Journal of Research Methodology, Volume 23, Number 3 (September-December 2010)

วารสารวิธีวิทยาการวิจัย ปีที่ 23 ฉบับที่ 3 (กันยายน-ธันวาคม 2553)

การประเมินความยุติธรรมของคะแนนจากการสอบด้วยคอมพิวเตอร์สำหรับคัดเลือกนักเรียนเข้ามหาวิทยาลัยแพทยศาสตร์ของสหรัฐอเมริกา

สังวรณ์ งัดกระโทก

Mark Reckase

บทคัดย่อ

การศึกษานี้ใช้วิธีการประเมินความยุติธรรมของคะแนนประเมินความเท่าเทียมกันของคะแนนสอบที่ใช้สมัครเข้าเรียนมหาวิทยาลัยแพทย์ของประเทศสหรัฐอเมริกา ระหว่างคะแนนจากการทดสอบด้วยรูปแบบการตอบลงบนกระดาษ และการทดสอบด้วยคอมพิวเตอร์ และประเมินความเท่าเทียมกันของคะแนนสอบด้วยรูปแบบคอมพิวเตอร์ จำนวน 3 ฉบับ ที่มาจากแบบทดสอบต่างฉบับกัน ในวิชาวิทยาศาสตร์ชีวภาพ วิทยาศาสตร์กายภาพ และตรรกะทางภาษา การศึกษาครั้งนี้ประเมินความเท่าเทียมกันของคะแนนจากวิธีการปรับเทียบคะแนนจากคะแนนการทดสอบรูปแบบการตอบลงบนกระดาษกับคะแนนจากการทดสอบด้วยคอมพิวเตอร์ และการปรับเทียบคะแนนจากการทดสอบด้วยคอมพิวเตอร์จากแบบทดสอบสามฉบับของแต่ละรายวิชา ผลการวิเคราะห์พบว่าฟังก์ชันการปรับเทียบคะแนนระหว่างกลุ่มชายและหญิงไม่แตกต่างกัน ทั้งสามวิชา

Evaluating Score Equity of Computerized MCAT

Sungworn Ngudgratoke
Mark Reckase

ABSTRACT

This study used a score equity assessment (SEA) to evaluate the comparability of scores on the Medical College Admission Test MCAT between the paper & pencil-based MCAT and the computer- based MCAT and across the three different test forms of the computer-based MCAT. Specifically, this study evaluated if the equating functions were invariant across gender groups (male vs. female) for the Biological Science (BS), Physical Science (PS), and Verbal Reasoning (VR) in two different conditions. One was when scores on the computerized MCAT test were linked to scores on the paper & pencil-based MCAT test. Another was when the three different forms of computerized MCAT tests were linked. The results indicated that overall population invariance of linking functions holds across gender groups.

Background

The Medical College Admission Test (MCAT) is a standardized, multiple-choice examination designed to assess problem solving, critical thinking, and writing skills in addition to the examinee's knowledge of science concepts and principles prerequisite to the study of medicine. Scores are reported in each of the following areas: Verbal Reasoning, Physical Sciences, Writing Sample, and Biological Sciences. Medical college admission committees consider MCAT scores as part of their admission decision process and almost all U.S. medical schools require applicants to submit MCAT scores during the application process.

There has been a dramatic change in MCAT exam since January 2007. The traditional MCAT test was delivered through a mode of paper & pencil. However with the advantages of computerized-based testing over paper-pencil testing only the computerized exam has been delivered since in January 2007. Changing from paper & pencil-based test (PBT) to computerized-based test (CBT) is expected to reduce the test length by about one-third, without changing the content. In addition the shorter version of the MCAT test would retain its predictive power. The change also allows testing time for MCAT to be reduced by 30 minutes.

It is necessary that scores on the computerized MCAT be fully comparable to and interchangeable with scores on the paper-pencil MCAT. This assumption is essential for allowing test users to track scores across test administration conditions and to enable U.S. medical schools to treat scores from paper & pencil tests and computerized tests equally when making an admission decision. Moreover, there is often a demand that scores for the computerized tests be on the same scale as the paper-pencil tests to the extent possible, so that scores from computerized test can be used more or less interchangeably with scores from paper & pencil test.

