Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 36

Factors Affecting Accuracy of Comparable Scores for Augmented Tests under Common Core State Standards

Ja Young Kim[†]

Won-Chan Lee

Deborah J. Harris

August 2013

[†]Ja Young Kim is Research Associate, ACT, Inc., 500 ACT Drive, P.O. Box 168, Iowa City, IA 52243 (email: jayoung.kim@act.org). Won-Chan Lee is Associate Professor, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu). Deborah J. Harris is Chief Research Scientist, ACT, Inc., 500 ACT Drive, P.O. Box 168, Iowa City, IA 52243 (email: deborah.harris@act.org).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

All rights reserved

Contents

List of Tables

List of Figures

Abstract

This paper examined various factors affecting linking accuracy of the augmented tests containing both common assessments and state-specific item sets in the context of the Common Core State Standards (CCSS). Linking was conducted using the common-item nonequivalent groups design with pseudo and simulated data. The factors included group ability differences, differential effect sizes, latent-trait correlations, differential latent-trait correlations between the common assessment and state-specific item set, and different equating methods (classical and IRT equating). Higher latent-trait correlation between the common assessment and state-specific item set was associated with smaller bias, and if the latent-trait correlation exceeded 0.8, IRT equating methods provided adequate linking even though the group ability difference was large and the differential effect size was moderately large. The same latent-trait correlations for the old and new tests resulted in smaller bias than differential latent-trait correlations for the old and new tests.

1 Introduction

The U.S. Department of Education has awarded Race to the Top grants to two state-led consortia: the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC), to develop assessments based on the Common Core State Standards (CCSS). A common assessment that measures the CCSS must provide results on the common standards that are comparable across states. If the common assessment does not cover state-specific standards, states are allowed to add up to an additional 15% of content or items (Achieve, 2010; Cizek, 2010). Therefore, it is likely that at least some states may design augmented tests that contain both the common assessment and state-specific items (Lazer, Mazzeo, Way, Twing, Camara, & Sweeney, 2010). Currently, 11 states plan to add state-determined standards to the CCSS (Center on Education Policy [CEP], 2012).

Large scale assessments require multiple forms that are similar in content and statistical specifications to maintain test security over time. In order to use scores obtained from alternate forms interchangeably, the reported scores from each form should be on a common scale which is usually accomplished by a process called test equating and linking. However, it is complicated to link the augmented tests that include both common assessment and state-specific item sets across different administrations within a state and to concord them across different states. This is because the state-specific item set is likely to vary across states and the items could change depending on how much the CCSS overlap the state-specific standards. Furthermore, students will perform differently on the common assessment and state-specific item set because the two parts of the augmented test are developed based on different standards; the common assessment is developed based on the CCSS while the state-specific item set is constructed based on state standards. At the same time, dimensionality issues also would exist because the augmented test measures two different standards.

Given the context above, this study was aimed to investigate the effect of the relationship between the common assessment and state-specific item sets on linking accuracy under different characteristics of examinees in proficiency on the two parts of the augmented tests using different equating methods.

2 Theoretical Framework

2.1 Definition of Equating and Linking

Kolen and Brennan (2004) refer to equating as a statistical process of adjusting scores to account for small differences in difficulty across test forms. Score equating is crucial for any large scale assessment programs that require multiple test forms in order to obtain comparability of scores on test forms. To conduct equating, the test forms should be similar in difficulty and content. Linking is an alternative to equating when two different tests are developed based on different specifications (Kolen & Brennan, 2004). If the states want to compare scores

on augmented tests that include both the common assessment and the state-specific item sets, linking generally may be used since content specifications for the state-specific item sets may differ over time depending on the content covered in the common assessment. There are three popular data collection designs for equating and linking: single group design, random groups design, and common item non-equivalent groups (CINEG) design. This study focuses on the CINEG data collection design, in which two test forms have a set of items in common, which are administered to different samples from different populations (Kolen & Brennan, 2004).

The U.S. Department of Education requires comparability of scores on the assessment across states within each consortium (Oregon Department of Education, 2011). To maintain comparability of scores on the common assessment, common items or common examinees across different test forms would be required. Since the PARCC proposed a plan to use multiple fixed test forms across years, grades, and states, common item links could be used to link the PARCC test forms (Luecht & Camara, 2011). Therefore, the CINEG design was used as the data collection method for the common assessment in this study. Note that the state-specific item sets would vary depending on the degree to which the CCSS are similar to the state-specific standards, whereas the common assessments measure the same CCSS and have similar statistical specifications, across years, grades, and states. Therefore, the common assessment has the same or parallel test forms, but the state-specific item sets are more varied.

2.2 Issues for Linking Augmented Tests in the CINEG Design

The CINEG design assumes that common items should represent the content and statistical characteristics of a total test. If a common item set does not represent the characteristics of the total tests, it is difficult to obtain accurate linking relationships under the CINEG design. Klein and Jarjoura (1985) demonstrated that the content representation in the common items is critical when non-equivalent groups perform differently with respect to various content areas covered in the full-test forms. In the context of the CCSS, if the state-specific items are not similar to the common assessment with respect to content, statistical, and item type characteristics, the common items on the common assessment would not accurately reflect the total augmented test characteristics.

