Invariance of Score Linkings Across Gender Groups for Forms of a Testlet-Based College-Level Examination Program Examination

Wen-Ling Yang and Rui Gao, Educational Testing Service

This study investigates whether the functions linking number-correct scores to the College-Level Examination Program (CLEP) scaled scores remain invariant over gender groups, using test data on the 16 testlet-based forms of the CLEP College Algebra exam. To be consistent with the operational practice, linking of various test forms to a common reference form is based on the Rasch model. Equatability indices proposed by Dorans and Holland (2000) were used to evaluate linking invariance over gender subpopulations. Overall, linkings based on the gender groups are very similar to linkings for the total group. At all score levels, differences between subgroup and total group linkings are smaller than the difference that will affect the pass/fail decision for CLEP candidates. On only one form, linking based on the male group would pass slightly more candidates than linking for the total group at the recommended CLEP cut score of 50 due to rounding. *Index terms: test equating, linking, invariance of linking, equatability*

To the greatest extent possible, equating functions should not be strongly influenced by the population of candidates on which they are derived. If the equating functions used to link the scores of two tests are not invariant across different subpopulations of candidates, the two tests really cannot be considered to be equatable (Dorans & Holland, 2000). The testlet-based College-Level Examination Program (CLEP) exams have multiple test forms designed to be similar in content and statistical properties. The number-correct scores on the alternate test forms are linked to a common reference form based on the Rasch model (i.e., one-parameter item response theory [IRT] model). This study investigated whether the functions that link the number-correct scores on a new form to the scores on a reference form remained invariant over gender subgroups. Because operational linkings for CLEP are based on the Rasch model by design, the Rasch model was used in this study to yield linkings that were consistent with the operational practice. Three types of equatability measures were used to assess to what degree the linking functions were invariant over subpopulations.

Linking/Equating Design for CLEP Testlet-Based Exams

CLEP is a widely accepted credit-by-examination program. It gives students an opportunity to demonstrate college-level knowledge that they have gained through prior study, independent study, professional experience, and/or cultural pursuits. College students who pass a CLEP exam will receive course credit, course exemption, and/or advanced placement toward a degree. Scores

Applied Psychological Measurement, Vol. 32 No. 1, January 2008, 45–61
DOI: 10.1177/0146621607311577

Table 1
Component Testlets for the 16 College Algebra Exam Forms

Form	Testlets				
1	A1	B1	C1	D1	V1
2	A1	B1	C1	D2	V1
3	A1	B1	C2	D1	V1
4	A1	B1	C2	D2	V1
5	A1	B2	C1	D1	V1
6	A1	B2	C1	D2	V1
7	A1	B2	C2	D1	V1
8	A1	B2	C2	D2	V1
9	A2	B1	C1	D1	V1
10	A2	B1	C1	D2	V1
11	A2	B1	C2	D1	V1
12	A2	B1	C2	D2	V1
13	A2	B2	C1	D1	V1
14	A2	B2	C1	D2	V1
15	A2	B2	C2	D1	V1
16	A2	B2	C2	D2	V1

on CLEP subject exams are reported on a scale of 20 to 80. The recommended minimum credit-granting score is a CLEP score of 50,[1] which represents the average test score of students who earn a grade of C in the corresponding college course.

Each of the CLEP testlet-based exams has multiple forms, all with the same number of testlets. The number of testlets varies from exam to exam, however. For instance, the CLEP College Algebra exam has 16 test forms, each consisting of five testlets. Each testlet is a collection of questions from a coherent content domain, and testlets are the building blocks for the CLEP exams. For College Algebra, items from five different content domains are selected to form types of testlets (A, B, C, D, and V). Depending on the item pool size, multiple testlets of the same type (e.g., A1, A2, A3, etc.) may be available. By design, testlets of the same type are comparable in content and statistical properties, such that they can be used interchangeably in test assembly. A test form is essentially a combination of testlets of different types that together meet both the content and statistical specifications of the exam. For the CLEP College Algebra exam, since two alternate testlets are available for each type except V, 16 test forms can be assembled, as shown in Table 1.

The 16 test forms overlap with one another at the testlet level to varying degrees, and the testlet-based test assembly approach results in test forms that are comparable in content and statistical properties. The computerized delivery software assigns a test form at random to a test taker. Test scores on different forms are equated to the same reference form to adjust for inevitable differences in form difficulties that arise in test construction.

To derive comparable scores across test forms on the CLEP 20-80 scale, the PARSCALE program was used to calibrate all items in the 16 test forms with the Rasch model (Hambleton & Swaminathan, 1990):

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}},$$

where $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly, D is a scaling factor, and b_i is the item difficulty for item i. b_i represents the point on the ability scale at which a candidate has a 50% probability of answering item i correctly.

For each of the testlet-based forms of the CLEP College Algebra exam, a unique conversion was established to link the number-correct scores on the form to the 20-80 CLEP score scale. The following diagram depicts how the observed number-correct scores on a testlet-based new test form were linked to scores on the common reference form[2] and then placed onto the 20-80 CLEP score scale:

$$\begin{aligned} &\textit{Raw Number-Correct Score}_{new} \rightarrow \theta_{new} \rightarrow \theta_{reference} \rightarrow \textit{Raw Number-Correct Score}_{reference} \\ &\quad \rightarrow \textit{Raw Formula Score}_{reference} \rightarrow \textit{CLEP Scaled Score}. \end{aligned}$$

To be specific, for a particular CLEP testlet-based new form the observed number-correct scores on the form were treated as expected IRT true scores on the number-correct scale and were then converted to the ability scores (θ) corresponding to the expected IRT true scores via a test characteristic curve for that form. The Stocking and Lord (1983) transformation method was used to place all parameter estimates from separate calibrations on the same metric. Therefore, the ability scores on the testlet-based new form were on the same scale as the ability scores on the common reference form. Using the test characteristic curve for the common reference form, the ability scores (θ) on the reference form were converted to the expected IRT true scores, which were then treated as if they were reference-form raw number-correct scores. Assuming no omits or not-reached items, the reference-form raw number-correct scores were further transformed into the reference-form raw formula scores by using the following equation for formula scoring:

$$FS = R - \left(\frac{n-R}{k-1}\right),$$

where R is the number-correct score, n is the total number of items on the reference form, and k is the number of multiple-choice options.

