of the item pool. In the year-to-year plan, anchor items were selected from the core and anchor items in the previous year's administration. Item calibration and score equating in each year were performed based on the scale from one year prior. The base scale was established using the ability scale in 2007. Then an equating chain was established where tests were equated from 2008 to 2007, 2009 to 2008, and 2010 to 2009. All tests can thus be linked back to the 2007 score scale. For example, in 2008, the full anchor set used in the item-pool plan had 30 anchor items. We selected 20 items out of those 30 anchor items for the year-to-year plan. The 10 items excluded were either not from the previous year or were used as field-test items previously. Because of this, the number of anchor items in the year-to-year plan was the same or less than that in the item-pool plan.

The percentage of anchor items common to the two equating plans ranged from 53% to 100% across the four years and all grades. For grades 3, 6, 7, and 9, the overlap percentages of anchor items range from 71% - 100%, 62% - 75%, 66% - 75%, and 54% - 69%, respectively. The year 2007 had the smallest difference between the anchor-item sets used in the two plans (92 % overlap on average across all the grades), while the year 2010 had the greatest difference (64% overlap on average across all the grades).

Item-Pool Equating Plan



Year-to-Year Equating Plan

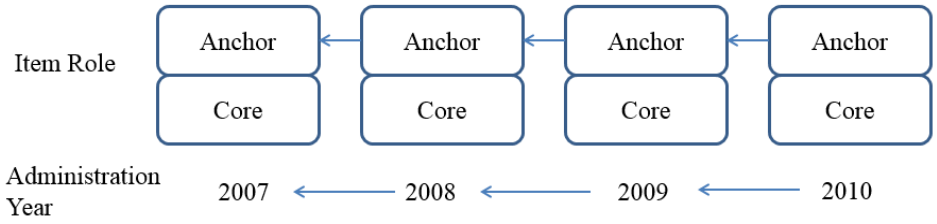


Figure 1. Equating Linkage Plans.

Item Response Theory Equating

Test score equating based on IRT methods was performed. Item parameters were estimated using the 3-parameter logistic (3-PL) IRT model (Lord, 1980) for multiple-choice items and the 2-parameter logistic (2-PL) IRT model for gridded-response items. MULTILOG 7 (Thissen, 2003) was used to conduct IRT calibration to place both core and anchor item parameter estimates on the same scale (Kim & Cohen, 1998). Maximum likelihood estimation was used (e.g., Rupp, 2003) for estimating student abilities or theta scores (θ). Analysis of scale dimensionality used the Q1 and Q3 statistics (Yen, 1981, 1984). Both indicated that unidimensional IRT models fitted well to the achievement data.

The test-equating process followed three steps. First, item calibration was performed after implementation of new tests each year. New item parameter estimates for both anchor and core items were obtained using IRT procedures, given the student response data. Second, old anchor-item parameter estimates were readily available in an item pool for the item-pool plan, and from the previous year's test for the year-to-year plan. To place new and old item parameter estimates onto a common scale, the test characteristic curve (TCC) method by Stocking and Lord (S-L; Stocking & Lord, 1983) was used to obtain scaling coefficients M1 (slope) and M2 (intercept) to link the new and old anchor-item parameter estimates. The M1

and M2 coefficients were linearly transformed to be on the scale score units, and then used together to adjust the core-item parameter estimates obtained during the calibration. Third, student-ability theta scores were estimated using the adjusted core-item parameters. The theta scales were constructed to have means of 0 and standard deviations of 1 at the population level. The scale scores, ranging from 100 to 500 within each grade, were computed using linear transformations of each year's theta estimates.

Evaluation

We evaluated the two equating plans based on differences in Stocking and Lord transformation coefficients (M1 and M2), test characteristic curves, conditional IRT standard errors of measurement (SEMs), effect sizes (ESs) for differences in scale-score means, ratios of scale-score standard deviations (SDs), scale-score differences, and the classification rates of students into the five levels of achievement used for reporting purposes.

The relative precision of measurement can be assessed using the conditional SEM in IRT, calculated as:

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}}, \qquad (1)$$

where $I(\theta)$ is the test information function at ability level θ (Cowell, 1991), defined as a function of item parameters and the probability of a correct response. The larger the conditional SEM, the less precise or the less reliable the test score is at ability level θ.

Considering the large sample sizes in this study, conducting statistical tests for the comparison of mean scale scores tends to conclude significant results, even with a tiny difference in means. An alternative measure of the mean difference is the effect size (ES). Effect size is a standardized measure of group mean differences. One common definition of the effect size is to use Cohen's *d* (Cohen, 1988). Specifically in this study, the effect size *d* is defined as:

$$d = \frac{\bar{x}_{yy} - \bar{x}_{ip}}{SD_{ip}}, \qquad (2)$$

where $\bar{x}_{yy}$ is the mean scale score from the year-to-year (*yy*) plan, $\bar{x}_{ip}$ is the mean scale score from the item-pool (*ip*) plan, and SD_{ip} is the standard deviation from the item-pool plan.