Therefore, it seems crucial to investigate how different degree of dissimilarity in the construct (latent traits) measured by the common assessment and state-specific item set affects linked scores across the augmented tests. Note that the common assessment does not necessarily mean the same test form. It could be the same single form, but it most likely consists of parallel forms with a set of common items. Also, characteristics of examinees would have impacts on equating and linking results. Dorans, Liu, and Hammond (2008) investigated the effects of characteristics of populations on equating results. They observed consistent results across different equating methods with small ability differences and divergent results across the equating methods with large abil-

ity differences. Given the previous research, examinee characteristics such as examinee proficiency on the common assessment and the state-specific item set also can influence the linked scores of the augmented tests.

3 Method

3.1 Data Source and Analysis

Large scale math and science tests were used in this study. Since no data were available for the augmented tests, the math test was considered a common assessment, and the science test was assumed to be a state-specific item set. Therefore, an augmented test used in this study consisted of the common assessment from the math test items and state-specific item set from the science test items. The reason for using the math and science tests to construct the augmented test was that the math and science tests were reasonably correlated (disattenuated correlation = 0.87), while the two tests measured different standards. Both the common assessment and state-specific item set consisted of multiple-choice items. Pseudo and simulated data were used in this study.

For pseudo data, as shown in Figure 1, the 60-item math test was split in half to make two parallel common assessments with 40 common items[1] and 10 unique items for each of the common assessments. Among the 40 common items, 20 common items were used as linking items. Two sets of nine state-specific items were selected from the science test and augmented to the corresponding common assessments to construct two augmented tests, New Test and Old Test. As mentioned earlier, states that adopted the CCSS are allowed to add a maximum of 15% of the items to address state standards, which implies that the common assessments must comprise at least 85% of the total test. In this study, the total number of items in each augmented test is 59: 85% of the augmented test is the common assessment (n = 50) and the remaining 15% is the state-specific item set (n = 9). Examinees taking New Test and Old Test are from the examinee groups of P (New Test group) and Q (Old Test group), respectively. The group Q taking Old Test was set as the reference group. The combined group (P and Q) took both New Test and Old Test, so examinee groups of P and Q were used as the sample to link New Test and Old Test. The sample size for each of the groups P and Q is 1000.

[1] The 40 common items were used as a common-item pool to manipulate different conditions of the number of common items considered in the full study design (see Kim, 2013). In the current study, the number of common items was not considered as a study factor since the results for this factor were obvious (i.e., the larger number of common items always performed better). Therefore, only the reuslts for the 20-common item condition are reported here.

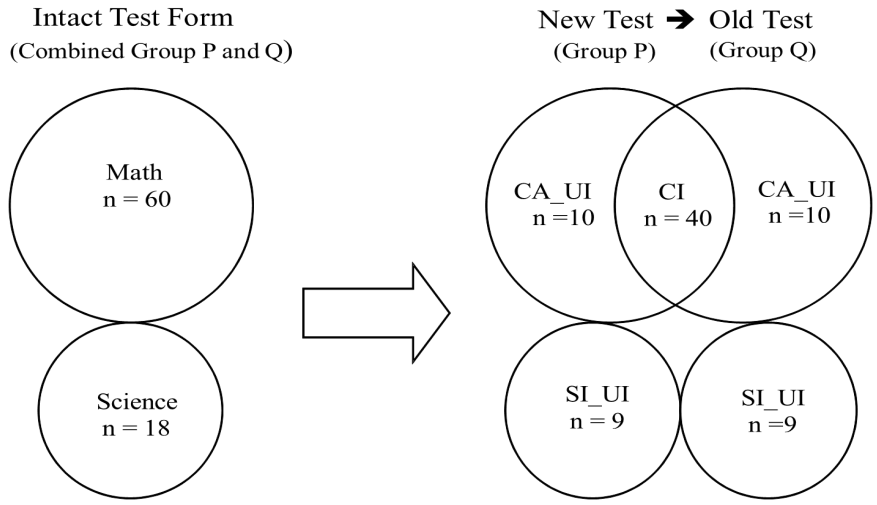

Figure 1: Construction of Augmented Tests

Note. New Test taken by the examinee group P and Old Test taken by the examinee group Q. CA = Common Assessment; SI = State-specific Items; n = Number of Items; CI = Common Items; UI = Unique Items.

Simulated data were used to manipulate different latent-trait correlation levels between the common assessment and state-specific item sets. The reason for using simulated data in addition to the pseudo data is that it is hard to obtain different levels of latent-trait correlations with only the pseudo data in which the latent-trait correlations, as measured by disattenuated correlations, ranged from 0.8 to 0.9. The simulation process involved four steps. First, pairs of theta values (N = 3000) were drawn from a bivariate normal distribution for New Test and Old Test populations. True item parameters for Old Test were the same as the pseudo test item parameters. However, true item parameters for New Test were not the same as the pseudo test item parameters because in the pseudo-data analyses, New Test and Old Test shared 40 common items even though only 20 were used as linking items. Therefore, to manipulate the smaller number of common items, the true item parameters for New Test were selected from additional test items. Second, item responses for each test were generated based on the true item parameters and examinees theta values. The three parameter logistic model (3PL) was fitted simultaneously to estimate item parameters for the common assessment items and the state-specific items. Third, linking was conducted to obtain linked scores. Fourth, the above steps were repeated 50 times. The composite raw scores for each set of data are from 0 to 59 (i.e., 60 score points), which are the summed number correct scores of

the common assessment and state-specific item set. The overall statistics are weighted by frequencies for Old Test.

3.2 Study Factors

Five factors were considered in this study: group ability difference (measured by common-item effect size), differential effect size, latent-trait correlations between the common assessment and state-specific item set, differential latent-trait correlations, and different equating methods. In the pseudo-data analyses, common-item effect size, differential effect size, and different equating methods were investigated. In the simulated data analyses, all the five factors were examined with more focus on the latent-trait correlations and differential latent-trait correlations.