Finally, using a linear conversion associated with the reference form, these raw formula scores on the reference-form scale were placed onto the CLEP 20-80 score scale.

Data and Study Design

To evaluate the IRT-based linking outcomes for the testlet-based CLEP exam, test data from the 16 forms of the CLEP College Algebra examination were used to study linking invariance across gender subpopulations. The CLEP College Algebra exam covers material usually taught in a one-semester college course in algebra. About half of the exam consists of routine problems requiring basic algebraic skills, and the remainder involves solving nonroutine problems that require candidates to demonstrate their understanding of concepts.

All the CLEP College Algebra items were 0/1-scored multiple-choice items. The exam contained about 60 items, including 50 or so operational items and 10 or so pretest items, administered in a 90-min testing session. Operational test data clearly showed that the CLEP College Algebra exam was not speeded. Across test forms, more than 98% of the test takers were able to complete the entire exam within the designated testing time. In addition, operational College Algebra items were screened for differential item functioning (DIF) in gender subgroups. The DIF analysis showed no substantial gender DIF issue on this exam.

Table 2 shows the sample sizes of the total group and the gender subgroups for each of the 16 forms of the CLEP College Algebra exam. There were about 1,000 candidates for each test form,

Table 2
Sample Sizes of Total and Gender Subgroups on the CLEP College Algebra Exam

Test Form		Total Group n	Male Group n_m	Male Group Proportion (n_m/n)	Female Group n_f	Female Group Proportion (n_f/n)
1	A1B1C1D1V1	995	415	.42	580	.58
2	A1B1C1D2V1	1,041	450	.43	591	.57
3	A1B1C2D1V1	1,035	456	.44	579	.56
4	A1B1C2D2V1	1,003	408	.41	595	.59
5	A1B2C1D1V1	1,079	439	.41	640	.59
6	A1B2C1D2V1	1,013	452	.45	561	.55
7	A1B2C2D1V1	957	415	.43	542	.57
8	A1B2C2D2V1	980	420	.43	560	.57
9	A2B1C1D1V1	1,045	441	.42	604	.58
10	A2B1C1D2V1	1,018	455	.45	563	.55
11	A2B1C2D1V1	1,017	428	.42	589	.58
12	A2B1C2D2V1	1,003	431	.43	572	.57
13	A2B2C1D1V1	980	424	.43	556	.57
14	A2B2C1D2V1	987	417	.42	570	.58
15	A2B2C2D1V1	1,009	423	.42	586	.58
16	A2B2C2D2V1	959	422	.44	537	.56
Overall		16,121	6,896	.43	9,225	.57

Note. CLEP = College-Level Examination Program.

with more females than males. Over the various test forms, the male group comprised 41% to 45% of the total group, and the female group comprised 55% to 59% of the total group.

The average number-correct scores and standard deviations for groups taking different forms of the College Algebra exam are summarized in Table 3. It shows that the male group had slightly higher mean scores than the female group on all but one form—Form 14. The female group scored higher than the male group by about half a raw score point on Form 14. The average raw scores across various test forms were similar to one another, both for the total group and for each of the gender subgroups. This provided evidence of random assignment of test forms to candidates (i.e., the groups taking different forms were fairly equivalent). Overall, Table 3 shows that the test forms were designed to be fairly similar to one another. Special attention was given to Form 14, where the group and/or the test form might have been somewhat different from the rest, when analyzing linking variances.

Using IRT-based equating and the reference-form raw score to scaled score transformation, raw scores on the CLEP testlet-based new forms were converted to scaled scores on the 20-80 CLEP scale. The average CLEP scaled scores and standard deviations for groups taking the various forms of the College Algebra exam are summarized in Table 4. As expected, the male group had higher mean CLEP scaled scores than the female group on all but Form 14.

For each of the 16 test forms, equatability measures were computed to assess the degree of linking invariance. In addition to assessing the invariance of score linking functions, which were in the metric of the CLEP scaled score, this study also examined whether linkings based on different subpopulations produced different pass/fail decision outcomes from linkings based on the total population.

Table 3
Average Raw Scores of Total and Gender Subgroups on the CLEP College Algebra Exam

			Total Group		Male Group		Female Group	
Test Form		*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	A1B1C1D1V1	995	27.21	10.58	27.77	10.79	26.81	10.41
2	A1B1C1D2V1	1,041	27.17	10.32	28.42	10.53	26.21	10.04
3	A1B1C2D1V1	1,035	27.60	9.98	28.50	10.05	26.89	9.87
4	A1B1C2D2V1	1,003	27.09	10.12	27.39	9.91	26.89	10.25
5	A1B2C1D1V1	1,079	26.93	10.47	28.30	10.55	26.00	10.32
6	A1B2C1D2V1	1,013	26.72	10.55	27.16	10.70	26.36	10.42
7	A1B2C2D1V1	957	26.53[a]	10.24	27.62	10.32	25.70[a]	10.10
8	A1B2C2D2V1	980	27.42	10.05	27.55	9.86	27.33[b]	10.18
9	A2B1C1D1V1	1,045	27.25	9.95	27.86	9.98	26.80	9.90
10	A2B1C1D2V1	1,018	26.96	10.23	28.25	10.42	25.92	9.95
11	A2B1C2D1V1	1,017	27.84	9.78	29.15	9.79	26.89	9.66
12	A2B1C2D2V1	1,003	26.81	9.72	27.42	10.05	26.35	9.44
13	A2B2C1D1V1	980	27.63	9.93	28.51	9.88	26.97	9.93
14	A2B2C1D2V1	987	27.03	10.08	26.73[a]	10.42	27.25	9.82
15	A2B2C2D1V1	1,009	27.91[b]	9.65	29.34[b]	9.71	26.88	9.48
16	A2B2C2D2V1	959	26.97	9.88	27.74	10.03	26.36	9.73

Note. CLEP = College-Level Examination Program.
a. The minimum of means across test forms.
b. The maximum of means.