Results

To conserve space, we only present full results for grade 3 from 2007 to 2010; effect sizes and SD ratios are reported for all grades. Full results for grades 6, 7, and 9 show similar patterns, and are available upon request.

Statistics from IRT Equating

Table 1 shows the S-L transformation coefficients – M1 and M2 – in scale-score units for each item-pool and year-to-year analysis. The S-L transformation coefficients obtained from both equating plans should be similar if equating results from the two plans are comparable. Both M1 and M2 values from the item-pool plan were slightly larger than those from the year-to-year plan. The discrepancies in M1 were 3.15 score points or less and in M2 no greater than 1.92 score points on the 400 point scale. The ratios of the slopes and of the intercepts from the two plans were close to 1.

Table 1. Summary of Transformation Coefficients for Grade 3.

Linkage Plan	Slope (M1)				Intercept (M2)			
	2007	2008	2009	2010	2007	2008	2009	2010
Year-to-year	55.64	51.51	51.47	54.08	335.39	340.29	340.56	346.06
Item-pool	55.77	53.18	54.62	55.09	335.41	342.21	342.47	346.11
Ratio	0.998	0.969	0.942	0.982	1.000	0.994	0.994	1.000

Note. Ratio = $Slope_{yy}/Slope_{ip}$ or $Intercept_{yy}/Intercept_{ip}$.

Figure 2 shows the TCC differences between the item-pool and year-to-year plans ($TCC_{yy} - TCC_{ip}$) in the scale-score metric which ranges from 100 to 500. The horizontal reference line is at 0, indicating no difference between TCCs. The four vertical lines divide the scale scores into the five achievement levels. In 2007, the TCC differences were essentially nil across the whole ability scale. In both 2008 and 2009, differences in the TCCs were negative at lower ability levels (i.e., year-to-year scale scores were lower), but positive for higher ability scores (i.e., year-to-year scale scores were higher). Even so, the TCC differences were within 0.5 scale score points across the entire ability range. In 2010, the TCC differences were all positive and within 1 scale-score point. Although TCC differences tended to be slightly greater in the later years, less than 1 point differences on scale-scale units would not have any practical impact on the equated scores.

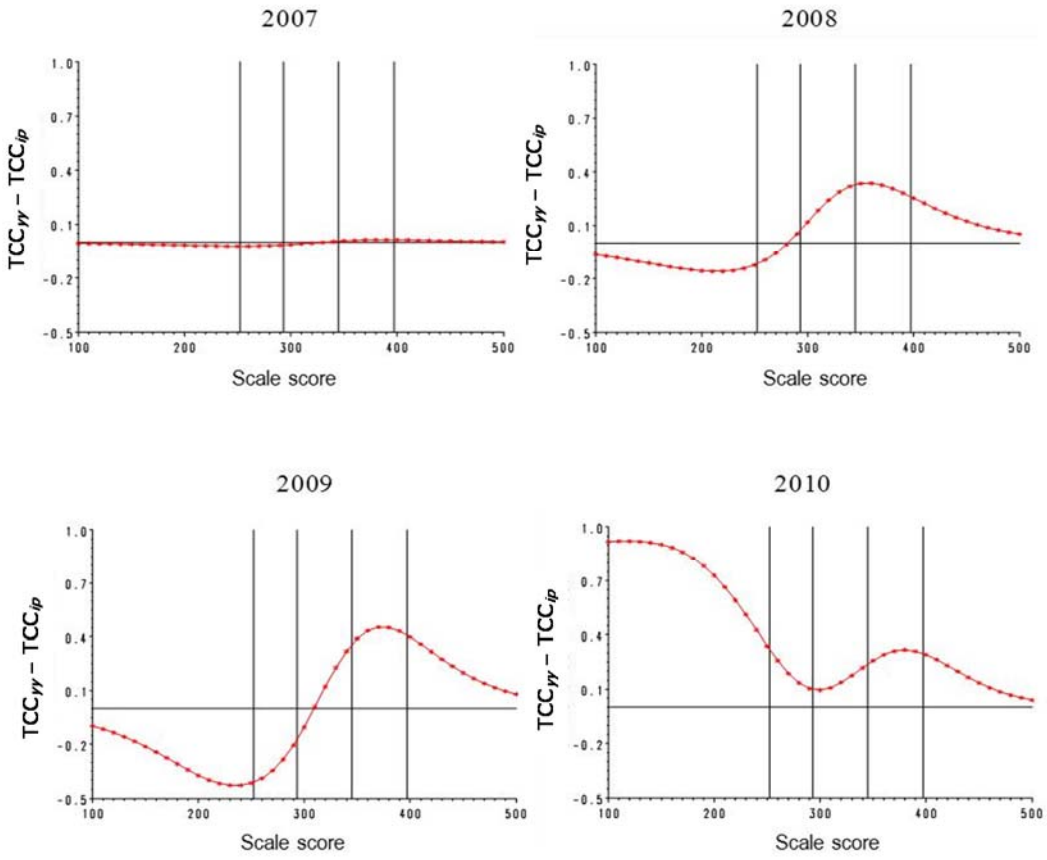


Figure 2. Differences between the TCCs for Grade 3 in 2007 through 2010.