To ensure that scores from computerized test can be used interchangeably with scores from paper-pencil test, test equating is used to adjust scores on two different testing conditions so that scores are comparable. For the MCAT exam, however, changing from paper & pencil testing to computerized testing will have implications

for test equating. This is because reducing the number of items, test specifications, and testing times as well as changing the testing environment from paper & pencil test to computerized test will likely influence the validity of MCAT scores and test score conversions from one test to another. It is suggested that when test structure, context, and administration conditions are changed there is a need to investigate if such changes still allow comparability between the paper & pencil and computerized tests (von Davier, 2007; Eignor, 2007; Liu, Cahn, & Dorans, 2006).

So, according to the changes in test administration for the MCAT, the major question is: How can we assess the comparability of scores from the computerized test compared to scores from the paper-pencil test? To answer this question, score equity assessment (SEA) introduced by Dorans (2004) can be used to evaluate if scores on paper & pencil test and scores on computerized test are comparable. In general, SEA focuses on whether or not scores that are supposed to be used interchangeably are in fact interchangeable (Liu, Cahn, & Dorans, 2006) and uses population invariance of linking functions across subpopulations, such as gender groups or ethnic groups, to assess the degree of interchangeability of test scores. Dorans and Holland (2000) consider subpopulation invariance to be the most important requirement of equating two tests, because "fulfilling this requirement will also imply that the tests measure the same thing and are equally reliable." If two tests measure different things and/or are not equally reliable, then the equating results will not be invariant for some subgroups. The concept of population invariance and the estimation of standard errors of population invariance statistics will be explained in the next section.

Objectives of this study

The purpose of this study was to apply SEA to evaluate the comparability of MCAT test scores between paper & pencil-based MCAT and computer-based MCAT and across three different test forms of computer-based MCAT. Specifically, this study evaluated if the equating functions were invariant across gender groups in two different conditions. One condition was when the computerized MCAT test was linked to the paper & pencil MCAT test. Another was when the three different forms of the computerized MCAT tests were linked.

The following sections describe the concept of population invariance and its measures. Next, the research methodology, data analyses, and results were presented. The last section discusses the summary and the conclusion as well as the future research.

Population Invariance Measures

In large scale assessment, multiple test forms are widely used because of test security. That is, examinees taking the test measuring the same construct are usually administered different test forms. For example one group of students take form X while another takes form Y. To compare performance on tests for students from these two groups, test equating methods are used to produce test scores from one test form that can be used interchangeably with scores from the other test form. In other words, in order to fairly compare performance of the two groups of students taking different test forms, test equating methods are used to adjust for difficulty in the two different test forms.

It is important to note that not every test form can be meaningfully equated to another form. Tests that are equated should meet the five requirements as stated by Doran and Holland (2000). Specifically, tests being equated to each other should measure the same construct, have equal reliability, and satisfy the symmetry, equity, and population invariance requirements. The population invariance requirement is the focus of this study. Dorans and Holland (2000) stated that if the population invariance holds, "the choice of (sub)population used to compute the equating function between the scores of test X and Y should not matter. In other words, the equating function used to link the scores of X and Y should be population invariant." They also proposed two measures used to determine if the population invariance requirement is met. One of the two measures they proposed, Root Mean Square Difference (RMSD), is defined as the root mean square difference between the subpopulation linking functions and the overall linking function. That is, if scores on test Y (e.g., the 'new' form) are equated to scores on test X (e.g., the 'old' form) on a target population of examinees P, then

$$\text{RMSD}(y) = \frac{\sqrt{\sum_c w_c \left[e_{P_c}(y) - e_P(y) \right]^2}}{\sigma_{XP}}, \quad (1)$$

where e_{P_c} is the equating function for subpopulation P_c (e.g., male vs. female) in population P; e_P is the linking function for the overall population P; w_c is the weight assigned to subpopulation P_c; and σ_{XP} is the standard deviation of the scores on X for population P. Note that the equating function is the equated score that can be obtained using an equating method such as an observed score equating method (e.g., equipercentile equating method) and the item response theory-based equating method (e.g., Stocking-Lord method).

