Practical Assessment, Research, and Evaluation

Volume 20 *Volume 20, 2015* Article 19

2015

Evaluating Common Item Block Options when Faced with Practical Constraints

Amanda Wolkowitz

Susan Davis-Becker

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 19, September 2015 ISSN 1531-7714

Evaluating Common Item Block Options when Faced with Practical Constraints

Amanda Wolkowitz & Susan Davis-Becker, *Alpine Testing Solutions, Inc.*

This study evaluates the impact of common item characteristics on the outcome of equating in credentialing examinations when traditionally recommended representation is not possible. This research used real data sets from several credentialing exams to test the impact of content representation, item statistics, and number of common items on equating results. The results of this research suggests that it may not be necessary to have a common item block that is strictly proportional in content or difficulty to the entire exam if the exam is unidimensional. The results also suggest that it may be beneficial to use all common items between two forms for equating instead of focusing on a smaller anchor block.

The purpose of equating scores from one exam form to another is to ensure that the final score on one form has the same meaning and interpretation as the comparable score on the other form of the same exam. The purpose of this research is to examine the extent to which a set of common items can deviate from being proportional to the content specifications of the exam and still be used as the basis for an accurate equating. Over the years, equating has been defined in many ways. Angoff (1971) defined equating as a conversion, much like one would convert yards to meters or pounds to kilograms. He stated, "to equate the forms [means] to convert the system of units of one form to the system of units of the other – so that scores derived from the two forms after conversion will be directly equivalent" (p. 85). Angoff (1971) qualified his equating definition with two restrictions: 1) the two forms of an exam must measure the same characteristics; and 2) the conversion equation between the two forms must be unique (up to the extent possible with random error). The first restriction implies that two forms of an exam that are to be equated must be built to the same content specifications and be parallel in construction. The second restriction implies that two forms of an exam must have a unique conversion. Specifically, the conversion equation should be independent of the abilities of examinees completing the forms at the time the conversion is developed.

Lord (1977) provided another definition. He stated, "Transformed scores y* and raw scores x can be called 'equated' if and only if it is a matter of indifference to each examinee whether he is to take test X or test Y" (p. 128). This definition is quite similar to Angoff's definition; it also implies that equating must occur between two forms measuring the same construct and that the conversion must hold regardless of the examinee completing the form.

Kolen and Brennan's (2014) definition is also comparable. They defined equating as "a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably" (p. 2). When defining equating it is important to note that equating is related to, but different than both linking and scaling. Specifically, equating is the strongest type of linking in that it links scores from one form of an exam to another form of the same exam. The equated scores may ultimately be placed (i.e.

"scaled") onto an established score scale for ease of interpretation (Kolen & Brennan, 2014). Weaker forms of score linkage include linking scores from one assessment to scores on a different assessment or linking scores from two different populations (Mislevy, 1992).

Just as Angoff (1971) and Lord (1977) defined equating, Kolen and Brennan also stated that two equated forms should be built with similar difficulty and content. Regardless of the exact definition adopted, equating means that the equated scores achieved by two examinees that complete different exam forms built to the same content specifications should have the same meaning.

One common theme exists among all of these definitions: the two forms being equated must be similar in difficulty and content. These definitions do not mention how equating is impacted by dimensionality. This leaves open the possibility that different forms of an exam built to multidimensional content specifications (e.g., a mathematics exam covering geometry and algebra ability) may still be equated to other forms of the same exam as long as the other forms are built with the same test characteristics as the base form to which it is equated. Although multidimensional equating has been challenged (Lumsden, 1961; 1976), there is a fine line between the definitions of unidimensionality and multidimensionality.

Test dimensionality refers to the number of abilities measured on an exam. An exam that is unidimensional measures only one latent ability (AERA, APA, & NCME, 2014). However, defining the latent ability is not so clear. For example, Linacre (1998) stated that "A data-set manifests one dimension so long as it is productive to think of it that way" (p. 268). Hambleton, Swaminathan, and Rogers (1991) defined a dataset as unidimensional if one dominant ability influences exam performance. If more than one dominant ability influences an examinee's performance on an exam, then the exam is multidimensional. Applying both of these definitions, if a high school mathematics exam measuring geometry and algebra ability has the purpose of measuring mathematics ability, it could be argued that the exam is unidimensional because it measures one latest ability of mathematical ability. On the other hand, it could be argued that the exam is multidimensional because different abilities are assessed in the two different content areas.

There is not one hard and fast rule for deciding whether a set of data is unidimensional or multidimensional. Several different tests may be applied to determine the dimensionality of an assessment. McDonald (1985) suggested examining the covariance matrix of the item residuals to detect any items that might be conditionally correlated. If such items exist, then it indicates additional levels of dimensionality in the data. Wright (1996) and Smith (1996) suggested using factor analysis to identify the dimension of a data set. Linacre (1998) recommended running a principal components factor analysis and then graphing the Rasch item measures against the standardized residual loadings. If multiple groups of items are shown on the graph, then the data may be multidimensional. Before declaring any dataset multidimensional, the items found to create the multiple dimensions should be analyzed to learn if it makes practical sense to group the items into the separate dimensions. If no logical explanation is found based on the content of the items, then other explanations might be found (e.g., compromised items or items with multiple keys). If this latter situation occurs, the data may still be considered unidimensional.

Test unidimensionality is an assumption of item response theory (IRT) equating, but not of classical test theory (CTT) equating. In credentialing exams, most organizations seek to develop unidimensional exams regardless of how the content is organized within the exam specifications. For example, an exam to become a licensed dentist may cover such topics as oral surgery, orthodontics, periodontics, endodontics, and radiology. The latent ability being measured by the exam is dentistry, but there are several "sub-abilities" of dentistry. To think of each sub-ability as a separate dimension would be impractical and likely too strict of an interpretation of the term "unidimensionality". If such an exam is considered unidimensional, then what role does the content specification play in equating?