Figure 3 depicts the conditional SEMs in the scale-score metric based on both item-pool and year-to-year plans. The differences between the two conditional SEM curves were very small in the middle of the scale-score range for all four years. The distances were relatively larger at both ends of the scale for later years, where the item-pool plan showed smaller SEMs than the year-to-year plan. Specifically, in 2007 and 2008, the conditional SEM functions from both plans were almost identical throughout the scale-score range. In 2009 and 2010, the conditional SEMs for scale scores between 100 and 180 based on the year-to-year plan were larger than those based on the item-pool plan by up to 25 scale-score points, or approximately 0.5 standard-deviation units. These differences may be due to the differing numbers of anchor items used, or to equating errors accumulated across years.

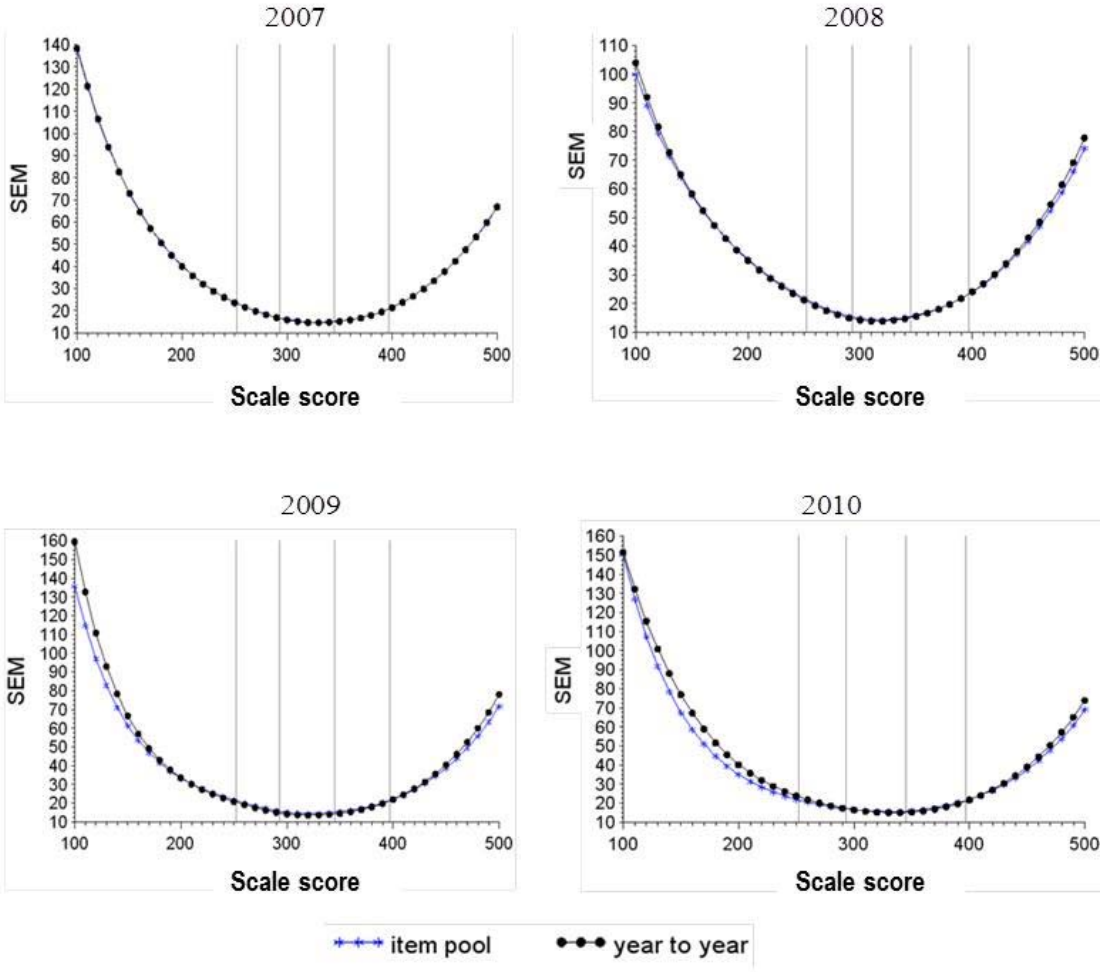


Figure 3. Conditional Standard Errors of Measurement for Grade 3 in 2007 through 2010.