Table 4
Average CLEP Scaled Scores of Total and Gender Subgroups on the College Algebra Exam

			Total Group		Male Group		Female Group	
Test Form		*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	A1B1C1D1V1	995	53.72	12.04	54.38	12.33	53.25	11.83
2	A1B1C1D2V1	1,041	54.04	11.84	55.57	12.12	52.87	11.51
3	A1B1C2D1V1	1,035	54.25	11.42	55.40	11.57	53.36	11.25
4	A1B1C2D2V1	1,003	54.05	11.66	54.57	11.48	53.69	11.80
5	A1B2C1D1V1	1,079	53.56	11.82	54.99	12.00	52.59	11.61
6	A1B2C1D2V1	1,013	53.68	12.01	54.13	12.22	53.32	11.85
7	A1B2C2D1V1	957	53.20[a]	11.60	54.41	11.79	52.28[a]	11.40
8	A1B2C2D2V1	980	54.58	11.49	54.76	11.33	54.45[b]	11.62
9	A2B1C1D1V1	1,045	53.73	11.34	54.43	11.47	53.21	11.24
10	A2B1C1D2V1	1,018	53.77	11.77	55.34	12.05	52.52	11.41
11	A2B1C2D1V1	1,017	54.50	11.20	56.12	11.32	53.34	11.01
12	A2B1C2D2V1	1,003	53.69	11.23	54.57	11.68	53.04	10.87
13	A2B2C1D1V1	980	54.32	11.22	55.18	11.24	53.67	11.18
14	A2B2C1D2V1	987	54.01	11.50	53.61[a]	11.95	54.31	11.19
15	A2B2C2D1V1	1,009	54.74[b]	10.96	56.34[b]	11.14	53.60	10.71
16	A2B2C2D2V1	959	54.04	11.32	54.96	11.57	53.32	11.10

Note. CLEP = College-Level Examination Program.
a. The minimum of means across test forms.
b. The maximum of means.

Method

The equatability indices that measure subpopulation invariance of linking functions, proposed by Dorans and Holland (2000) and described in von Davier, Holland, and Thayer (2004), were computed to assess the equatability of the forms of the CLEP College Algebra exam. The root mean square difference (RMSD) statistic describes the difference between the total and the subgroup linking functions across subgroups at each score level, and the root expected mean square difference (REMSD) is a measure of overall differences between the total and the subgroup linking functions across subgroups and across score levels. In addition to these two indices, the root expected square difference ($RESD_j$) statistic for individual groups/subpopulations used by Yang (2004) was computed to evaluate the linking difference between each subgroup and the total group across score levels.

Because CLEP scores are used for making pass/fail decisions for granting college-level credits, for each form of the College Algebra exam the recommended cut score for the exam (i.e., the CLEP credit-granting score) was also applied to investigate whether linkings based on subpopulations produced different pass/fail outcomes than those based on the total population. The pass/fail classification rates based on various linkings were compared, and the practical significance of the differences was evaluated. The classification outcomes are summarized in the Results section.

RMSD

This section discusses various equatability measures and explains the criterion used to evaluate the magnitude of equatability measures. The equatability measures were compared to the criterion to decide whether the linking differences were of practical significance.

Let P be the population of CLEP candidates with subpopulations P_j that partition P into a set of mutually exclusive and exhaustive subpopulations. In this study, the subpopulations are male and female groups, so there are $J = 2$ subpopulations. The formula for the RMSD statistic is defined as follows:

$$\mathrm{RMSD}(x) = \frac{\sqrt{\sum_{j=1}^{J} w_j \left[e_{P_j}(x) - e_P(x) \right]^2}}{\sigma_{Y_P}},$$

where x is a raw score level on the testlet-based CLEP exam, $e_P(x)$ denotes the function that places x on the CLEP scale for the total population P, $e_{P_j}(x)$ denotes the function that places x on the CLEP scale for the subpopulation P_j, w_j is the proportion of P_j in P, and $\Sigma w_j = 1$. The denominator, σ_{Y_P}, is the standard deviation of the CLEP scaled score in the total population P (Dorans & Holland, 2000).

RESD

The $RESD_j$ statistic is a weighted average of differences between a subpopulation linking function and the total group linking function (Yang, 2004). The formula of the $RESD_j$ is defined below:

$$\mathrm{RESD}_j = \frac{\sqrt{E_P\left\{ \left[e_{P_j}(x) - e_P(x) \right]^2 \right\}}}{\sigma_{Y_P}} = \frac{\sqrt{\sum_{x=0}^{Z} w_{xp} \left\{ \left[e_{P_j}(x) - e_P(x) \right]^2 \right\}}}{\sigma_{Y_P}},$$

where j denotes a subpopulation, $E_P\ \{\ \}$ denotes averaging over raw score levels weighted by the relative number of candidates at each score level in the total population P, Z is the maximum possible raw score, w_{xP} is $\frac{n_x}{n}$ in the total population P, and $\Sigma w_{xP} = 1$. Note that n_x is the number of candidates at a raw score level of x, and n is the total number of candidates. In this study, the RESD_j is in the metric of the standard deviation of the CLEP scaled score.