The RMSD is interpreted similarly to an effect size (Dorans & Holland, 2000). Specifically, it is the amount of invariance of the linking function for subpopulation c to the overall linking function. So, a value of .15 for RMSD(y), say y = 80, is interpreted as a root mean square difference of 15% of the standard deviation of test X in P in the linking functions at score 80 of test Y.

Another measure of population invariance, the Root Expected Mean Square Difference (REMSD), is defined as a summative measure of the values of the RMSD(y),

$$\text{REMSD} = \frac{\sqrt{\sum_c w_c E_P \left\{ \left[e_{P_c}(Y) - e_P(Y) \right]^2 \right\}}}{\sigma_{XP}}, \quad (2)$$

where Y is a random score from test Y from the population P; and $E_P\{\cdot\}$ is the averaging over the square difference between the subpopulation linking functions from the overall linking function.

The REMSD is interpreted similarly to the RMSD. For example, if REMSD = .30, then the root mean square difference across the scores of test Y is, on average, 30% of the standard deviation of test X in P in the linking functions.

Research Method

1. Data

The data used for assessing comparability of the gender subgroup equating functions across test forms of the computerized MCAT test were drawn from the three forms (Form 34, Form 35, and Form 36) of the April 2007 test administration. However, one test form (Form 32) of the paper & pencil data and one test form (Form 3) of the MCAT computerized testing data, both administered in August 2006, were used to investigate the comparability of the gender subgroup equating functions when the reported scores on the paper & pencil-based testing (PBT) and the computer-based testing (CBT) were linked.

Data for the biological science (BS), physical science (PS), and verbal reasoning (VR) test sections measuring three different knowledge and skills that are prerequisite for study in medicine were analyzed separately. Only scaled scores and a variable related to gender of MCAT test takers were used in this study. For the information about gender of the test takers, male (M) and female (F) were used; examinees providing "D" as their gender identification were deleted.

2. Data Analysis

To evaluate the equity of computerized MCAT scores across gender groups, the equipercentile equating method was used. Steps taken to analyze data were as follows.

1. Analyze descriptive statistics.

Descriptive statistics describing the distribution of MCAT scores including mean, standard deviation, minimum, and maximum were reported separately by test form, test section, gender, and mode of test administration (PBT vs. CBT).

2. Compute RMSD and REMSD

2.1 Pre-smooth MCAT scores using a log-linear model in order to remove irregularities in the data due to sampling variations and to remove peaks associated with formula scoring of the tests (Liu, Cahn, & Doran, 2006).

2.2 Equate scores on the computerized MCAT to scores on the paper & pencil MCAT using the total group (all examinees).

2.3 Repeat 2.2 for male and female subgroups separately.

2.4 Compute $RMSD_i$ and REMSD using (1) and (2), respectively. Note that to calculate $RMSD_i$ and REMSD, the weights in the RMSD and REMSD formula were set equally for male and female subgroups (i.e., $w_0 = w_1 = .5$) which means that the same weight was given to each of the equating functions for the sub-populations (von Davier, Holland & Thayer, 2004).

3. Step 2 was also applied when the three different forms of the computerized MCAT were linked.

4. Evaluate comparability of linking functions across gender groups

It is useful to examine whether or not the difference between male and female subgroup linking functions has any important consequences for reported scores (Holland & Dorans, 2006). Traditionally, this examination has been carried out by using a "difference that matters" (DTM) that is used to indicate whether the scores resulting from an equating function would change the reported score. Dorans and Feigelbaum (1994) called a difference in reported score points a DTM if "the testing program considered it to be a difference worth worrying about." For example, MCAT testing program reports the scaled score ranging from 1 to 15 with the increment of 1. The DTM is ±.5 reported score point because in the 1 to 15 scaled score of MCAT, scores are rounded and reported in steps of 1 point. Thus, differences of less than ±.5 points are not distinguished in the reporting scale and can be ignored.