Scale-Score (SS) Summary

Figure 4 presents standardized-mean-difference effect sizes for all grades from 2007 to 2010. Most ES values were negative, showing that mean scale scores from the item-pool plan were typically higher than those from the year-to-year plan. All ES values were very small ($|d| < 0.05$) according to Cohen's rules (Cohen, 1988). Figure 5 shows the ratios of SDs between the year-to-year and item-pool plans (SS_{yy}/SS_{ip}) for all grades. The ratios of SDs from the two plans were very close to 1, indicating the SDs from the two

plans were quite comparable. The scale-score SD differences were very small in the equating base year (2007) for all grades, and diverged slightly from 1 in the following years. The SD ratios showed no consistent trends, but indicated trivial differences in SDs between the two linkage plans.

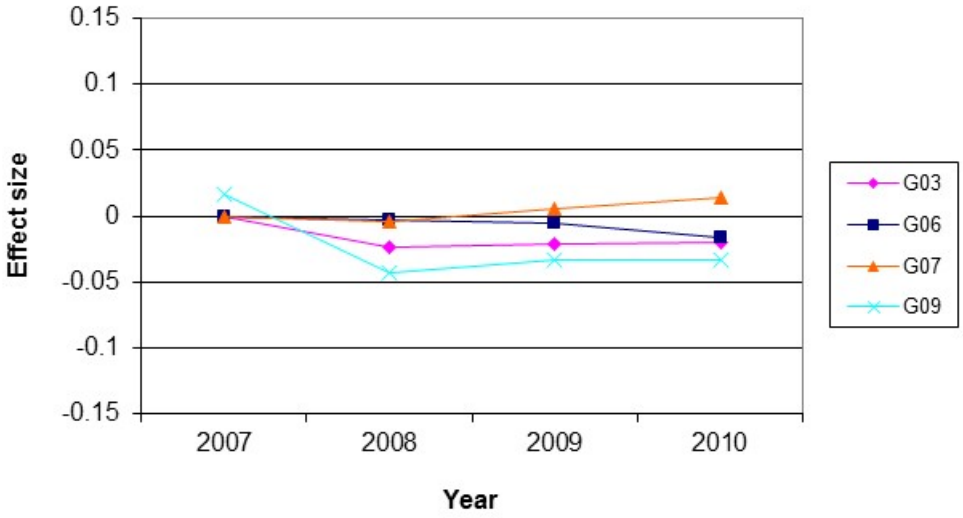


Figure 4. Effect Sizes by Grade and Year.

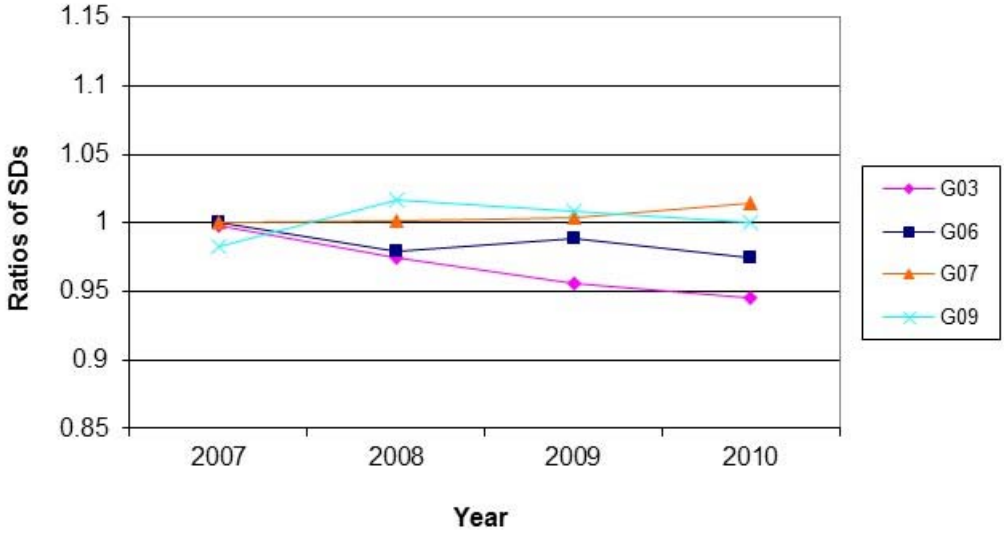


Figure 5. Ratios of the Standard Deviations of Scale Scores (SD_{yy}/SD_{ip}).

Figure 6 depicts the differences in scale scores from the two plans ($SS_{dif} = SS_{yy} - SS_{ip}$) versus the scale scores based on the item-pool plan (SS_{ip}) for only grade 3. Overall, SS_{ip} and SS_{dif} were linearly related. The two scores were similar (i.e., SS_{dif} was near 0) around the score mean of 300, but the score differences grew larger at the two ends of the scale. The enlarged score differences may be due to the fact that

the M1 (slope) coefficients were always larger for the item-pool plan. As a result, the item-pool plan produced more extreme scores than the year-to-year plan. Moving from 2007 to 2010, as the M1 difference increased and extreme scores became even more disparate, the lines on the plots became steeper.