The purpose of content specifications is to "delineate the aspects (e.g., content, skills, processes, and diagnostic features) of the construct or domain to be measured" (AERA, APA, & NCME, 2014, p. 76). In licensure and certification exams, the domain/construct/ability being measured is usually defined as unidimensional. Therefore, the purpose of the content specifications is to break down the one dimension into useful, organized components. One job

analysis committee may break down the measured ability into four major areas; whereas, another committee may organize it into six. Regardless of how the measured ability is organized into a content specification document, the exam still measures one latent ability.

If an exam is unidimensional, then the role that content representation plays in equating may be less critical than exams which are multidimensional. In CTT, multiple studies suggest that exams that are equated through nonequivalent groups with an anchor test (NEAT) design must be a mini-version of the full exam in terms of both content and difficulty (Angoff, 1968; Deng, Sukin, & Hambleton, 2009). Other researchers have suggested that that the common items should be proportionally equivalent to the exam itself in terms of content, but not necessarily difficulty (Keller & Keller, 2003). Although these findings seem necessary if an exam is truly multidimensional, it is arguable that proportional content representation in the anchor block may not be necessary for a unidimensional exam because all items represent the same latent ability.

The same question arises in IRT equating. In IRT equating, it is an assumption of the model that the data are unidimensional. If this assumption is met, then it again can be argued that content representation of a common item block may not be necessary because all items measure the same latent ability. Most researchers assume that the common items used for equating an exam need to be proportional in terms of content representation to the entire test (Keller & Keller, 2003); however, the robustness of violations of this assumption is an area still in need of research (Deng, Sukin, & Hambleton, 2009; Yang, 2000).

The purpose of this research is to examine the extent to which proportional content representation can be violated in a set of common items and still achieve accurate equating. This investigation has three practical implications. First, it is not always operationally possible for programs to create an anchor block that is proportional to the content specifications. This may occur when there is limited breadth within the item bank and one or more forms of the exam are not exactly proportional to the specifications. In this situation, the anchor block may be proportionally close to the content specifications, but not exactly representative. Second, due to limitations of depth within an item bank, programs may have to construct operational forms that have more common items than for which the exam design calls. In such cases, the question arises as to whether the equating should be based on the anchor block (proportionally representative in terms of content, but fewer items) or all the common items (not proportionally representative in terms of content, but more items). Third, some programs may slightly edit their content specifications over time to adjust the weight or emphasis of certain content areas. If strict content proportionality is used to create anchor blocks, then adjustments to any content area requires establishing a new base scale for equating. However, if equating efficiently works with using common items, then the same base form may potentially be used when a content area has been added or removed from the content specifications.

To investigate the extent to which content proportionality of a common item block could be violated and still result in accurate equating, this research equates two forms of an exam in which up to half of the items are used as the common item block and content proportionally is not of primary concern. This research does not seek to compare equating methods but rather serves to start an investigation into the necessity of proportional common item blocks from a new angle.

Method

Exams

Data from four credentialing exams were used in this study to investigate the need for the set of common items to have the same content representation as the total test. The exams varied in length from 35 to 100 items, the number of content domains listed in the blueprint varied from 4 to 11, and the content of the exams varied. Table 1 provides some basic information about each exam.

Population

One challenge to conducting this type of research with real data (i.e., non-simulated data) is differences in the population from one testing administration to the next. To avoid the error of having such differences influence the results of this research, data from one form and one administration for each exam was included in this analysis. Specifically, the sample of candidates who completed the same form of the exam was randomly divided into two groups. One group was

Table 1. Number of Candidates and Scored Items on the Four Exams Analyzed

Exam	Type of Exam	N Candidates	N Content Domains	N Scored Items	Internal Consistency Reliability
A	Licensure or Certification, depends on state	1649	6	100	0.91
B	Certification	824	5	35	0.68
C	Certification	1614	11	100	0.92
D	Licensure (one of several exams)	1140	4	80	0.77

randomly selected as the base group, and the other group was the new group (see Figure 1). Both groups completed the same form of the exam during the same administration period; therefore, the populations were assumed to be randomly equivalent. In addition, this design does not require equating between groups (i.e., all scores on the new form are equated to the same scores on the base form [identity equating]). Thus, the extent to which the proportionality of the content representation of the common items affects equating could be assessed by calculating the bias in the equating results.

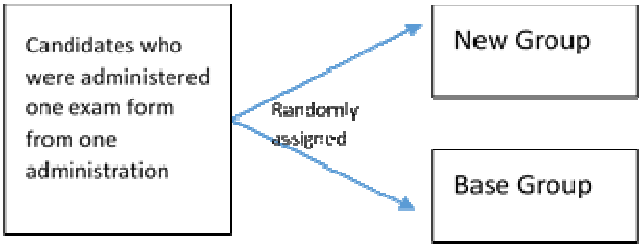

Figure 1. Diagram of the equating design.

Common item blocks

The content representation of the total test for each program was noted. Then, a two-dimensional, internal, common item block design was implemented as shown in Figure 2. The first dimension of this design was three different common item blocks based on how well the common items represented the exam in terms of content. The first block contained equating items proportional to the content representation of the entire test. In some cases, exact proportionality could not be achieved. However, each domain in the common item block was within 13% of the represented domain of the entire test (see Table 2). The second common item block contained equating items that were close to—but not proportional to—the content representation of the entire test. "Close to" was defined as a difference not exceeding 20% between the percent of items on the entire test representing one content domain and the percent of items in the common item block representing the same content domain. The small number of items in the common item block may make this difference seem large; however, it represents a small deviation in terms of the number of items in the equating block. For example, the largest difference of 20% was found in Exam B for the common item block that contained 25% of the total items. This difference was due to one domain having one item represented in the nine-item equating block instead of the 2.8 items needed to be representative of the entire test. In Exam D, two domains had 5 out of 20 items in an equating block instead of nine. These differences reflect the practical issues of equating that testing programs face. The third common item block contained items from only one domain of the exam. (Two domains had to be used for Exam C due to a lack of items from one domain.) This latter block was referred to as the "Not proportional" common item block.