To evaluate the sensitivity of subpopulation linking functions, RMSD and REMSD are commonly compared to the DTM. If an estimate of RMSD or REMSD is less than .5, a difference between male and female equating functions can be ignored. However, σ_{XP} in the formula (1) is used to quantify the sum differences between total population and subpopulation linked scores in standard deviation units. In the present study, the equating converted the MCAT 1-to-15 scaled scores on one test form to another. It is needed to transform DTM such that it has the same unit as

RMSD and REMSD do. So in this study, a DTM was standardized by dividing 0.5 by σ_{XP}. An estimate of RMSD or REMSD was then compared to the standardized DTM (SDTM). If an estimate of RMSD or REMSD is less than a chosen SDTM, a difference between male and female equating functions can be ignored (von Davier & Manalo, 2006).

Results

The results of this study are presented in the following order. Part I presents the evaluation of the comparability of linking functions derived by equating the reported CBT test scores to the reported PBT test scores, while part II presents the evaluation of the comparability of linking functions derived when the reported test scores from the three different test forms of the computer version of MCAT were linked. Both Part I and II were intended to present the estimated RMSD and REMSD and their corresponding SDTM values, all computed to evaluate the comparability of linking functions between male and female subgroups.

Part I. Evaluation of comparability of linking functions when PBT and CBT were linked.

Table 1 and Table 2 respectively present descriptive statistics for scaled scores from PBT and CBT tests that were administered in August 2006 and April 2007, respectively. As seen in these tables, there were more examinees taking the PBT exam than those taking the CBT exam and one of the explanations for that is that the computerized MCAT was just initialized at the time. The means of the scaled scores of BS, PS, VR for the male groups were higher than those for the female examinees across the modes of test administration, suggesting that overall males performed better than females across different modes of testing. For the BS and the PS, the scaled scores from the PBT test were higher than scaled scores from the CBT test, implying that the PBT test may be slightly easier than the CBT. However, for the VR section, males performed better than females when taking the CBT, while females seemed to perform better than males when taking the PBT.

Table 1 Descriptive Statistics for BS, PS, and VR section of PBT

Section	gender	n	Mean	SD	Maximum	Minimum
BS	Male	16,777	9.014	2.513	15	1
	Female	20,265	8.159	2.546	15	1
	Total	37,042	8.547	2.567	15	1
PS	Male	16,777	8.958	2.521	15	2
	Female	20,265	7.793	2.316	15	1
	Total	37,042	8.320	2.479	15	1
VR	Male	16,777	8.042	2.517	15	1
	Female	20,265	7.636	2.531	15	1
	Total	37,042	7.820	2.532	15	1

Table 2 Descriptive statistics for BS, PS, and VR section of CBT

Section	gender	n	Mean	SD	Maximum	Minimum
BS	Male	1,483	8.164	2.609	14	1
	Female	1,420	7.160	2.596	14	1
	Total	2,903	7.673	2.650	14	1
PS	Male	1,483	8.292	2.591	15	1
	Female	1,420	6.907	2.253	15	1
	Total	2,903	7.615	2.528	15	1
VR	Male	1,483	8.118	2.618	15	1
	Female	1,420	7.573	2.695	15	1
	Total	2,903	7.851	2.670	15	1

Even though the focal statistics to evaluate the comparability of the MCAT tests were RMSD and REMSD that were presented in the following sections, the transformed scores that were computed separately for the total group and the gender subgroups were also presented. The transformed scores in this study were actually the equated scores obtained by linking the CBT to the PBT using the equipercentile

equating method (Kolen & Brennan, 2004, p. 43-46). The transformed scores separately computed for the total group and subgroups add more detailed information regarding which gender subgroup would have an advantage or a disadvantage when the total group was used for equating as commonly performed in practices.