REMSD

REMSD (Dorans & Holland, 2000) is used to summarize linking differences across score levels and subpopulations. Its formula is as follows:

$$\text{REMSD} = \frac{\sqrt{\sum_{j=1}^{J} w_j E_P\left\{\left[e_{P_j}(x) - e_P(x)\right]^2\right\}}}{\sigma_{Y_P}}.$$

To be more expressive, the formula for REMSD can be rewritten as follows:

$$\text{REMSD} = \frac{\sqrt{\sum_{j=1}^{J} w_j \sum_{x=0}^{Z} w_{xP}\left[e_{P_j}(x) - e_P(x)\right]^2}}{\sigma_{Y_P}} \text{ or } \frac{\sqrt{\sum_{x=0}^{Z} w_{xP} \sum_{j=1}^{J} w_j\left[e_{P_j}(x) - e_P(x)\right]^2}}{\sigma_{Y_P}}.$$

REMSD is a double-weighted average of differences between subpopulation linking functions and the total group linking function. In this study, it is a measure of overall equatability in the metric of the standard deviation of the CLEP scaled score.

Hypothetical Total Group

If all the candidates had taken the same test form instead of the 16 different forms of the CLEP College Algebra exam, the size of the total group for the form would be 16,121, which is the sum of the observed sample sizes across the 16 forms (see Table 2). Using such a hypothetical total group for each of the forms allowed workarounds for potential problems due to sampling variability, especially when observed sample sizes were small, for computing equatability indices.

The frequency distribution of the hypothetical total group was estimated via the probability density function for the ability (θ) estimate, produced by IRT-based equating with the Rasch model. A standard normal distribution with a mean of 0 and standard deviation of 1 was assumed for the θs of the hypothetical total group. The frequency estimation procedure for the hypothetical total group is summarized in the appendix. In computing equatability indices, the estimated frequencies and the proportions of gender groups in the hypothetical total group were used as weights. The drawback of using a hypothetical group is that errors may occur in estimating its frequency distribution, which may then affect the equatability outcomes.

In addition to using the hypothetical total group data to control for sampling errors in computing equatability indices, equatability indices were also computed using the data from each of the observed total groups. By contrasting the two sets of outcomes for each form, the appropriateness of the hypothetical and observed total group data could be evaluated.

Difference That Matters With the CLEP Exams

As mentioned earlier, CLEP scores are used for making decisions about granting college-level course credits. The pass/fail decision depends on how a CLEP candidate's score compares to the

recommended cut score. For CLEP candidates a change in the pass/fail decision is the *difference that matters* (DTM). Therefore, evaluation of linking differences focused on CLEP test score levels near the pass/fail cut score and comparison of the pass/fail classification outcomes resulting from various linkings.

On the CLEP 20-80 score scale, half a score unit at the pass/fail threshold is crucial because it may result in a reverse decision on pass/fail status. Therefore, the DTM for CLEP was quantified to be half a CLEP scaled score unit when that score was at the threshold. For comparisons across various exam forms, the DTM was further expressed in the standard deviation unit such that it represented the standard-score equivalent of half a CLEP score. This standardized DTM is called *SDTM* in this article.

The SDTM in standard deviation units is useful in evaluating subpopulation invariance of functions that link number-correct scores and the CLEP scaled scores. RMSD/REMSD statistics can be compared to the SDTM to determine whether the linking differences are practically significant. The linking differences across subpopulations are considered negligible if the differences represented by the RMSD/REMSD are less than the SDTM. The practice of ignoring differences that are less than half a score reporting unit has been used for years in the equating practices of major testing programs, such as the SAT (Dorans & Feigenbaum, 1994).

Computationally, dividing .5 by the standard deviation of the CLEP scores in the total population gives the SDTM. Over the 16 forms of the College Algebra exam, the estimated standard deviation of the CLEP scores in the hypothetical total group ranged from 9.46 to 9.54. Accordingly, the SDTM based on the hypothetical total group data ranged from .052 to .053. This article uses SDTM* to denote the SDTM based on the hypothetical total group data throughout, which should be differentiated from the SDTM that was based on the observed total-group data.

The SDTM based on the observed total group data can be obtained by dividing .5 by the sample standard deviation of the observed total group, treating the sample standard deviation as the standard deviation of the total population for each form. As shown in Table 4, the observed standard deviation of the CLEP scores ranged from 10.96 to 12.04 over the 16 forms. Therefore, the SDTM based on the observed total group data ranged from .042 to .046 for these 16 forms.

Results and Discussion

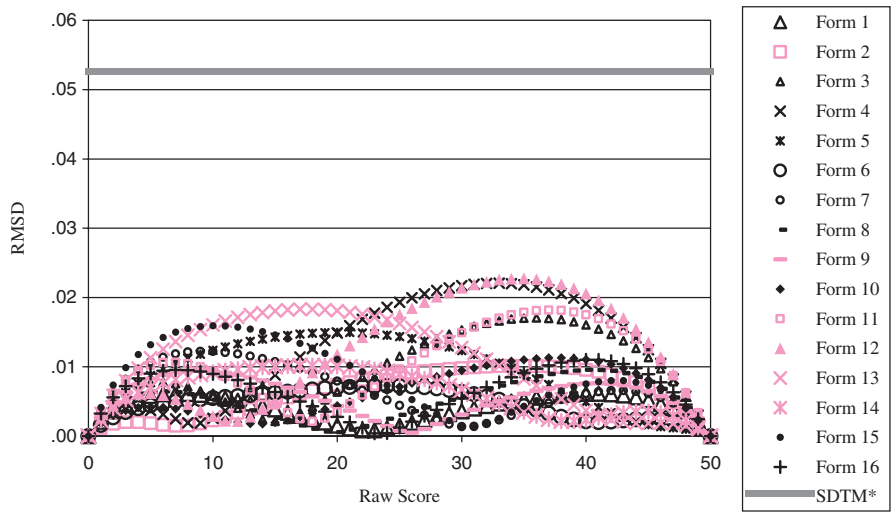
Equatability outcomes based on the hypothetical total group for the 16 forms of the CLEP College Algebra exam are presented in this section, followed by an evaluation of the impact of linking variation on pass/fail classifications.