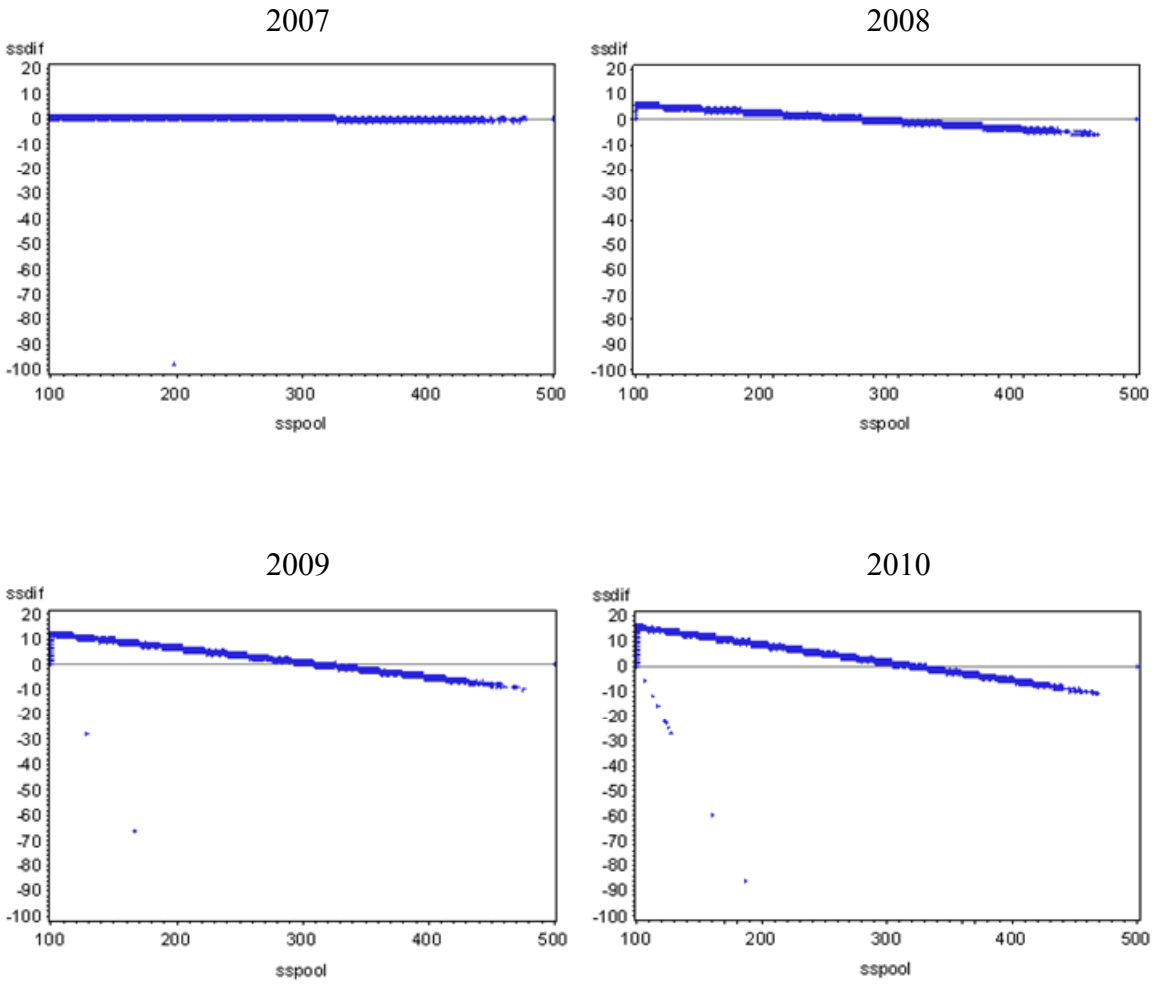


Figure 6. Scale-score Differences (SS_{yy}-SS_{ip}) for Grade 3 in 2007 through 2010.