Table 3 presents the transformed scores, and the estimates of RMSD as well as their corresponding SDTM criteria, reported at each score point of the MCAT scaled scores. Even though the differences between male and female equating functions were observed, the results indicate that those differences were ignorable because they were relatively small when compared with the SDTM criteria. This suggests that the invariance of the male and the female linking functions was met when the three sections of the computer-based MCAT were linked to the same sections of the paper & pencil MCAT.

Specifically, for the BS section, the estimates of RMSD ranged from 0.011 to 0.050. All the estimates of RMSD were less than the SDTM of 0.195, suggesting that the invariance of the linking across gender subgroups was met. The estimates of RMSD for the PS section ranged from 0.015 to 0.155. All the estimates of RMSD were less than the SDTM of 0.202. For the VR section, the estimates of RMSD ranged from 0.009 to 0.053. All the estimates of RMSD were also less than the SDTM of 0.198.

Table 4 presents the estimated REMSD statistics for the BS, the PS, and the VR sections of MCAT. Obviously, the estimates of REMSD for the BS, the PS, and the VR were 0.030, 0.074 and 0.036, respectively, which were less than their corresponding SDTM criteria of 0.195, 0.202, and 0.198. The estimates of REMSD give the additional information to illuminate that the differences in the equating functions for the BS, the PS, and the VR sections of MCAT between males and females were ignorable when the scores obtained from the CBT were linked to those obtained from the PBT.

Table 3 Estimates of RMSD for BS, PS, and VR sections

Section	Scaled Score	Transformed Score			RMSD	SDTM
		Total	Male	Female		
BS	1	1.523	1.585	1.507	0.017	0.195
	2	2.647	2.761	2.618	0.032	0.195
	3	3.808	3.885	3.825	0.022	0.195
	4	4.960	4.962	5.056	0.026	0.195
	5	6.046	5.991	6.200	0.045	0.195
	6	7.053	6.983	7.219	0.050	0.195
	7	8.006	7.953	8.153	0.043	0.195
	8	8.937	8.914	9.054	0.033	0.195
	9	9.859	9.867	9.953	0.026	0.195
	10	10.773	10.807	10.851	0.023	0.195
	11	11.660	11.714	11.732	0.025	0.195
	12	12.494	12.551	12.571	0.026	0.195
	13	13.390	13.404	13.440	0.014	0.195
	14	14.245	14.248	14.305	0.017	0.195
	15	15.067	15.066	15.107	0.011	0.195
PS	1	1.648	2.108	1.356	0.155	0.202
	2	2.708	3.012	2.580	0.094	0.202
	3	3.678	3.730	3.672	0.015	0.202
	4	4.666	4.612	4.747	0.028	0.202
	5	5.679	5.588	5.816	0.047	0.202
	6	6.719	6.631	6.885	0.054	0.202
	7	7.782	7.729	7.956	0.052	0.202
	8	8.840	8.836	9.014	0.050	0.202
	9	9.848	9.877	10.034	0.054	0.202
	10	10.766	10.798	10.992	0.065	0.202
	11	11.593	11.607	11.883	0.083	0.202
	12	12.393	12.380	12.713	0.091	0.202
	13	13.232	13.204	13.479	0.071	0.202
	14	14.117	14.096	14.371	0.073	0.202
	15	15.045	15.048	15.291	0.070	0.202
VR	1	1.285	1.391	1.209	0.037	0.198
	2	2.293	2.328	2.293	0.010	0.198
	3	3.278	3.237	3.351	0.024	0.198
	4	4.232	4.139	4.362	0.045	0.198
	5	5.151	5.041	5.305	0.053	0.198
	6	6.056	5.961	6.199	0.048	0.198
	7	6.975	6.911	7.090	0.037	0.198
	8	7.922	7.887	8.008	0.026	0.198
	9	8.892	8.877	8.957	0.018	0.198
	10	9.871	9.868	9.922	0.014	0.198
	11	10.841	10.849	10.880	0.011	0.198
	12	11.778	11.808	11.794	0.009	0.198
	13	12.646	12.734	12.589	0.029	0.198
	14	13.477	13.632	13.416	0.047	0.198
	15	14.472	14.678	14.365	0.065	0.198