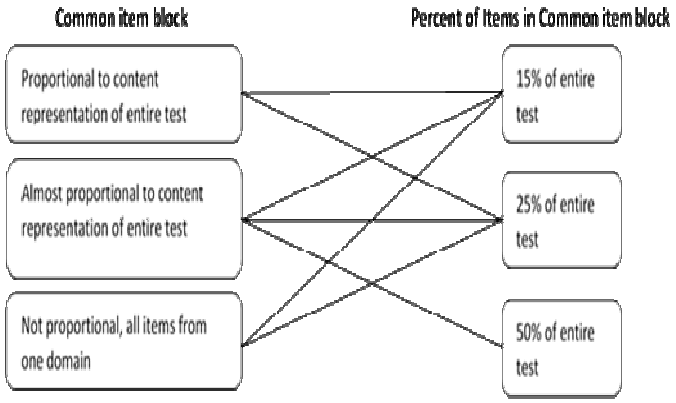

Figure 2. Diagram of the seven different common item blocks.

The second dimension of the design concerned the number of equating items. For each of the three common item blocks, 15% and 25% of the total items were selected to be part of the common item block. To investigate if using all common items as the common item block is more effective than a content-proportional anchor block with fewer items, one additional equating was performed. This equating

included a common item block containing 50% of the items on the total test; these items were close to proportional to the content representation of the entire test.

The equating items were selected at random from the entire test (without concern for item difficulty). Overall, this equating design led to seven different common item blocks for each exam.

Table 2 displays the maximum difference found between the content representation of the entire test and the common item block across all domains of an exam. For example, the blueprint for Exam A assigns 23% of the exam to domain 3. The common item block that is "close to proportional" and contains 25% of the total items on the exam had 40% of the items assigned to domain 3. The difference between these two values, i.e. 17%, was the greatest difference found amongst all domains of Exam A. This value of 17% is listed in Table 2 for Exam A under the heading: Close to Proportional – 25%.

As defined above, "proportional" common item blocks were within 13% of the content representation of the total test. In addition, all "close to proportional" common item blocks were within 20% of the content representation of the total test. The "not proportional" common item blocks were those blocks containing items from only one domain (with the exception of program C); thus, they deviated by 100% from the content representation of the total test in all but one content domain.

Table 2. Maximum Difference Between the Percent of Items in Each Content Area of the Total Test and the Percent of Items in Each Content Area of the Common Item Block

Exam	Proportional		Close to Proportional			Not Proportional	
	15%[1]	25%[2]	15%	25%	50%[3]	15%	25%
A	4%	2%	17%	17%	10%	100%	100%
B	12%	7%	14%	20%	12%	100%	100%
C	2%	4%	9%	10%	9%	100%	72%
D	2%	1%	11%	19%	7%	100%	100%

[1]15% of the exam equals 15 items for Exams A and C, 6 items for Exam B, and 12 items for Exam D.

[2]25% of the exam equals 25 items for Exams A and C, 9 items for Exam B, and 20 items for Exam D.

[3]50% of the exam equals 50 items for Exam A and C, 18 items for Exam B, and 40 items for Exam D.

Dimensionality

The dimensionality of each exam was assessed using the Rasch Principle Components Analysis (PCA) of Residuals. This analysis was performed on the full dataset (i.e., not half of the dataset representing the base or new group) using Winsteps® (Linacre, 2014). The amount of variance explained by the data and the greatest secondary dimension (i.e., unexplained variance in the first contrast) were initially reviewed. Then, the item measures were plotted against the standardized residual loading from the first contrast of a Rasch PCA of Residuals. The plot was examined for dimensionality by: 1) locating the items for each domain on the plot and determining if they grouped together, and 2) determining if a horizontal line could clearly separate a group of items (regardless of domain) from the other items (Linacre, 1998).

Equating Methods

Tucker linear (TLIN) and Rasch equating were the CTT and IRT methods performed to equate the new form of each exam to the base form using each of the seven common item blocks. Thus, a total of 14 equating procedures were performed for each exam. TLIN was selected as the CTT method because the method uses a common item block, has shown to be an effective method for equating when the two groups being equated are similar in ability (Puhan, 2012), and the assumptions of the model were met by the datasets included in this study. The primary assumptions of TLIN equating are that the two forms to be equated have equal reliability and do not differ greatly in terms of difficulty. Based on the design of this study (i.e., one dataset randomly split into two groups), both of these requirements were met. The TLIN equating was performed using the Common Item Program for Equating (CIPE) program (Kolen, 2004).

Rasch true score equating was selected as the IRT equating method due to its simplicity and the fact that most of the programs involved in this study use the Rasch model for equating their test forms. This form of equating assumes that the items assessed on an exam are unidimensional and locally independent. The dimensionality of the full dataset was assessed as previously described. The local independence assumption was assessed by analyzing the off-diagonal values of the correlation matrices at different intervals along the ability continuum. If the local independence

assumption is met, these values should be close to zero (Hambleton et al., 1991).

Evaluation Criteria

There are several ways to evaluate the accuracy of an equating procedure. These include using the root mean squared deviation, the two-tailed student's t-test, and examining the correlation between the subtotal scores on the common items and the total scores on the exam (i.e., index of equating efficiency, Budescu, 1985; Klein & Jarjoura, 1985; Kolen & Harris, 1990; Livingston, Dorans, & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990). This study evaluates the equating procedures by examining the bias and standard error of equating at the established cut score for the exam as well as across the ability spectrum. The study evaluates the magnitude of the correlation between the scores on the common item block and the total score (i.e., the index of equating efficiency, Budescu, 1985).

Results

The examinees from one exam administration were randomly divided into two groups: base group and new group. The number of examinees in each group and the mean score on the exam is provided in Table 3. The mean total scores for the two groups were within 0.75 points of each other. The alpha reliability of the base form and new form of the exams were also similar to each other and within 0.03 of each other (see Table 4).