A few points in the plots deviated vastly from the other points. These score pairs all occurred for students who obtained the lowest possible scale score based on year-to-year equating, SS_{yy} =100. So for example in the 2010 plot the coordinate (160, -60) represents a student who was assigned scores of 160 in the item-pool analysis (SS_{ip} = 160) and 100 in the year-to-year analysis (SS_{yy} = 100). The largest difference between scale scores obtained from the two plans was 98 scale-score points. For these low achieving students, the year-to-year plan seemed to produce much lower score estimates than the item-pool plan. In practice, it is common that only a limited number of students are available at the lowest end of the scale. Thus, obtaining accurate scale-score estimates for low achieving students is generally difficult. That is, measurement errors are typically larger at the extremes of the score scale. The conditional SEM plots in Figure 3 also show that the greatest SEMs were associated with the lowest scale score. Specifically, when the scale score was equal to 100, the SEMs varied between about 105 and 160 scale-score points across

the four years. Further investigation revealed that some of those low performing students had unusual response patterns that are correlated to large conditional SEMs (SEM_{yy} > 138 scale-score points). Because person ability score estimates on the low end of the scale were largely unreliable, the use of different equating plans could significantly affect the scoring of those low achieving students. However, given the use of empirical data in this study, it is difficult to identify whether the discrepancies in scale-score estimates between the two equating plans were due to the design of equating, differences in anchor sets, or other uncontrollable factors. Practitioners in large-scale assessments should keep in mind the difficulty of estimating those low achieving students' abilities and the possible impact of the choice of an equating plan on them.

Achievement Levels

Table 2 shows tabulations of grade-three students in the five achievement levels obtained from the year-to-year and item-pool plans. The two plans performed similarly, which led to high agreement rates at each AL and across years. The total agreement rates were calculated by summing the agreement percentages on the diagonal of each table. The total agreement rates from 2007 to 2010 were 99.8%, 96.9%, 95.5% and 94.4%, respectively. The weighted Kappa values (Cohen, 1968) from 2007 to 2010 were .999, .975, .964, and .955, respectively. Values of Kappa greater than .80 indicate almost perfect agreement (Landis & Koch, 1977). However, the rates of consistent classification by the two plans decreased slightly as we moved farther from the base equating year. As previously noted, at the lower achievement levels (1 and 2) the year-to-year plan produced higher score estimates than the item-pool plan. For example, in 2007, 0.05 percent of third-grade students were classified into level 2 using the year-to-year plan, but into level 1 using the item-pool plan. However, no students were classified into level 1 using year-to-year scores and level 2 using item-pool scores. At the higher achievement levels 4 and 5, the year-to-year plan tended to produce lower score estimates, therefore leading to more students being classified into lower achievement levels. The percentages of inconsistently classified students increased slightly in later years.

It should be noted that in large-scale testing programs, a small percentage such as one percent could imply relatively a big chunk of students (e.g., approximately 2000 students in each grade in the current study). Thus, the choice of equating plans could have impact on those students (e.g., those students in one percent who may be classified as not reaching the specified state performance level) who would be classified differently depending on the choice of an equating plan. As the gap in achievement classifications by the two equating plans increased though small in later years (i.e., the inconsistent classification rates increased from 0.2% in 2007 to 5.6% in 2010), the potential impact on students and stakeholders would need to be taken into consideration when determining the equating plan to be used in large-scale assessments.

Table 2. Counts (and Percentages) of Students in Five Achievement Levels Obtained from Scores based on Year-to-Year vs. Item-pool Plans for Grade 3.

		Item-Pool											
		2007						2008					
	Level	1	2	3	4	5	Row Total	1	2	3	4	5	Row Total
Year-to-year	1	**24544**	0	0	0	0	24544	**20319**	0	0	0	0	20319
		(12.16)					(12.16)	**(9.95)**					(9.95)
	2	99	**29723**	0	0	0	29822	404	**27353**	399	0	0	28156
		(0.05)	**(14.72)**				(14.77)	(0.20)	**(13.40)**	(0.20)			(13.79)
	3	0	85	**67492**	68	0	67645	0	0	**66937**	3010	0	69947
			(0.04)	**(33.43)**	(0.03)		(33.51)			**(32.78)**	(1.47)		(34.26)
	4	0	0	0	**55419**	107	55526	0	0	0	**57128**	2603	59731
					(27.45)	(0.05)	(27.51)				**(27.98)**	(1.27)	(29.25)
	5	0	0	0	0	**24325**	24325	0	0	0	0	**26027**	26027
						(12.05)	(12.05)					**(12.75)**	(12.75)
	Column Total	24643	29808	67492	55487	24432	201862	20723	27353	67336	60138	28630	204180
		(12.21)	(14.77)	(33.43)	(27.49)	(12.10)	(100)	(10.15)	(13.40)	(32.98)	(29.45)	(14.02)	(100)
		2009						2010					
	Level	1	2	3	4	5	Row Total	1	2	3	4	5	Row Total
Year-to-year	1	**18292**	0	0	0	0	18292	**16670**	0	0	0	0	16670
		(8.92)					(8.92)	**(8.10)**					(8.10)
	2	1442	**25066**	0	0	0	26508	1946	**24276**	0	0	0	26222
		(0.70)	**(12.22)**				(12.92)	(0.95)	**(11.80)**				(12.75)
	3	0	845	**64826**	3200	0	68871	0	1624	**64479**	3078	0	69181
			(0.41)	**(31.60)**	(1.56)		(33.57)		(0.79)	**(31.34)**	(1.50)		(33.63)
	4	0	0	0	**58577**	3834	62411	0	0	0	**59529**	4869	64398
					(28.56)	(1.87)	(30.42)				**(28.94)**	(2.37)	(31.31)
	5	0	0	0	0	**29053**	29053	0	0	0	0	**29240**	29240
						(14.16)	(14.16)					**(14.21)**	(14.21)
	Column Total	19734	25911	64826	61777	32887	205135	18616	25900	64479	62607	34109	205711
		(9.62)	(12.63)	(31.60)	(30.12)	(16.03)	(100)	(9.05)	(12.59)	(31.34)	(30.34)	(16.58)	(100)