3.2.1 Common-item Effect Size (CES)

The group ability difference was measured by Common-item Effect Size (CES). Different conditions of the common-item effect size were created using a sampling process with the external variables including students' number correct score in a large scale reading test and state of residency. Three levels of the common item effect size were created: 0.0, 0.2, and 0.4. The common-item effect size represents a standardized mean difference in common item scores between New Test and Old Test. The effect size was calculated for New Test minus Old Test common item scores as follows:

$$\text{CES} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2}}}, \tag{1}$$

where $\overline{X}$ is the mean for the common item scores. N is the number of examinees, and s^2 is for the variance of scores. New Test and Old Test are indicated by 1 and 2, respectively.

3.2.2 Differential Effect Size (DES)

The effect size based on the common items that were obtained only from the common assessment does not represent student performance on the state-specific item set. Therefore, it is problematic to use only the common-item effect size as an index of group differences. It is likely that examinees perform differently on the common assessment and state-specific item set because the two parts of the augmented test are developed based on different standards: The common assessment is developed based on the CCSS, and the state-specific item set is constructed based on state standards. Therefore, it would be reasonable to examine differential effect sizes of the common assessment and state-specific item set for linking the augmented tests. The differential effect size between the common assessment and state-specific item set can be calculated for the effect

size of the common assessment minus the effect size of the state-specific item set as follows:

$$\text{DES} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(N_1-1)s^2_{X1}+(N_2-1)s^2_{X2}}{N_1+N_2}}} - \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{(N_1-1)s^2_{Y1}+(N_2-1)s^2_{Y2}}{N_1+N_2}}}, \tag{2}$$

where $\overline{X}$ and $\overline{Y}$ are the means for the common assessment scores and state-specific item scores, respectively; s^2_X and s^2_Y are the variances of the common assessment scores and state-specific item scores, respectively; N is the number of examinees; and New Test and Old Test were indicated by 1 and 2, respectively. Two levels of the differential effect size were created: 0.0 and 0.2.

Figure 2 contains a visual illustration for the combination of the common-item effect size and the differential effect size conditions. The bigger circle represents the common assessments (CA) and the smaller circle represents the state-specific item sets (SI). The standardized mean is represented as SM. [CES 0.2 DES 0.0] represents that the standardized mean based on the common item score for New Test group is 0.2 points higher than the standardized mean based on the common item score for Old Test group. Also, the standardized mean based on the state-specific item score for New Test group is also 0.2 points higher than the standardized mean based on the state-specific item score for Old Test group. In this situation, the differential effect size (DES) is 0.0 because the common item effect size and the state-specific item effect size are the same.

Similarly, [CES 0.2 DES 0.2] represents that the standardized mean based on the common item score for New Test group is 0.2 points higher than the standardized mean based on the common item score for Old Test group. However, the standardized mean based on the state-specific item score for New Test group is 0.4 points higher than the standardized mean based on the state-specific item score for Old Test group. Under this situation, DES is 0.2 because the difference in the effect sizes for the common item effect size and the state-specific item effect size is 0.2.

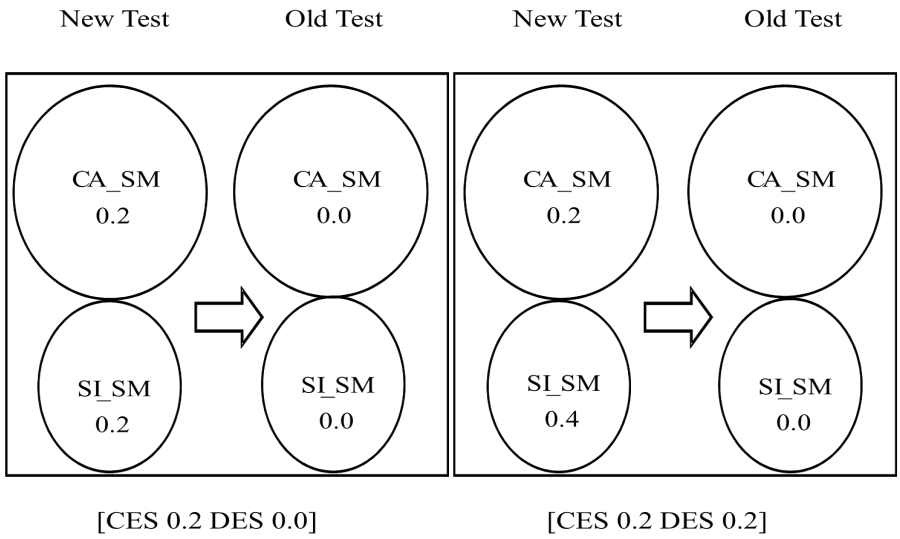

Figure 2: Visual Explanation of CES and DES.

Note. CA_SM = Standardized mean on common assessment; SI_SM = Standardized mean on state-specific item set.

3.2.3 Latent-Trait Correlations between Common Assessment and State-Specific Item Set

Five conditions of the latent-trait correlations between the common assessment and state-specific item set were included: $r = 0.6$, $r = 0.7$, $r = 0.8$, $r = 0.9$, and $r = 0.98$. The conditions of the latent-trait correlation were determined based on the median value of the disattenuated correlation between math and science raw scores for six national achievement tests administered in the 2005-2006 period, which was generally 0.8. Therefore, 0.8 was set as a mid-level of the latent-trait correlation values in this study.