Table 4 Estimates of REMSD for BS, PS, and VR

Test Section	REMSD	SDTM
BS	0.030	0.195
PS	0.074	0.202
VR	0.036	0.198

Part II. Evaluation of comparability of linking functions when the three different test forms of the computerized MCAT were linked

Table 5 shows the descriptive statistics of the scores on the three test forms (Form 34, Form 35, and Form 36) of the computerized MCAT, separately for the three test sections (BS, PS, and VR). As seen in Table 5, even though the scaled scores were equally distributed between male and female subgroups, the male subgroup performed slightly better than did the female subgroup in all test sections.

Table 5 Descriptive statistics for the three computerized MCAT test forms

Test form	Gender	N	M	SD	Maximum	Minimum
BS 34	Male	1333	9.345	2.267	15	1
	Female	1540	8.699	2.255	15	1
	Total	2873	8.999	2.283	15	1
BS 35	Male	1127	9.749	2.471	15	1
	Female	1059	8.688	2.766	15	1
	Total	2186	9.235	2.671	15	1
BS 36	Male	1123	9.207	2.291	15	1
	Female	1175	8.657	2.387	15	1
	Total	2298	8.926	2.356	15	1
PS 34	Male	1333	9.361	2.455	15	1
	Female	1539	8.117	2.255	15	2
	Total	2872	8.694	2.430	15	1
PS 35	Male	1097	8.828	2.591	15	2
	Female	1092	7.712	2.419	14	1
	Total	2189	8.271	2.567	15	1
PS 36	Male	1080	9.455	2.594	15	1
	Female	1154	8.262	2.399	15	2
	Total	2234	8.838	2.565	15	1
VR 34	Male	1333	8.462	2.507	15	1
	Female	1540	8.138	2.473	15	1
	Total	2873	8.288	2.494	15	1
VR 35	Male	1112	8.397	2.347	15	1
	Female	1079	7.996	2.360	15	1
	Total	2191	8.199	2.362	15	1
VR 36	Male	1131	8.141	3.009	15	1
	Female	1169	7.620	2.977	15	1
	Total	2300	7.876	3.003	15	1

The estimates of RMSD and the transformed scores that were computed separately for the total group and the gender subgroups were presented in the table 6 to 8. Note that for this part the estimated RMSD and REMSD values were obtained when the three different test forms of the computerized MCAT were linked to each other.

When the three different test forms (Form 34, 35 and 36) of the BS section were linked to each other, as seen in Table 6 nearly all the estimated RMSD were less than the SDTM criteria. However when the Form 35 was linked to the Form 34 of the BS section of the computerized MCAT, the differences between male and female equating functions were obviously observed at the scaled score of 1 to 3. The estimates of RMSD associated with these score points were 0.375, 0.372, and 0.269, respectively, which were larger than the SDTM of 0.219. At the scaled score of 1 to 3, males would have had lower scores if male-only equating was carried out-using the total group equating was beneficial to males who were relatively lower achievement test takers.

In addition, when the Form 36 was linked to the Form 35, the differences between male and female equating functions were also pronounced at the scaled score of 1 and 2. Specifically, the estimates of RMSD associated with the scaled score of 1 and 2 were larger than the SDTM criteria of 0.187. This suggests that, at the scaled score of 1, the males would have had scaled scores of 2 rather than 1 if males-only equating was carried out. However at the scaled score of 2 females would have had scaled score of 1 rather than 2 if female-only equating was carried out.

When the Form 36 was linked to the Form 34, the population invariance requirement holds throughout the wide range of the MCAT scaled score. Specifically, the differences between the male and female equating functions were ignorable, evidenced by the fact that all the estimates of RMSD were less than the SDTM criteria of 0.219.