Note. Level = achievement level. Numbers in bold type are the consistent classification.

Summary and Discussion

Linkage plans should be tailored to meet practical needs. Considering impact of varying factors (e.g., update of an item pool, change in anchor items) on equating results in practice, it is important to evaluate the empirical performance of different linkage plans in educational assessment programs. This study investigated the performance of item-pool and year-to-year equating linkage plans using large-scale assessment data for grades 3, 6, 7, and 9 from 2007 to 2010. Despite the frequent use of these two equating plans in state assessment programs, a real data investigation of the two plans based on large-scale assessments has not been seen frequently in the literature. Based on results from the present study, the two equating plans seemed to give slightly increasing different results based on a four-year evaluation. The item-pool equating plan seemed to be more discriminating by giving a higher ability student an increasingly higher scores, whereas giving a lower ability student increasingly lower scores. However, the differences in equating results between the two plans appeared to be very small in terms of each criterion evaluated, and may not have considerable practical consequences for score reporting in educational practice. Based on the four-year evaluation in our study, the year-to-year equating plan might be considered as an alternative to the item-pool equating plan when the equating chain is not too long (our evaluation of four-year equating showed little differences between the two linkage plans).

The item-pool equating calibrates item parameter estimates to an established item pool, whereas the year-to-year equating calibrates item parameter estimates to common items in an existing test form from one year prior. Item-pool equating has been widely used in large-scale assessments, in part because more freedom can be exercised in the selection of anchor items in the item-pool than in the year-to-year plan. However, the cost of having readily available anchor items is the added time and expense of creating and maintaining the item pool. In practice, the year-to-year plan may be easier to apply and cost effective, because it does not require the construction of an item pool. However, this plan should be used with caution because equating results only depend on the prior year's parameter estimates for anchor items. Problems such as scale shifts (Guo et al., 2011) can occur in successive equating years due to systematic equating errors or other errors not corrected in earlier years. In the present study, although slightly greater dispersion of equating results was observed when more years were included in the equating chain, the differences were too small to conclude that there were practically significant differences between the two plans. This would suggest that, with a relatively short equating chain (i.e., four years such as in this study), year-to-year equating could be a suitable alternative to item-pool equating. Until more research has been conducted to establish scale stability for a series of equating links (e.g., Zhang, McDermott, Fantuzzo, & Gadsden, 2013), we recommend that researchers interested in year-to-year equating always select anchor items for equating cautiously and use this linkage plan only in situations where the number of equating links is not large.

Because of the constraints in the selection of anchor items in the year-to-year plan, this resulted in an equal or lower numbers of anchor items in the year-to-year plan compared to the item-pool plan. A noted limitation of this study is the potential confounding between the unequal numbers of anchor items and the actual difference of the two linkage plans. This caused the difficulty to identify whether the differences in the S-L transformation coefficients and equating results were due to the two linkage plans or because the anchor sets used in the two plans were different in the number of anchor items, their content, and statistical properties. In the current study, it is nearly impossible to select an identical anchor item set for the two linkage plans, due to the way empirical data were collected, periodical maintenance of the item pool, and other practice issues. In a

study by Kim and Cohen (1998), equating outcomes, such as linking coefficients, parameter estimates, and root mean square differences, were estimated more accurately with a larger number of common items. Therefore, these equating results should be interpreted with caution.

As an investigation of real assessment data, our observations and the implications drawn are restricted to data from assessment programs similar to the one examined in this study. The many complex issues involved in operational test-score equating in practice makes it difficult to conduct a simulation study that considers every issue in test equating. For example, updating an item pool is regularly done in educational practice. However, it is not easy to simulate an item pool containing a large number of calibrated items that also incorporate necessary parameter updates and changes to items after each equating process. In addition, constructing anchor item sets comparable for the two linkage plans is difficult, given that the two plans follow different linking procedures. For these reasons, the scope of our comparison of the two equating plans was limited to empirical data. However, the data sets were from a large-state testing program and the results from the current comparison could benefit practitioners involved in similar large-scale state assessment programs.