Table 5 compares the average difficulty of the entire exam to the average difficulty of the seven different common item blocks for both groups. The average difficulty of the items in the common item blocks was approximately the same for the two groups (maximum difference was 0.03). The difficulty of the common item blocks was also within 0.10 score points of the difficulty of the entire test for most common item blocks. The average item difficulties for the common item blocks with nine proportional items (25%) for Exam B were more difficult than the overall test. This deviation in difficulty is likely due to the small number of total test items (N = 35) from which the common items were selected. Similarly, the common item blocks with 12 equating items (15%) selected to be "close to proportional" for Exam D were more difficult than the average difficulty of the exam.

The average item-score correlation for the anchor items ranged from 0.18 (base form of Exam D with 15% of the anchor items proportional to content) to 0.44 (base form of Exam A with 15% of the anchor items proportional to content) (see Table 6). Overall, the anchor items for Exam A and Exam B tended to have better item-score correlations than the average item on the entire respective exam, whereas the average item-score correlations for the anchor items on Exam C and Exam D were often less than the average item-score correlation for the items on the entire exam.

Table 3. Mean Total Score on Each Exam

Exam	Cut Score	Base Group			New Group		
		N	Mean	SD	N	Mean	SD
A	65 out of 100	825	66.88	13.69	824	66.63	13.98
B	23 out of 35	412	24.73	3.49	412	24.81	3.61
C	70 out of 100	807	75.50	13.63	807	75.33	13.59
D	55 out of 80	570	58.40	7.55	570	58.55	8.00

Table 4. Reliability of the New and Base Forms

Exam	Base	New
A	0.91	0.91
B	0.68	0.70
C	0.92	0.92
D	0.77	0.80

Table 5. Difficulty of the Common Item Blocks (values listed are the average p-values of the common items)

Exam	All Items	Proportional to Content				Close to Proportional to Content						Not Proportional to Content			
		15%		25%		15%		25%		50%		15%		25%	
		Base	New	Base	New	Base	New	Base	New	Base	New	Base	New	Base	New
A	0.67	0.68	0.67	0.71	0.71	0.68	0.68	0.70	0.70	0.66	0.66	0.71	0.70	0.68	0.67
B	0.71	0.65	0.64	0.55	0.54	0.64	0.64	0.67	0.67	0.75	0.75	0.79	0.79	0.71	0.72
C	0.75	0.83	0.83	0.76	0.76	0.75	0.74	0.77	0.77	0.77	0.76	0.79	0.79	0.76	0.76
D	0.73	0.76	0.76	0.75	0.76	0.62	0.63	0.72	0.73	0.73	0.73	0.65	0.65	0.64	0.64

Table 6. Item-Score Correlations of the Common Item Blocks (values listed are the average item-score correlations of the common items)

Exam	All Items	Proportional to Content				Close to Proportional to Content						Not Proportional to Content			
		15%		25%		15%		25%		50%		15%		25%	
		Base	New	Base	New	Base	New	Base	New	Base	New	Base	New	Base	New
A	**0.31**	0.44	0.43	0.34	0.34	0.40	0.39	0.34	0.34	0.33	0.34	0.39	0.39	0.38	0.37
B	**0.33**	0.39	0.41	0.34	0.35	0.29	0.28	0.34	0.35	0.35	0.38	0.31	0.30	0.38	0.35
C	**0.36**	0.40	0.40	0.34	0.35	0.35	0.35	0.35	0.36	0.37	0.37	0.35	0.37	0.35	0.35
D	**0.24**	0.18	0.21	0.23	0.24	0.24	0.25	0.23	0.24	0.22	0.25	0.20	0.21	0.21	0.22

Dimensionality

The Rasch PCA of Residuals resulted in the Rasch dimension explaining 21.5%, 49.3%, 22.4%, and 18.0% of the variance in Exams A, B, C, and D. Exams A, C, and D had eigenvalues of 3.1 or less, whereas Exam B had an eigenvalue of 4.6. The unexplained variance in the first contrast did not exceed the variance explained by either the person or item measures in any of the four exams. The scatterplots for exams A, C, and D did not show evidence of items from the same domain grouping together (see Figure 3). The grouping of data points observed in the scatterplot for Exam B did show signs of multidimensionality; however, further investigation into the items did not reveal that any one content domain was causing the grouping. Other