3.2.4 Differential Latent-Trait Correlations between Common Assessment and State-Specific Item Set

The latent-trait correlations between the common assessment and state-specific item set would not always be the same across different augmented tests because state-specific item sets would vary depending on state objectives and standards covered by the common assessment. Specifically, if a state within a consortium adds different state-developed items over time, the latent-trait correlations between the common assessment and state-specific item would differ across years. Therefore, it seems reasonable to investigate how the differential latent-trait correlations across different augmented tests would affect the linking accuracy.

Also, the state-specific standards usually have been added to the CCSS within the same subject area. Therefore, the latent-trait correlations between the common assessment and state-specific item set would be reasonably correlated in real situations (i.e., two English tests); and, of the five values of the latent-trait correlations described earlier (0.6, 0.7, 0.8, 0.9, and 0.98), only the highest pairs of values using $r = 0.8$, $r = 0.9$, and $r = 0.98$ were included in the differential latent-trait correlation conditions. Four levels of the differential latent-trait correlations were included: [r(new) = 0.8 r(old) = 0.9], [r(new) = 0.9 r(old) = 0.8], [r(new) = 0.9 r(old) = 0.98], and [r(new) = 0.98 r(old) = 0.9]. r(new) and r(old) indicate the latent-trait correlations between the common assessment and state-specific item set for New Test and Old Test, respectively.

3.2.5 Equating Methods

For each of the conditions mentioned above, linking was conducted using the CINEG design with frequency estimation (FE), chained equipercentile equating (CE), IRT true score equating (TS), and IRT observed score equating (OS) methods. Linking was conducted using Equating Recipes (Brennan, Wang, Kim, & Seol, 2009) and PIE (Hanson, Zeng, & Cui, 2004). Synthetic weights of one were given to the examinee group taking New Test in order to construct a synthetic population. Before linking, items in New Test and Old Test were calibrated using BILOG-MG (Zimowski, Muraki, & Mislevy, 2003). Items from the two tests were put on the same scale using the Stocking and Lord (Stocking & Lord, 1983) scale transformation method.

3.3 Evaluation

For the pseudo-data analyses, the criterion to evaluate the linking accuracy of the augmented tests was established by using a single-group equating method. The rationale for using a single group equating as a criterion linking relationship is that two sources of linking errors –group ability differences and the use of common items as opposed to full-length tests to determine the linking relationship– are not involved in the single group equating. Different criteria for the linking relationships were used for different equating methods. A single group equipercentile equating relationship without smoothing was used as a criterion for the two classical equating methods: frequency estimation and chained equipercentile equating. Further, single group IRT observed and true score equating relationships were used as criteria for the IRT observed and true score equating methods, respectively.

For the simulated data analyses, true linking relationships were obtained using simple structure multidimensional IRT observed-score equating (Lee & Brossman, 2012) based on the true item parameters and a population bivariate normal distribution for the two latent variables of the common assessment and state-specific item set. More specifically, the latent-trait variables for the common assessment (represented as 1) and the state-specific item set (represented as 2) are denoted as θ_1 and θ_2, and they follow the bivariate normal distribution

denoted as BN (μ_1, μ_2, σ_1, σ_2, r). The terms μ_1, μ_2, σ_1, and σ_2 represent the means and standard deviations for each of the two latent-trait variables of θ_1 and θ_2, respectively, and r stands for the latent-trait correlation between θ_1 and θ_2. Different criteria of linking relationships were used for different population distributions. More specifically, based on Study Factors 3.2.1 through 3.2.4, a total of 120 (3 x 2 x 5 x 4) pairs of population distributions for New Test and Old Test were considered in the simulation study. Therefore, 120 criteria of linking relationships were used for the corresponding population distributions of the simulated data analyses. This study assumes that each of the common assessment and state-specific item sets measure a single latent-trait variable, and the two latent-trait variables measured by the common assessment and state-specific item set are correlated. A total of 41 x 41 quadrature points and weights were created for the population bivariate distributions. Theta values range from -4 to 4 for each of the common assessment and state-specific item sets.

In order to conduct the simple structure multidimensional IRT observed-score equating, a fitted marginal observed-score distribution was obtained for each of New Test and Old Test populations using the bivariate normal distribution and true item parameters. First, using an extended version of the Lord-Wingersky algorithm (Lord & Wingersky, 1984; Hanson, 1994; Thissen, Pommerich, Billeaud, & Williams, 1995), conditional raw-score distributions were computed for each pair of theta values. Then, the marginal observed-score distribution was obtained for each test by aggregating the conditional distributions over the entire bivariate theta distribution. Using the marginal observed-score distributions for the two tests, equipercentile equating was conducted on a synthetic population with a weight of one on New Test group. The indices for evaluating the linking results from the pseudo and simulated data analyses are bias (Bias), standard error (SE), and root mean squared error (RMSE). These summary statistics were calculated for each of the raw score points (i.e., conditional statistics) and across all score points (i.e., overall statistics). Conditional statistics are shown in Equations 3, 4, and 5.