Equatability of the CLEP College Algebra Exam

The RMSD outcomes for various forms of the College Algebra exam are presented in Figure 1, and Figure 2 zeros in on the linking differences between the total group and the gender subgroups. The $RESD_j$ and REMSD results are shown in tables.

RMSD results. Figure 1 depicts the RMSD outcomes at various raw score levels for the 16 forms, which are compared to a thick line representing a range of SDTM* values across various forms. To facilitate the interpretation of the RMSD outcomes, Figure 2 further depicts the direction and magnitude of linking differences between the total group and each of the gender subgroups. Linking differences were transformed to have the same unit as the SDTM* in Figure 2 to facilitate comparisons. Statistics in both Figures 1 and 2 are compared to the SDTM* for practical significance because they are based on the hypothetical total group data and have the same unit as

Figure 1
Linking Differences at Raw Score Levels for Various Forms of the CLEP College Algebra Exam



Note. CLEP = College-Level Examination Program; RMSD = root mean square difference; SDTM = standardized difference that matters.

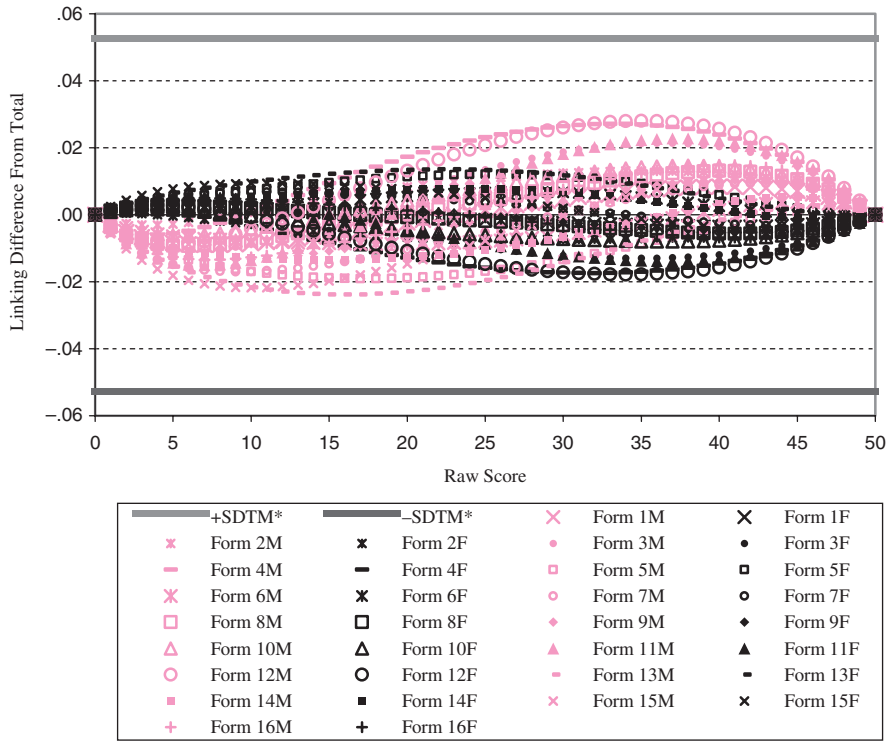
the SDTM*, which is in the metric of the standard deviation of the CLEP scores in the total population.

Figure 1 shows that none of the RMSD measures across forms was larger than .025. All of the RMSD values were much smaller than the SDTM*, which ranged from .052 to .053. This suggests that on all forms linking differences at various score levels were negligible. The multiple forms of the testlet-based CLEP College Algebra exam are designed to be highly comparable in content and statistical performance because of the common content and statistical specifications. This may account for the consistent findings in linking variations over gender subpopulations across various forms in Figure 1.

Figure 2 shows that neither the linking for the female group nor the linking for the male group deviated from the linking for the total group by more than .03 units at various raw score levels for each form. Because the linking differences across forms were smaller than the SDTM*, they were negligible. In Figure 2, the lines for the female group are generally closer to the zero line for the total group than the lines for the male group. This indicates that the linkings for the female group were more similar to the linkings for the total group, which was mainly because there were more females than males in the total group.

The largely consistent patterns in linking differences between gender subgroups across forms shown in Figure 2 may simply suggest that in general the College Algebra forms are similar to one another and the groups taking various forms were fairly randomly equivalent. The causes for the general patterns, however, may not have a simple or straightforward answer. In practice, equating results depend on various characteristics of equating samples (e.g., examinees' ability levels and their interaction with test forms) while the other equating conditions (e.g., test form and equating method) are held constant. Equating sample characteristics are generally reflected in score distributions of samples. Thus, the patterns of linking differences observed on the CLEP College Algebra exam could be attributed to differences in score distributions between male and female groups.

Figure 2
Linking Differences Between Each Gender Group and Total Group at Raw Score Levels for Various Forms of the CLEP College Algebra Exam



Note. CLEP = College-Level Examination Program; SDTM = standardized difference that matters.

The raw score means by gender across forms presented in Table 3 show that the male group had slightly higher mean scores than the female group on all but one form, and scores for the male group were slightly more spread out than the female group on most of the forms. This finding helps to explain why the linkings for males exceed those for females in the higher score regions and females exceed males in the lower score regions on most of the College Algebra forms. Nevertheless, despite the interest in finding reasons for such linking differences, the small differences, which were so small that they were all negligible from a practical point of view, should not be distracting.

REMSD and RESD$_j$ results. Table 5 shows the REMSD and the RESD$_j$ statistics for the 16 forms, respectively. As with the RMSD statistics presented earlier, the statistics in Table 5 are based on the hypothetical total group. Over the 16 forms, the REMSD ranged from .0038 to .0177, well below the SDTM*, which ranged from .052 to .053. This suggests that the linkings between the number-correct scores and the CLEP scaled scores were invariant over gender subpopulations for the CLEP College Algebra exam. The linking differences between the total and gender groups were negligible.