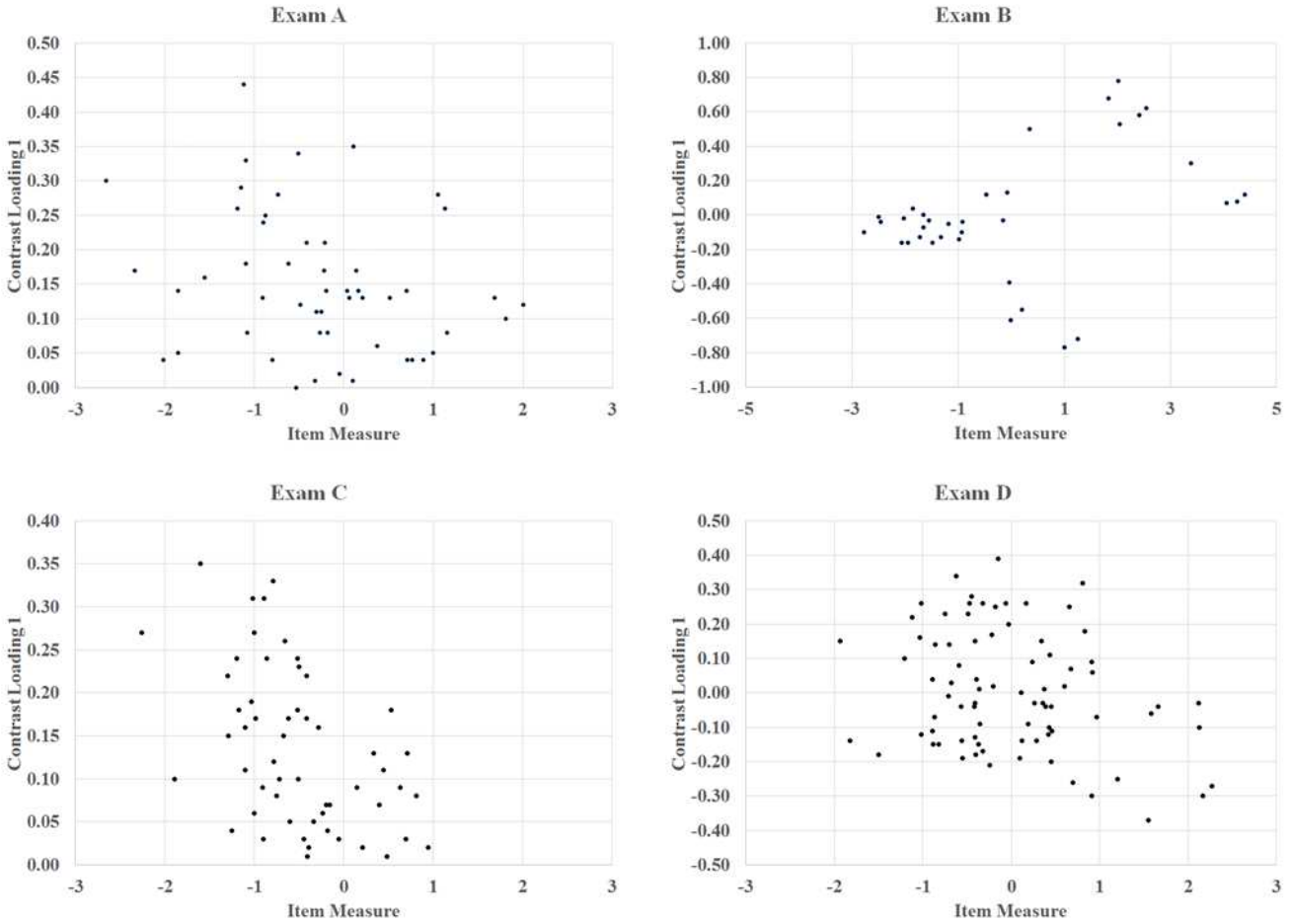

Figure 3. Scatterplots of the item measure versus 1st contrast loading for each exam.

possible reasons for the multidimensional look of Exam B could be the short length of the exam (N = 35) or item compromise. Overall, the Rasch PCA of Residuals analysis suggests that all the datasets can be thought of as unidimensional even though multiple content domains are defined as components of the latent ability.

CTT Equating Result

The accuracy of the equating using each of the seven common item blocks was measured by bias and standard error. The two groups of examinees completed each exam during the same exam administration and completed the same form of the exam; therefore, the most accurate equating would be identity equating (or not equating). This would have a bias of 0. For each of the four exams, TLIN equating was performed and bias was calculated at the cut score of the exam. The differences between the equated scores and the "true" scores (i.e., scores resulting from identity equating [original scores]) were examined.

Table 7 displays the results of the equated cut scores. Among the four exams, there was no clear trend as to which common item block performed the best. At the cut score, the absolute bias for all common item blocks was less than 0.50. This suggests that all blocks did a respectable job of equating the two forms.

Table 7. Equated Cut Score Using TLIN Equating

Exam	Cut Score	Common Items Proportional to Content		Common Items Close to Proportional to Content			Common Items Not Proportional to Content	
		15% Block	25% Block	15% block	25% block	50% block	15% block	25% block
A	65	64.89	64.85	65.00	64.79	64.93	65.06	64.75
B	23	23.14	23.08	22.97	23.01	22.93	23.00	22.88
C	70	70.17	70.27	70.28	70.07	70.29	69.68	70.23
D	55	54.94	54.55	54.77	54.75	55.10	54.99	55.00

Beyond the equating at the cut score, the bias was estimated for each equating procedure across the entire score scale. The absolute bias in the equating results was estimated for each exam and each common item block (see Table 8). Although no equating block outperformed another block when equating at the established cut score for the exams, the equating block containing 50% of the items that were close to proportional to the content of the exam had the least average absolute bias among the seven equating blocks across all but one of the exams (also displayed graphically in Figure 4).

Table 8. Average Absolute Bias Across All Equated Scores Using TLIN Equating

Exam	Common Items Proportional to Content		Common Items Close to Proportional to Content			Common Items Not Proportional to Content	
	15% block	25% block	15% block	25% block	50% block	15% block	25% block
A	1.23	0.76	0.76	0.24	0.13	0.32	0.80
B	0.31	0.39	0.56	0.37	0.15	0.34	0.69
C	0.36	1.06	0.76	0.42	0.40	0.22	0.48
D	0.73	1.19	1.15	1.18	0.20	1.01	0.67

Further evidence supporting this finding is seen in the analysis of the standard errors of equating (Figure 5). For all four exams, the least amount of standard error of equating across the ability continuum occurred when 50% of the items were used in the equating block. The remaining equating blocks had similar standard errors to each other and the error increased at the extremes.

IRT Results

Rasch true score equating assumes that the datasets are unidimensional and the items are locally independent. The datasets were treated as unidimensional as previously discussed. Local independence was checked by observing the off diagonal values of correlation matrices for examinees scoring in the bottom 25% of the total group and in the top 25% of the total group. The assumption was checked using all examinees (i.e., not the split group). Across all four exams, less than 20% of the inter-item correlations had values greater than 0.20[1]. These values were deemed acceptable to meet the requirements to perform IRT equating.

The accuracy of Rasch equating at the cut score was evaluated for each of the seven common item blocks. Bias was measured as the difference between the ability levels of the base and new group at the unequated cut score. For example, the cut score for the base group completing Exam A was 65 or an ability

[2] The inter-item correlations should be close to zero to indicate local independence. The assumption is somewhat robust, so a value of 0.20 was selected as the criteria upon which to evaluate whether the assumption was met or violated.