$$Bias(x_i) = \sqrt{\left[\frac{\sum_{j=1}^{J} \hat{e}_j(x_i)}{J} - e(x_i)\right]^2} \tag{3}$$

$$SE(x_i) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left[\hat{e}_j(x_i) - \left(\frac{1}{J}\sum_{j=1}^{J}\hat{e}_j(x_i)\right)\right]^2} \tag{4}$$

$$RMSE(x_i) = \sqrt{\left[\frac{\sum_{j=1}^{J} \hat{e}_j(x_i)}{J} - e(x_i)\right]^2 + \frac{1}{J}\sum_{j=1}^{J}\left[\hat{e}_j(x_i) - \left(\frac{1}{J}\sum_{j=1}^{J}\hat{e}_j(x_i)\right)\right]^2} \tag{5}$$

In Equations 3 through 5, J is the number of replications (10 for the psedo data and 50 for the simulated data); $\hat{e}_j(x_i)$ is an estimated equated score at score point i; and $e(x_i)$ is a criterion equated at score point i. The overall statistics include the weighted average root mean squared bias, Wbias $= \sqrt{\sum_i w_i Bias(x_i)^2}$, the weighted average standard error, WSE $= \sqrt{\sum_i w_i SE(x_i)^2}$, and the weighted average RMSE, WRMSE $= \sqrt{\sum_i w_i RMSE(x_i)^2}$. In addition to the summary statistics mentioned above, an index of an acceptable level of error called a Difference That Matters (DTM) was used to determine a practically acceptable level of the error for adequate linking in the simulated data analyses. The DTM value of 0.5 was used. The value of 0.5 for DTM is determined based on the rationale such that if the difference between rounded scale scores is greater than 0.5, the reported scale scores could be changed (Dorans & Feigenbaum, 1994). In this study, the adequacy of linking was determined by comparing the DTM to $Bias(x_i)$. If the $Bias(x_i)$ was greater than the DTM, the linking was considered unacceptable.

4 Results

Descriptive statistics for the pseudo tests are provided in Table 1. The New Test group showed slightly higher means for the common assessment, state-specific item set, and combined test than did the Old Test group. Generally, the descriptive statistics for New Test and Old Test were similar, indicating two pseudo tests were equivalent. Table 2 presents descriptive statistics for item parameters of New Test and Old Test for the simulated data analyses. Old Test has higher means for discrimination, difficulty, and guessing parameters than did New Test.

Table 3 contains the averaged summary statistics for the effects of common-item effect size and differential effect size. The pseudo-data and simulated data analyses showed the same pattern of the results for the effect size conditions, therefore, only the results from the simulated data analyses were presented in this section. The statistics were averaged over the latent-trait correlation conditions and equating methods. CES 0.0 indicates that Old Test and New Test groups are equivalent. CES 0.2 is for moderately large group ability difference, and CES 0.4 is for large group ability difference. DES 0.0 is almost no differential effect size, and DES 0.2 is moderately large differential effect size, respectively. As indicated in Table 3, WBIAS and WRMSE increased as the group ability difference increased, while WSE seemed similar across different group ability conditions. Also, larger WBIAS and WMSE were associated with the larger differential effect size. The values of WSE were similar when the differential effect sizes were 0.0 and 0.2.

Table 1: Descriptive Statistics for Single Group Pseudo-Data Analyses

		New Test	Old Test
Score	N	10,105	10,105
Common Assessment	scale	0 - 50	0 - 50
	Mean	25.378	25.280
	SD	0 .011	9.821
	Skew	0.232	0.257
	Kurt	-0.714	-0.691
State-specific Item Set	sclae	0 - 9	0 - 9
	Mean	4.974	4.073
	SD	2.034	1.967
	Skew	0.069	0.069
	Kurt	-0.619	-0.478
Combined	Scale	0 - 59	0 - 59
	Mean	30.252	29.954
	SD	11.399	11.146
	Skew	0.236	0.269
	Kurt	-0.702	-0.658
Common Item	Mean	9.944	9.944
(n = 20)	SD	4.182	4.182
	Skew	0.217	0.217
	Kurt	-0.678	-0.67

Table 2: Descriptive Statistics for Item Parameters for New Test and Old Test in Simulated Data Analyses

	New Test			Old Test		
	a	b	c	a	b	c
Mean	0.954	0.259	0.167	1.062	0.414	0.179
Median	0.865	0.491	0.146	1.002	0.349	0.163
SD	0.362	1.120	0.071	0.327	0.881	0.076

Note. a = Discrimination; b = Difficulty; c = Guessing.

Table 3: Averaged Summary Statistics for Group Ability Difference and Differential Effect Size

		Statistics		
Differential Effect Size	Group Ability Difference	WBIAS	WSE	WRMSE
DES 0.0	CES 0.0	0.056	0.220	0.230
	CES 0.2	0.159	0.230	0.285
	CES 0.4	0.289	0.225	0.378
DES 0.2	CES 0.0	0.266	0.223	0.350
	CES 0.2	0.415	0.225	0.473
	CES 0.4	0.552	0.227	0.599

Table 4 presents the summary of WBIAS to compare the effects of group ability difference and differential effect size depending on different equating methods. [CES 0.0 DES 0.2] indicates a smaller value of the group ability difference associated with a larger value of differential effect size. [CES 0.2 DES 0.0] is for a larger value of the group ability difference associated with a smaller value of differential effect size. If larger bias was observed for [CES 0.0 DES 0.2] than for [CES 0.2 DES 0.0], the effect of DES would be larger than the effect of CES. Conversely, if [CES 0.2 DES 0.0] yielded larger bias compared to [CES 0.0 DES 0.2], the effect of CES would be larger than the effect of DES. Also, the results from the simulated data with $r = 0.8$ were selected to compare the pseudo-data results because the disattenuated correlation between the common assessment and state-specific item set was about 0.8 for the pseudo data.