The RESD$_j$ outcomes for the gender subpopulations in Table 5 are further examined. As expected, the very small RESD$_j$ for both the male and female groups suggested a negligible linking difference between each subgroup and the total group. The linking in the female group was

Table 5
Equatability Measures of the CLEP College Algebra Exam
Using Weights Based on the Hypothetical Total Group Data

		Equatability Index		
		$RESD_j$ for Individual Subgroup		
Test Form		Male	Female	REMSD
1	A1B1C1D1V1	.0052	.0021	.0038
2	A1B1C1D2V1	.0102	.0066	.0084
3	A1B1C2D1V1	.0145	.0096	.0120
4	A1B1C2D2V1	.0213	.0144	.0177
5	A1B2C1D1V1	.0154	.0105	.0128
6	A1B2C1D2V1	.0070	.0055	.0062
7	A1B2C2D1V1	.0091	.0042	.0067
8	A1B2C2D2V1	.0071	.0028	.0051
9	A2B1C1D1V1	.0073	.0028	.0052
10	A2B1C1D2V1	.0093	.0059	.0076
11	A2B1C2D1V1	.0145	.0093	.0118
12	A2B1C2D2V1	.0202	.0137	.0168
13	A2B2C1D1V1	.0181	.0116	.0147
14	A2B2C1D2V1	.0099	.0065	.0082
15	A2B2C2D1V1	.0121	.0058	.0091
16	A2B2C2D2V1	.0084	.0034	.0061

Note. The equatability indices were computed using estimated frequencies for a hypothetical total group as weights. The hypothetical total group encompassed all candidates taking various alternate forms ($N = 16{,}121$). Its frequency distribution was estimated based on the θ estimate, which resulted from item response theory–based equating and had a standard normal distribution. In addition, the expected proportional weights were used for the two gender groups in computing the equatability indices. Specifically, the weight was .428 for the male group and .572 for the female group. CLEP = College-Level Examination Program; $RESD_j$ = root expected square difference; REMSD = root expected mean square difference.

a bit more similar to that in the total group than was the linking in the male group was because the majority of the total group was female. This was consistent with the findings from Figures 1 and 2.

Table 6 shows the equatability results, using weights based on observed data for the total group and the subgroups. The results in Table 6 are very similar in magnitude to those in Table 5. The REMSD in Table 6 ranges from .0033 to .0143, well below the SDTM, which ranges from .042 to .0406. The similarities between Tables 5 and 6 suggest that the assumptions made about the hypothetical total group are likely to hold, and the observed sample sizes of various forms are probably large enough that they are not subject to much sampling error.

Impact of Linking Variation on Pass/Fail Classification Consistency

Although the invariance of raw score to CLEP scaled score linking is important because it indicates whether the scores are equatable, it is of practical interest to further study the impact of linking differences on pass/fail classifications given a cut score for the CLEP exam. Ultimately, it is the pass/fail decision that matters to CLEP candidates. Specifically, the pass/fail classification outcomes based on the linkings in the gender subpopulations were compared to the classification

Table 6
Equatability Measures of the CLEP College Algebra Exam Using Weights Based on Observed Data

			Equatability Index		
			$RESD_j$ for Individual Subgroup		
Test Form		*N*	Male	Female	REMSD
1	A1B1C1D1V1	995	.0046	.0018	.0033
2	A1B1C1D2V1	1,041	.0082	.0052	.0066
3	A1B1C2D1V1	1,035	.0125	.0080	.0102
4	A1B1C2D2V1	1,003	.0172	.0114	.0140
5	A1B2C1D1V1	1,079	.0115	.0078	.0095
6	A1B2C1D2V1	1,013	.0052	.0039	.0045
7	A1B2C2D1V1	957	.0078	.0034	.0057
8	A1B2C2D2V1	980	.0066	.0027	.0048
9	A2B1C1D1V1	1,045	.0065	.0025	.0046
10	A2B1C1D2V1	1,018	.0079	.0048	.0064
11	A2B1C2D1V1	1,017	.0132	.0083	.0107
12	A2B1C2D2V1	1,003	.0173	.0116	.0143
13	A2B2C1D1V1	980	.0141	.0091	.0115
14	A2B2C1D2V1	987	.0078	.0050	.0063
15	A2B2C2D1V1	1,009	.0100	.0047	.0074
16	A2B2C2D2V1	959	.0079	.0033	.0058

Note. Observed frequencies for the total group and observed proportional weights for gender subgroups were used in computing the REMSD statistics. CLEP = College-Level Examination Program; $RESD_j$ = root expected square difference; REMSD = root expected mean square difference.

outcomes based on the linkings in the total group. The recommended cut score of 50 on the CLEP scale was used to make pass/fail decisions.

Table 7 highlights the unrounded linking outcomes on the CLEP score scale near the recommended cut-score level for various linkings across the College Algebra exam forms. It also provides differences between subgroups and total group linkings. Note that the unrounded CLEP scores around 49.5, instead of 50, are presented in Table 7 because a CLEP score of 49.5 would be rounded up to become 50.

Table 7 shows that differences between the unrounded subgroups and total group linkings near the cut-score levels across forms were all smaller than .5, which suggests that the differences are not practically significant. This is consistent with the small RMSD values presented earlier in Figures 1 and 2. As a result, for each of the College Algebra forms the pass/fail rate was not expected to vary across different linkings. However, because operationally the pass/fail decision was made with rounded CLEP scores (i.e., candidates earning a rounded CLEP score of 50 or above would pass the exam, whereas candidates with a rounded score below 50 would fail), the pass/fail rate on a form may vary across linkings due to rounding.

For each of the College Algebra forms, Table 8 highlights the rounded linking outcomes at or near the recommended cut-score level for various linkings. The far right-hand column of Table 8 shows the percentage of candidates in the total group earning a CLEP score at or near the cut score of 50 for each form. The percentage data in Table 8 was used to evaluate the impact of differences between rounded linking outcomes on the pass/fail decision.