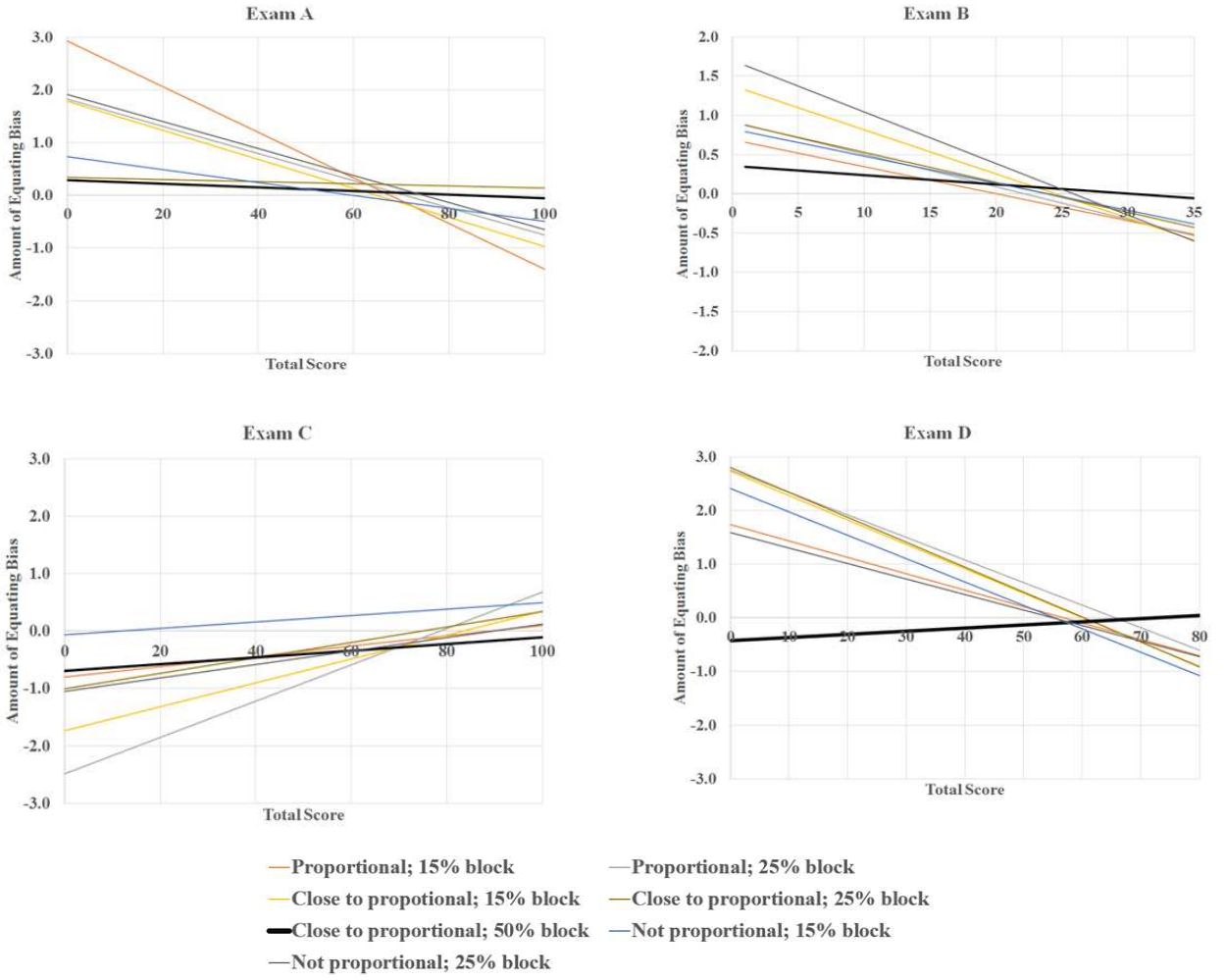

Figure 4. Absolute bias for all scores using TLIN equating.

level of 0.76. After performing Rasch equating, a cut score of 65 corresponded to an ability level of 0.75 for one of the equating blocks in the new group. Thus, the bias for this equating block was 0.01 on the Rasch scale. The most accurate equating would yield a bias of 0.00.

Table 9 shows the ability level of the new group at the unequated cut score for each exam and common item block. In general, equating with a common item block consisting of 15% of the total items did not perform as well as the other common item blocks. However, there was no clear "best" common item block. As with the CTT equating, all methods reproduced the equating accurately at the cut score. In addition, all estimated ability levels at the equated cut score were within 0.10 of the "true" ability level regardless of the common item block used.

Table 9. Ability Level of New Group at Raw Cut Score

Exam	Theta Cut Score	Common Items Proportional to Content		Common Items Close to Proportional to Content			Common Items Not Proportional to Content	
		15% block	25% block	15% block	25% block	50% block	15% block	25% block
A	0.76	0.74	0.75	0.75	0.75	0.76	0.75	0.76
B	0.92	0.83	0.88	0.93	0.92	0.96	0.89	0.92
C	0.98	0.92	0.95	0.96	0.96	0.96	0.99	0.96
D	0.90	0.94	0.94	0.94	0.91	0.89	0.89	0.90

The finding that no equating block outperformed another was also seen when examining the exam characteristic curves (TCC) and standard errors across the ability continuum for all seven common item blocks and four exams. Plots of the TCCs and standard