One should be cautious about the potential scale shrinkage in complex equating linkage plans. Scale shrinkage means that, as the number of equating links increases, the standard deviations of scale scores consistently decreases, leading to scale scores regressed to the mean. In the current study, we noticed a decrease in standard deviations of scale scores from 2007 to 2010 for grade 3 in the year-to-year plan and for grade 6 in the item-pool plan. This scale-shrinkage effect seemed not to be linkage-plan specific, and was not observed in other grades examined. Prior research has indicated that scale shrinkage can occur when equating is conducted within and across grades (Camilli, 1988, 1999; Yen, 1985). The shrinkage phenomenon may relate to academic growth (Burket, 1984), or measurement errors (Camilli, 1988), or both. It is unclear whether the scale shrinkage is an artifact of the equating plans, and whether the increase in the number of equating links would make the shrinkage effect more striking. Further investigation of this issue is needed, but meanwhile empirical researchers and psychometricians should be aware of possible scale shrinkage and its potential influence when implementing test scaling and equating in practice.

Given the rarity of research using real large-scale assessment data to compare the year-to-year and item-pool equating plans, our study presents a unique opportunity to observe real performance of the two equating plans for a state-wide testing program. However, this case study is limited in terms of the conditions examined. Our investigation included four years of mathematics test data, which did not allow us to examine effects of a longer chain on equating results. If we had included more years of data, the comparison between the two equating plans might show larger differences. Furthermore, results based on only mathematics assessments might not be generalizable to other subjects or ability constructs. While other content subjects merit further examination, current discussion on mathematics assessments provides useful information for future research. Complementing what was learned in this study, a well-constructed simulation study can help disentangle the effects of critical factors in equating such as the effect of anchor items, scale shrinkage, unusual scale-score patterns, and more. Future simulation studies could also provide opportunities to investigate the effects of the violations of the assumptions (e.g., model-data fit, multidimensionality, etc.) in IRT equating plans.

References

Arai, S., & Mayekawa, S.-i. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika, 38*(1), 1-16.

Battauz, M. (2013). IRT test equating in complex linkage plans. *Psychometrika, 78*(3), 464-480.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285.

Burket, G. R. (1984). Response to Hoover:" The most appropriate scores for measuring educational development in the elementary schools: GE's". *Educational Measurement: Issues and Practice, 3*(4), 15-16.

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational and Behavioral Statistics, 13*(3), 227-241.

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement, 36*(1), 73-78.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213-220.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*: Lawrence Erlbaum.

Cowell, W. R. (1991). *A procedure for estimating the conditional standard errors of measurement for GRE general and subject tests* (GRE board report No. 87-03P). Retrieved from http://www.ets.org/Media/Research/pdf/RR-91-25-Cowell.pdf

Educational Testing Service. (2009). *California Standards Tests technical report spring 2008 administration*. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/cxcsttechrpt08.pdf

Florida Department of Education. (2006). *Reading and mathematics technical report for 2006 FCAT test administrations*. Retrieved from http://fcat.fldoe.org/pdf/fc06tech.pdf

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*(4), 369-377.

Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika, 75*(3), 438-453.

Guo, H., Liu, J., Dorans, N., & Feigenbaum, M. (2011). *Multiple linking in equating and random scale drift*. Retrieved from http://www.ets.org/Media/Research/pdf/RR-11-46.pdf

Haberman, S., & Dorans, N. J. (2009). *Scale consistency, drift, stability: Definitions, distinctions and principles.* Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, San Diego, CA.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-30). New York, NY: Springer.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*(2), 147-154.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd Ed.). New York: Springer.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational Measurement, 19*(4), 279-293.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics, 33*(1), 159-174.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38-60.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational and Behavioral Statistics, 8*(2), 137-156.

Puhan, G. (2008). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education, 22*(1), 79-103.

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3*(4), 365-384.

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

Tate, R. L. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and psychological measurement, 63*(6), 893-914.

Thissen, D. (2003). MULTILOG (Version 7.03). Chicago: Scientific Software International.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*(4), 333-344.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement, 35*(3), 185-198.

Von Davier, A. A. (2011). Quality control and data mining techniques applied to monitoring scaled scores. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.), *Proceedings of the 4th international conference on educational data mining*. Eindhoven, The Netherlands.

Wang, L., Qian, J., & Lee, Y. H. (2013). Exploring Alternative Test Form Linking Designs With Modified Equating Sample Size and Anchor Test Length. *ETS Research Report Series, 2013*(1), i-17.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*(4), 399-410.

Zhang, X., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2013). Longitudinal stability of IRT and equivalent-groups linear and equipercentile equating. *Psychological Reports, 113*(1), 291-313.

Corresponding Author:
Xinya Liang
Educational Statistics and Research Methods
University of Arkansas, Fayetteville
AR 72701.
Email: xl014@uark.edu,
Fax: 479-575-3319
Phone: 479-575-7948