Table 7
Differences Between Unrounded Linking Outcomes Near the CLEP Cut-Score Level

Form	Number-Correct Score	Corresponding CLEP Scaled Score (Unrounded) Based on Linking			Linking Difference	
		Linking in Total Group	Linking in Male Group	Linking in Female Group	Male – Total	Female – Total
1	24	49.95	49.96	49.94	.01	–.01
	23	48.83	48.83	48.82	.00	–.01
2	24	50.30	50.40	50.23	.10	–.07
	23	49.16	49.26	49.09	.09	–.07
3	24	50.06	50.18	49.97	.12	–.09
	23	48.93	49.03	48.84	.10	–.08
4	24	50.41	50.63	50.27	.21	–.15
	23	49.26	49.46	49.12	.20	–.14
5	24	50.17	50.00	50.28	–.16	.11
	23	49.06	48.89	49.17	–.17	.11
6	24	50.52	50.44	50.58	–.08	.06
	23	49.39	49.31	49.45	–.08	.06
7	24	50.28	50.22	50.32	–.06	.04
	23	49.15	49.08	49.20	–.07	.04
8	24	50.63	50.66	50.61	.03	–.02
	23	49.49	49.51	49.48	.02	–.01
9	24	49.93	49.91	49.93	–.02	.00
	23	48.80	48.77	48.81	–.03	.00
10	24	50.28	50.35	50.22	.07	–.06
	23	49.13	49.20	49.08	.06	–.05
11	24	50.04	50.12	49.96	.09	–.07
	23	48.90	48.97	48.83	.07	–.07
12	24	50.39	50.58	50.26	.19	–.13
	23	49.24	49.41	49.11	.17	–.13
13	24	50.14	49.95	50.27	–.19	.13
	23	49.03	48.83	49.16	–.20	.13
14	24	50.49	50.39	50.57	–.10	.07
	23	49.36	49.25	49.44	–.11	.07
15	24	50.25	50.17	50.31	–.09	.05
	23	49.13	49.03	49.18	–.10	.06
16	24	50.61	50.61	50.61	.01	.00
	23	49.46	49.46	49.47	–.01	.00

Note. CLEP = College-Level Examination Program.

As shown in Table 8, when the converted CLEP scores based on various linkings were rounded, the pass/fail rates based on different linkings for the same form were the same in general. Only on Form 8, the pass/fail rate was not invariant across various linkings. Specifically, the pass/fail rate based on the male group linking was different from the pass/fail rate based on the total group linking for Form 8 when the CLEP scores were rounded, despite the negligible linking difference exhibited in the unrounded CLEP scores (see Table 7). For candidates earning a raw score of 23 on Form 8 (see Table 8), if the total group linking were used they would have a rounded CLEP score

Table 8
Differences Between Rounded Linking Outcomes Near the CLEP Cut-Score Level

Form	Number-Correct Score	Corresponding CLEP Scaled Score (Rounded) Based on Linking			% of Candidates in the Total Group at the Raw Score Level
		Linking in Total Group	Linking in Male Group	Linking in Female Group	
1	24	50	50	50	2.91
	23	49	49	49	2.51
2	24	50	50	50	2.79
	23	49	49	49	4.03
3	24	50	50	50	3.57
	23	49	49	49	2.32
4	24	50	51	50	4.39
	23	49	49	49	2.59
5	24	50	50	50	3.52
	23	49	49	49	2.87
6	24	51	50	51	3.16
	23	49	49	49	2.67
7	24	50	50	50	2.61
	23	49	49	49	3.03
8	24	51	51	51	3.16
	23	**49**	**50**	**49**	**3.47**
9	24	50	50	50	3.92
	23	49	49	49	2.68
10	24	50	50	50	3.54
	23	49	49	49	4.03
11	24	50	50	50	3.34
	23	49	49	49	2.95
12	24	50	51	50	3.39
	23	49	49	49	3.69
13	24	50	50	50	3.67
	23	49	49	49	2.86
14	24	50	50	51	3.85
	23	49	49	49	2.74
15	24	50	50	50	3.77
	23	49	49	49	3.07
16	24	51	51	51	3.55
	23	49	49	49	3.13

Note. Bold numbers indicate forms with differences in pass/fail rates (after rounding) for linkings based on different groups.

of 49, but the rounded CLEP score would be 50 if the male group linking were used. The difference in the pass/fail rate on Form 8 was clearly due to rounding instead of linking variation. Fortunately, the percentage of candidates who could have been inadvertently advantaged by the rounding practice was small, less than 3.5% in the total group (see Table 8).

In short, the pass/fail rates that resulted from linkings in the gender subpopulations were generally consistent with the outcomes that resulted from linkings in the total group, and the only inconsistent case (on Form 8) can be attributed to rounding. Therefore, it is reasonable to conclude that

for the CLEP College Algebra exam there is no substantial impact of linking variation, if there is any, across gender subpopulations on the pass/fail classification.

Summary and Suggestions

This research demonstrates how population invariance checks could be applied to evaluate linking outcomes involving IRT-based equating for testlet-based exams. Overall, the linking outcomes were invariant over gender subpopulations for all of the 16 forms of the CLEP College Algebra exam. The linking differences between the gender groups and the total group were very small for all of the forms. Even for Form 14, for which the candidate group and/or the test form looked somewhat different from the rest, the equatability indices were small enough to suggest negligible linking differences. Equatability outcomes based on the hypothetical and the observed total groups were very similar, which provided evidence of the tenability of the equatability measures in this study.

The correlation coefficient corrected for attenuation between CLEP College Algebra testlets that represent various major content areas ranged from .89 to .97, which suggested a strong dominant dimension. By testlet model design, test content across the 16 College Algebra forms are well balanced. In addition, literature has shown the robustness of the IRT unidimensionality assumption, such that violations of unidimensionality might not have a substantial impact on equating (Camilli, Wang, & Fesq, 1995; Dorans & Kingston, 1985; Reckase, Ackerman, & Carlson, 1988). As a result, a unidimensional IRT model is likely to be sufficient for linking CLEP College Algebra forms.