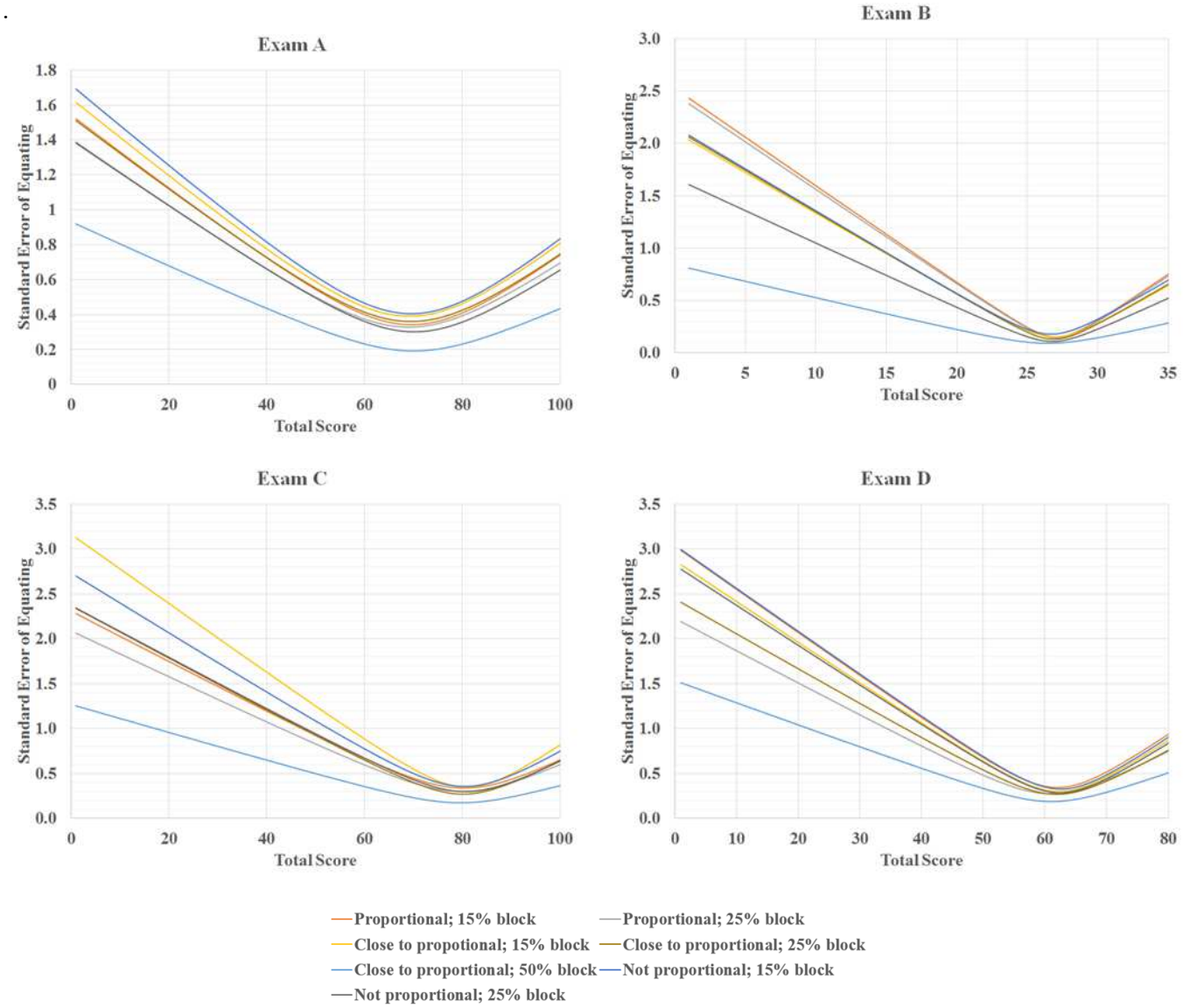

Figure 5. Standard errors of equating for the TLIN method

errors revealed coincidental graphs indicating equivalent scores and standard errors across the ability continuum for all common item blocks. These results suggest that if the exams to be equated are unidimensional and the ability of the two groups involved in the equating are approximately equal, then proportional content representation in the common item block may not be necessary.

Discussion

The main purpose of this research was to investigate the extent to which content proportionality of a common item block could be violated and still result in accurate equating. In this study, the examinees from one administration of four different exams were randomly split into two groups. As a result of this design, the two groups were of very similar ability, and the selected common item blocks had similar difficulty between the two groups.

Although it is not realistic to equate two forms of an exam administered during the same period with the same items, this study completed the equating in this way to determine how well different common item blocks would reproduce the scores on the base form of the exam. Although it is good practice to have the common item blocks be a mini-version of the entire test (Dorans, Moses, & Eignor, 2011; Kolen & Brennan, 2014), it may not be necessary when the

dataset is unidimensional. The primary purpose of having the common item block proportional to the content representation of the entire test is to capture any group differences and differential performance on items in different content areas (Dorans et al., 2011; Kolen & Brennan, 2014). However, if an exam is sufficiently unidimensional, then it can be argued that all items on the exam measure the same dominant construct or latent ability; therefore, including more items in the common item block better reflects any group differences than a smaller number of items.

Klein and Jarjoura (1985) completed a study with a similar goal. In their study, the exam consisted of six "tightly defined" content areas and TLIN equating was used. They found that proportional content representation was more important for equating accuracy than having a larger number of items in the common item block that were not proportional. The issue of unidimensionality was not discussed in the study because TLIN equating was the equating method of choice and does not require such an assumption being met. However, in the present study of four credentialing exams, unidimensionality was assessed and all datasets were deemed to have one dimension. If datasets were multidimensional, then different results would likely have surfaced and perhaps confirmed Klein and Jarjoura's finding that content proportionality is needed for CTT equating.

The primary role of any common item block is to quantify the difference between the two groups of examinees. According to Holland and Dorans (2006), "The most important properties of the anchor test are its integrity and stability over time and its correlation with the scores on the two tests being equated" (p. 201). It goes without question that the items selected to be part of the common item set should be evaluated for drift to ensure that they have not changed their statistical properties over time. In this study, this was not a concern because there was no difference in time between the two groups studied.