For the FE method with the pseudo and simulated data, [CES 0.2 DES 0.0] resulted in larger WBIAS and WRMSE compared to [CES 0.0 DES 0.2], indicating that the group ability difference had more effect on the linking results than did the differential effect size. For the CE method with the pseudo data, [CES 0.2 DES 0.0] resulted in similar magnitude of WBIAS to that of WBIAS for [CES 0.0 DES 0.2]. However, for the simulated data, larger WBIAS was observed for [CES 0.0 DES 0.2] than for [CES 0.2 DES 0.0]. When the CE was used, the results were inconclusive for the pseudo-data analyses while a pattern of results was obvious for the simulated data analyses, indicating that the CE was affected more by the differential effect size than by the group ability difference. The OS and TS methods generally resulted in larger WBIAS for [CES 0.0 DES 0.2] than for [CES 0.2 DES 0.0] for both pseudo-data and simulated data analyses. Based on this result, the IRT equating methods seemed more sensitive to the differential effect size than to the group ability difference while the frequency estimation and chained equipercentile methods were affected more by the group ability difference than the differential effect size.

Table 5 presents the summary of averaged statistics for group ability differ-

Table 4: WBIAS for Comparison of Group Ability Difference and Differential Effect Size

		Equating Methods			
Class of Data		FE	CE	TS	OS
Pseudo Data	[CES 0.0 DES 0.2]	0.895	1.014	0.327	0.318
	[CES 0.2 DES 0.0]	1.125	1.014	0.243	0.225
Simulated Data	[CES 0.0 DES 0.2]	0.263	0.264	0.267	0.272
(r= 0.8)	[CES 0.2 DES 0.0]	0.285	0.173	0.084	0.087

ence and equating methods. When groups of examinees were equivalent, all four equating methods perform similarly. However, the FE and CE methods provided larger WSE compared to the TS and OS methods. When group ability differences were moderately large and large, the TS and OS methods resulted in smaller WBIAS and WRMSE compared to the FE and CE methods.

Table 6 presents the summary of averaged statistics for the latent trait correlations between the common assessment and state-specific item set for Old Test and New Test. The magnitudes of WBIAS and WRMSE generally decreased as the latent-trait correlation values increased. The magnitudes of WBIAS and WRMSE were slightly larger for [r(new) = 0.9 r(old) = 0.8] than for [r(new) = 0.8 r(old) = 0.8]. Similarly, the differential latent-trait correlation [r(new) = 0.8 r(old) = 0.9] resulted in larger values of WBIAS and WRMSE than the latent-trait correlation value of [r(new) = 0.8 r(old) = 0.8]. These results seem to imply that if the latent-trait correlations were different for Old Test and New Test, more linking error would occur compared to the situation when the latent-trait correlations were the same for Old Test and New Test.

Table 5: Averaged Summary Statistics for Group Ability Difference and Equating Methods

		Statistics		
Group Ability Difference	Equating Method	WBIAS	WSE	WRMSE
CES 0.0	FE	0.156	0.247	0.307
	CE	0.159	0.273	0.335
	TS	0.175	0.188	0.263
	OS	0.167	0.174	0.255
CES 0.2	FE	0.431	0.251	0.505
	CE	0.308	0.284	0.431
	TS	0.202	0.194	0.294
	OS	0.208	0.180	0.288
CES 0.4	FE	0.709	0.252	0.754
	CE	0.464	0.284	0.552
	TS	0.252	0.191	0.327
	OS	0.257	0.178	0.324

Table 6: Averaged Summary Statistics for Latent-Trait Correlations and Differential Latent-Trait Correlations

	Statistics		
Latent-Trait Correlation	WBIAS	WSE	WRMSE
$[r(\text{old}) = 0.6\ r(\text{new}) = 0.6]$	0.340	0.224	0.429
$[r(\text{old}) = 0.7\ r(\text{new}) = 0.7]$	0.311	0.219	0.403
$[r(\text{old}) = 0.8\ r(\text{new}) = 0.9]$	0.309	0.226	0.398
$[r(\text{old}) = 0.9\ r(\text{new}) = 0.8]$	0.288	0.235	0.387
$[r(\text{old}) = 0.8\ r(\text{new}) = 0.8]$	0.287	0.226	0.389
$[r(\text{old}) = 0.9\ r(\text{new}) = 0.98]$	0.287	0.228	0.380
$[r(\text{old}) = 0.9\ r(\text{new}) = 0.9]$	0.267	0.220	0.368
$[r(\text{old}) = 0.98\ r(\text{new}) = 0.9]$	0.265	0.225	0.363
$[r(\text{old}) = 0.98\ r(\text{new}) = 0.98]$	0.253	0.222	0.358

Figure 3 displays the conditional bias for different levels of the latent trait correlations using the FE method when the group ability is moderately large (CES 0.2). Since patterns of the results for the FE and CE methods were similar,

only the FE results are presented here. There are two subfigures: the left one is for the condition where the differential effect size is almost zero (DES 0.0), and the other is for the condition where the differential effect size is moderately large (DES 0.2). The bold line in the middle of each plot represents the 0.5 DTM criterion. The conditional bias was largest for r= 0.6, and smallest for r= 0.98 over most of the score range. Generally, as the latent-trait correlation increased, bias decreased slightly across different score levels. When comparing the magnitude of bias to the DTM, the lines of the plots for all the five latent-trait correlation conditions were below the DTM line when the differential effect size was almost zero. However, when the differential effect size was moderately large, bias for the latent-trait correlation values tended to exceed the DTM line over most of the score range.