Nonetheless, future research should further investigate the robustness of the unidimensionality assumption (e.g., by examining local independence) for CLEP linkings, especially for the other CLEP exams that are prone to multidimensionality problems, because unidimensional IRT procedures are likely to yield inconsistent ability estimates if a test is influenced by more than one equally potent dimension (Reckase, 1979). In such studies, equatability indices can still be used to detect possible linking differences and to evaluate the adequacy of the use of the Rasch model. Despite the fact that significant linking differences might not be attributable to either problematic equating or inadequate use of the Rasch model because of their confounded effects on equating outcomes, it is still important to detect and document such linking differences for operational enhancement. If there is no significant linking difference in such cases, the use of the Rasch model is likely to be sufficient.

In estimating frequency distributions for the hypothetical total groups in this study, a standard normal distribution was assumed for the ability of the hypothetical total group. This assumption of a standard normal ability distribution seems reasonable based on the observed raw and scale score distributions as well as the wide-ranging CLEP candidate background, training in college-level Algebra, and preparation level for the exam. Table 3 of this study shows the observed total-group raw score means and standard deviations across forms. Based on the data of about 1,000 candidates for each form, it seems fair to assume a normal ability distribution for the underlying candidate population. A recent large CLEP operational test dataset accumulated over time further shows a scale score mean and median of 51, only 1 point off the midpoint of 50 on the 20-80 CLEP score scale. This further supports the normality assumption. In addition, CLEP candidates generally vary widely in age, educational background, course-taking experience, and preparation level for the CLEP exam because there is no prerequisite for taking the CLEP exam. As a result, CLEP test takers include college students, homeschooled students, working parents or adult students, international students, people looking to advance or change their careers, people seeking additional education or a credential, military service members and veterans, and so forth. This adds support for the normality assumption.

Nevertheless, one could still raise a concern about the aforementioned normality assumption, especially when the mean scores shown in Table 3 of this study were all slightly above the

midpoint of the 0-50 raw score scale, which suggests that the underlying ability distribution could be slightly negatively skewed. To ease this concern, future studies should thoroughly examine the normality assumption for ability distribution and work with a nonnormal distribution if necessary.

Appendix
Procedure for Deriving the Frequency Distributions for the Hypothetical Total Groups

A standard normal distribution (with $\mu = 0$ and $\sigma = 1$) was assumed for the ability (θ) of the hypothetical total group for each of the 16 forms of the College-Level Examination Program (CLEP) College Algebra exam. Based on this assumption, for each raw score level of an exam form, the standard normal probability density function $f(\widehat{\theta}) = (\frac{1}{\sqrt{2\pi}})\, e^{\frac{-\widehat{\theta}^2}{2}}$ was used to estimate the probability density in the hypothetical total group for the ability estimate ($\widehat{\theta}$) resulting from the IRT-based equating with the Rasch model. Based on the resulting probability density estimates, the relative frequencies were then calculated for different raw score levels for the hypothetical total group. This estimation procedure was applied to all 16 forms of the CLEP exam. See Table A1 for an example of such estimation outcomes.

Assuming that all the examinees took the same exam form instead of the 16 different forms, the size of the hypothetical total group for each of the 16 forms would be 16,121, which is the sum of the observed sample sizes across the 16 forms. By multiplying the estimated relative frequencies for a form by 16,121, the frequency estimates were obtained for the hypothetical total group for each of the 16 forms.

Table A1
Example Outcomes of Estimating Frequencies at Various Raw Score Levels for the Hypothetical Total Group on a CLEP College Algebra Exam Form

Raw Score	θ Estimate	Standard Normal Probability Density	Relative Frequency
30	.418705	.3655	.0393
29	.321705	.3788	.0408
28	.2259	.3889	.0418
27	.1309	.3955	.0426
26	.036507	.3987	.0429
25	–.05752	.3983	.0429
24	–.151501	.3944	.0424
23	–.2457	.3871	.0417
22	–.340322	.3765	.0405
21	–.435708	.3628	.0390

Note. CLEP = College-Level Examination Program.

Notes

1. The American Council on Education (ACE) conducts periodic reviews of the College-Level Examination Program (CLEP), including the processes to determine the recommended credit-granting score. The ACE recommends a credit-granting score of 50 for various CLEP exams, effective July 2001.

2. Although the common reference form used in this study differed from the testlet-based new forms in test construction, length, and scoring method, this reference form was useful in maintaining the credit-granting standard for the CLEP College Algebra exam before a new standard could be established on the testlet-based form via standard setting. Because a standard-setting study was conducted recently and since May 2005 all the testlet-based forms have been equated to a new testlet-based reference form, the linking procedure described in this study is no longer used operationally for the College Algebra exam.

References

Camilli, G., Wang, M. M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement*, *32*, 79-96.

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: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281-306.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, *22*, 249-262.

Hambleton, R. K., & Swaminathan, H. (1990). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193-203.

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

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating and their relationship to population invariance. *Journal of Educational Measurement*, *41*, 15-32.

Yang, W.-L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement*, *41*, 33-41.

Acknowledgments

The authors sincerely thank Neil Dorans for his insightful suggestions and comments on the design and analyses of this study. We also would like to acknowledge Annie Nellikunnel's assistance in running computer programs for IRT item calibrations and equatings as well as Brad Moulder's input from both methodology and CLEP operational perspectives. We also thank Dan Eignor, Rosemary Reshetar, and Cathy Wendler for their reviews of an earlier draft and APM *editors and reviewers for their thoughtful comments and suggestions. We are also grateful to the College Board for the use of the CLEP data for this research.*

Author's Address

Address correspondence to Wen-Ling Yang, Educational Testing Service, Rosedale Road, Princeton, NJ 08541; e-mail: wyang@ets.org.