In terms of the correlation, a strong relationship between the common item scores and the total test scores is essential for accurate equating. In addition, longer common item sets are better than shorter ones because they have higher reliability (Holland & Dorans, 2006). Table 10 shows the correlations between the common items and the total scores on the base and new forms. The highest correlations were with the common item blocks made up of 50% of the total test followed by 25% of the total test. However, the correlations did not differ significantly based upon whether the common items were proportional, close to proportional, or not proportional to the content representation of the entire test. As such, this study did not find that having a common item block as a mini-test increased the correlation between the scores on the common item block and the total test.

Table 10. Correlations Between the Scores on the Common Item Block and Total Score for the Base and New Forms

Exam	Form	Proportional		Close to Proportional			Not Proportional	
		15%	25%	15%	25%	50%	15%	25%
A	Base	0.86	0.87	0.81	0.85	0.96	0.80	0.89
B	Base	0.63	0.67	0.67	0.77	0.91	0.62	0.86
C	Base	0.85	0.89	0.83	0.90	0.96	0.81	0.88
D	Base	0.56	0.78	0.69	0.77	0.90	0.60	0.73
A	New	0.85	0.87	0.82	0.84	0.96	0.80	0.89
B	New	0.63	0.65	0.66	0.76	0.93	0.60	0.84
C	New	0.85	0.89	0.82	0.91	0.96	0.84	0.88
D	New	0.62	0.79	0.68	0.78	0.92	0.62	0.72

This study did find that all four exams were unidimensional and yielded more accurate equating for the TLIN procedure when more items were used in the common item block. In particular, the equating procedures performed with 50% of the total items selected as common items performed better when looking at all possible exam scores and with less error than any other common item block.

In terms of the IRT equating, all blocks performed equally well. This is not surprising given the two groups were of very similar abilities and the difficulties of the content areas were approximately equal between the two groups. Thus, the base scale upon which the new form was placed was stable and similar for all equating blocks. In practice, it seems that the greater number of stable items that can be anchored to a base scale for IRT equating, then the more accurate the equating would be. When all stable common items are not anchored to the base scale, error is introduced into the equating; the common items on the new form that are not anchored are being estimated to new values that differ from the base form calibration. Although this latter argument could not be directly assessed in this research, it is an area for further study.

Conclusion

In this research, we examined the extent to which a set of common equating items can stray from being proportional to the content of the entire test if the exam measures a truly unidimensional trait. Four different credentialing exams were used and seven different common item blocks were studied for each exam. The results indicated that the TLIN method of equating with a common item block of 50% of the total items on the exam that were close to—but not proportional to—the content representation of the total test, outperformed the six other equating blocks (including those with proportional content representation). These remaining common item blocks performed approximately the same. The results of the Rasch equating indicated that all common item sets performed the equating equally well.

There are three primary practical implications of these findings. The first is that it may not be necessary to have a common item block that is strictly proportional in content or difficulty to the entire exam if the exam is unidimensional. Second, it may be beneficial to use all common items between two forms for equating instead of a smaller anchor block. Third, these findings support the possibility of continuing to equate new forms of an exam that have only slightly changed from the blueprint (e.g., addition or removal of a content area) to the base form in order to maintain the calibrated item bank scale. These implications may be beneficial to programs that are challenged to meet the recommended requirements for common item blocks due to weaknesses in the item bank.

This research is not intended to provide a definitive guide for future equating. Instead, it reexamines the concept of using a mini-test as an anchor test from an operational perspective and opens the discussion as to how different the common item set can be from the total test to have effective equating when the dataset is unidimensional. As in any equating, it is highly recommended that the equating results stemming from a non-representative common item set be evaluated for reasonableness and accuracy. Moreover, since the research presented here was only conducted on four exams, more research in this area should be conducted before using non-representative common items as the primary method for establishing an equating block.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Angoff, W. H. (1968). How we calibrate College Board scores. *College Board Review*, 68, 11-14.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508-600). Washington, DC: American Council on Education.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement*, 22, 13-20.

Deng, N., Sukin, T., & Hambleton, R. (2009). Judging the content and statistical equivalence of MCAS operational and linking items. *Center for Educational Assessment MCAS Validity Report* No. 20. (CEA-709). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21-42). New York, NY: Springer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE Publications, Inc.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 187-220). Westport, CT: Praeger Publishers.

Keller, R. R., & Keller, L. A. (2003). The effect of anchor test composition on the accuracy of classifying examinees. Available at http://www.measuredprogress.org/scoring-and-reporting-equating.

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. (2004). Common Item Program for Equating (CIPE) [computer program]. Version 1100. Available from http://www.education.uiowa.edu/centers/casma

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking (3rd ed.). New York, NY: Springer.

Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, 27(1), 27-39.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, 2(3), 266-283.

Linacre, J. M. (2014). Winsteps® (Version 3.81.0) [Computer Software]. Beaverton, OR: Winsteps.com

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3, 73-95.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14(2), 117-138.

Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin*, 58(2), 122-131.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, 27, 251-280.

McDonald, R. P. (1985). Factor analysis and related methods. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: ETS Policy Information Center.

Puhan, G. (2012). Choosing among Tucker or chained linear equating in two testing situations: Rater comparability scoring and randomly equivalent groups with an anchor. *Journal of Educational Measurement*, 49(3), 312-329.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education*, 3, 53-71.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, 3(1), 25-40.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, 3(1), 3-24.

Yang, W. (2000). The effects of content homogeneity and equating method on the accuracy of common-item test equating. Paper presented at the annual meeting of the *American Educational Research Association*, New Orleans.

Citation:

Wolkowitz, Amanda & Davis-Becker, Susan. (2015). Evaluating Common Item Block Options when Faced with Practical Constraints. *Practical Assessment, Research & Evaluation*, 20(19). Available online: http://pareonline.net/getvn.asp?v=20&n=19

Corresponding Author

Amanda Wolkowitz
Psychometrician
Alpine Testing Solutions, Inc.
51 West Center Street, #514
Orem, UT 84057

email: Amanda.Wolkowitz [at] alpinetesting.com