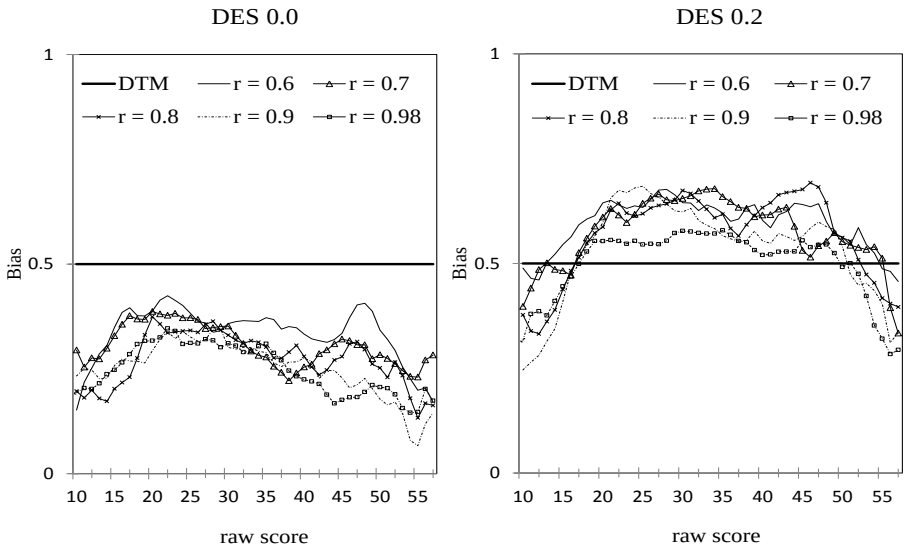

Figure 3: Conditional Bias of Latent-Trait Correlations Using FE for CES 0.2

Figure 4 plots the conditional bias for the latent-trait correlations using the TS method with the large group ability difference (CES 0.4). Since results for the TS and OS methods were similar, only the results from the TS method were discussed here. As shown in the left subfigure for DES 0.0, even though the group ability difference was large, all the latent-trait correlation conditions resulted in adequate linking when the differential effect size was almost zero. However, the latent-trait correlation values below or equal to 0.8 (e.g.,$r = 0.6$, $r = 0.7$, and r= 0.8) did not provide adequate linking in the high score range (40 to 57) when the differential effect size was moderately large. Based on the results

shown in Figures 3 and 4, the moderately large differential effect size would result in inadequate linking for all the values of the latent-trait correlations when the group ability difference was moderately large for the FE method. For TS, the latent-trait correlation values below 0.8 resulted in inadequate linking when the group ability difference was large, and the differential effect size was moderately large.

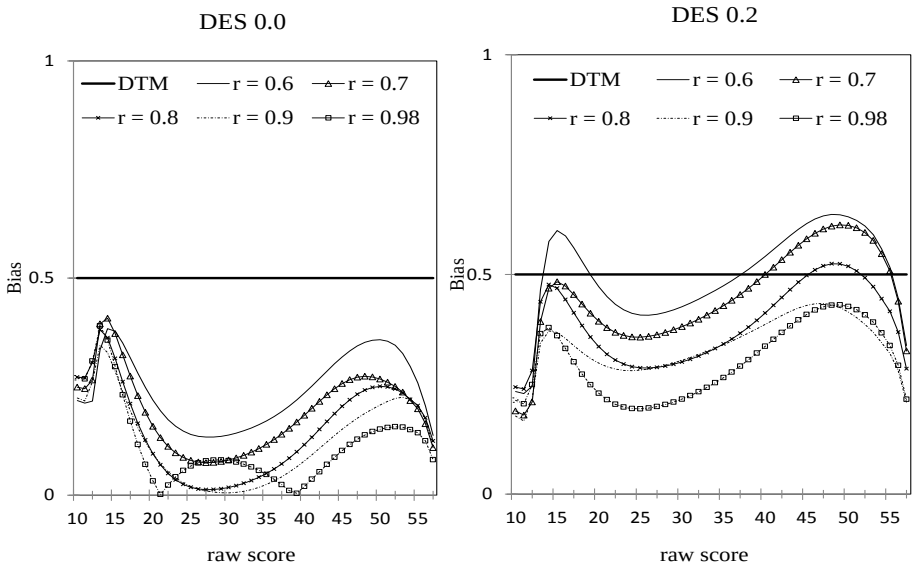

Figure 4: Conditional Bias of Latent-Trait Correlations Using TS for CES 0.4

5 Discussion

When states that adopt the CCSS need to obtain comparable scores on the augmented tests that contain both common assessment and state-specific item sets from different administrations, this study provides several technical guidelines. First, when groups of examinees are equivalent across old and new tests, the choice of a method for equating would not significantly affect results since all four methods (FE, CE, TS, and OS) considered in this study perform similarly. However, the FE and CE methods tend to result in larger standard error of linking compared to the TS and OS methods. In addition, when test groups are equivalent, or differ only slightly, in proficiency, but differential effect size for the common assessment and state-specific item set is large, the FE method might be preferred over the CE, TS, and OS methods. In contrast, when group ability difference is large, but differential effect size for the common assessment

and state-specific item set is small, the CE, TS, and OS methods would be preferred over the FE method. This is because the FE method tends to be more sensitive to the group ability difference than the differential effect size, while the CE, TS, and OS methods are more sensitive to the differential effect size than the group ability difference.

Based on the results in this study, inadequate linking would result when the differential effect size is moderately large under two conditions – moderately large group ability difference with the FE method and large group ability difference with the TS method. However, the latent-trait correlation values above 0.8 result in adequate linking for the TS method even though the differential effect size is moderately large and the group ability difference is large. In addition, when the latent-trait correlation is different for old and new tests, less accurate linking would result compared to the case when the latent-trait correlation is similar for the old and new tests. Therefore, it is recommended that the latent-trait correlation between the common assessment and state-specific item set remains above 0.8. Also, it would be important to keep the 15 percent augmentation guidelines clear and consistent to obtain a similar level of relationship between the common assessment and state-specific item set across years, or even across states within a consortium.

As a limitation, it was difficult to determine whether the differential effect size came from the difference in test form difficulty or the difference in the constructs measured by the common assessment and state-specific item sets. Specifically, when there is a difference in the effect sizes on the common assessment and state-specific item set, there might be several reasons. First, the common assessment and state-specific item sets measure different constructs because they are developed based on different standards. Second, when students have been taught with more emphasis on state-developed standards than the CCSS, the students might perform better on the state-specific item sets than the common assessment. Third, it is possible that the difficulty levels for the common assessment and the state-specific item set might be different. The results for the differential effect size conditions therefore need to be interpreted with caution.

In addition, this study was conducted with only paper-pencil based tests consisting of multiple-choice items. According to Year One Report for PARCC made by the U.S. Department of Education (May, 2012), the PARCC proposed a plan that includes both computer-based innovative item types and extended performance tasks for grades 3-11 in mathematics and English language. Therefore, there might be more complicated psychometric issues in using the innovative item types for the PARCC assessment than in using traditional item types. In this sense, generalizing from the results based upon the current study to the whole CCSS context should be limited. Future research could investigate the effects of using different types of items, including constructed response items and technology enhanced items on linking results for the augmented tests.

In terms of criterion linking relationships, using different criterion linking relationships would cause different results for some conditions. For example, it was hard to determine whether the CE was more sensitive to the group ability

difference or to the differential effect size for the pseudo-data analyses, while it was obvious that the CE was more sensitive to the differential effect size than the group ability difference for the simulated data analyses. Note that the criterion linking relationships were different for the pseudo-data and simulated data analyses –single group linking relationship was used as the criterion for the pseudo-data analyses, while a simple structure multidimensional IRT observed-score linking relationship was used as the criterion for the simulated data analyses. Furthermore, in the pseudo-data analyses, the criterion linking relationship differed for each equating method. Therefore, generalization of the conclusions about performances of different equating methods should be exercised with caution.

Further, it would be interesting to see how the use of a different criterion may alter the results of the simulation study. An additional analysis was conducted where the unidimensional IRT linking with no group difference was used as a criterion for the simulated data analyses. The results of this additional analysis led to the same conclusions. Future research could be conducted to examine how different criterion linking relationships would affect linking results and how different criterion linking relationships would be influenced by different levels of group ability difference, differential effect size, and latent-trait correlation between the common assessment and state-specific item sets.

This study investigated various factors affecting linked scores of augmented tests that include both the common assessment and the state-specific item set. The guidelines in this study can provide ideas about what factors need to be considered when states, assessment consortia, and testing companies use equivalent scores on augmented tests over time, across grades, and states within a consortium.

6 References

Achieve. (2010, August). *Common Core State Standards & Accountability.* (Downloaded on January, 2011 from http://www.achieve.org/).

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes.* [Computer program]. Iowa City: University of Iowa.

CEP. (2012). *Year Two of Implementing the Common Core State Standards: States' Progress and Challenges.* (Available online: http://www.cep-dc.org/displayDocumentID=391).

Cizek, G. J. (2010). *Translating Standards into Assessments: The Opportunities and Challenges of a Common Core.* (Available online: http://www.brookings.edu/media/Files/events/2010).

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.),*Technical issues related to the introduction of the new SAT and*

PSAT/NMSQT (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*, 81–97.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items.* Unpublished research note.

Hanson, B. A., Zeng, L., & Cui, Z. (2004). *PIE* [Computer Software]. Iowa City: University of Iowa.

Kim, J. (2013). *Factors affecting accuracy of comparable scores for augmented tests under Common Core State Standards.* Unpublished doctoral dissertation. University of Iowa.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common item equating with nonrandom groups. *Journal of Educational Measurement, 22*, 197–206.

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

Lazer, S., Mazzeo, J., Way, W. D., Twing, J. S., Camara, W., & Sweeney, K. (2010). *Thoughts on linking and comparing assessments of Common Core Standards.* (Available online: http://www.ets.org/Media/Home).

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed format tests using a simple structure multidimensional IRT framework. In Kolen, M. J. & Lee, W. (Ed.) *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (*CASMA Monograph Number 2.2*, pp. 118-129).

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement, 8*, 453-461.

Luecht, R. M., & Camara, W. J. (2011). Evidence and Design Implications Required to Support Comparability Claims. Washington: PARCC.

Oregon Department of Education. (2011). *Development of the Common Assessment for 2014-2015.* (Available online: http://www.ode.state.or.us).

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

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39-49.

U.S. Department of Education. (2012) *Race to the Top Assessment*. (Available online: http://www2.ed.gov/programs/parcc-year-1.pdf).

Zimowski, M., Muraki, E., Mislevy, R., & Bock D. (2003). *Bilog-MG* [Computer software]. Mooresville, IN: Scientific Software International.