For the PS section of the computerized MCAT, table 7 presents the transformed scores that were separately computed for the total group and the gender subgroups,

and the estimated RMSD as well as the SDTM criteria. All of these statistics were obtained when the three different test forms of the PS section were linked to each other. The results show that the differences between male and female equating functions were relatively small. Specifically, the estimates of RMSD that were obtained when Form 35 was linked to Form 34 ranged from 0.010 to 0.076, and they were less than the SDTM criteria of 0.206. The estimates of RMSD that were obtained when the Form 36 was linked to the Form 35 ranged from 0.008 to 0.148, and they were less than the SDTM criteria of 0.195. Also, when the Form 36 was linked to the Form 34, the estimates of RMSD ranged from 0.017 to 0.153, and they were less than the SDTM of 0.206.

Therefore when the three different forms of the PS section were linked to each other, the observed male and females equating functions were insignificant.

For the VR section of the computerized MCAT, table 8 presents the transformed scores that were separately computed for the total group and the gender subgroups, and the estimated RMSD as well as the SDTM criteria. All of these statistics were obtained when the three different test forms of the VR section were linked to each other. The results were similar to the results in the table 7 in that when the three different test forms of the VR section were linked to each other separately by gender subgroups, the male and female equating functions were quite similar and all the RMSD values were less than the SDTM criteria.

Specifically, when the Form 35 was linked to the Form 34, the estimates of RMSD ranged from 0.007 to 0.107 and they were less than the SDTM criteria of 0.200. The estimates of RMSD that were obtained when the Form 36 was linked to the Form 35 ranged from 0.003 to 0.191 and they were less than the SDTM criteria of 0.212. When the Form 36 was linked to the Form 34, the estimates of RMSD ranged from 0.004 to 0.177 and they were also less than the SDTM criteria of 0.200.

Table 6 Estimates of RMSD for the BS section

Forms linked	Scaled Score	Transformed score			RMSD	SDTM
		Total	Male	Female		
35 to 34	1	1.730	0.705	2.375	0.375	0.219
	2	2.876	1.812	3.430	0.372	0.219
	3	3.884	3.106	4.267	0.269	0.219
	4	4.764	4.288	5.031	0.169	0.219
	5	5.569	5.234	5.772	0.121	0.219
	6	6.289	6.038	6.538	0.110	0.219
	7	7.027	6.836	7.267	0.095	0.219
	8	7.844	7.694	8.052	0.079	0.219
	9	8.719	8.608	8.889	0.063	0.219
	10	9.615	9.552	9.745	0.045	0.219
	11	10.490	10.490	10.570	0.025	0.219
	12	11.383	11.418	11.407	0.013	0.219
	13	12.277	12.324	12.289	0.015	0.219
	14	13.164	13.190	13.186	0.011	0.219
	15	14.221	14.199	14.271	0.017	0.219
36 to 35	1	0.755	1.626	0.598	0.234	0.187
	2	1.657	2.493	1.215	0.250	0.187
	3	2.597	3.155	2.216	0.179	0.187
	4	3.594	3.975	3.333	0.122	0.187
	5	4.683	4.956	4.484	0.089	0.187
	6	5.852	6.131	5.615	0.097	0.187
	7	7.050	7.392	6.751	0.120	0.187
	8	8.225	8.552	7.902	0.122	0.187
	9	9.358	9.608	9.067	0.102	0.187
	10	10.452	10.625	10.233	0.074	0.187
	11	11.535	11.634	11.386	0.047	0.187
	12	12.672	12.689	12.529	0.038	0.187
	13	13.823	13.827	13.739	0.022	0.187
	14	14.757	14.802	14.698	0.020	0.187
	15	15.334	15.346	15.326	0.004	0.187
36 to 34	1	1.506	1.266	1.563	0.076	0.219
	2	2.529	2.341	2.602	0.063	0.219
	3	3.522	3.318	3.607	0.068	0.219
	4	4.446	4.257	4.550	0.067	0.219
	5	5.305	5.190	5.401	0.046	0.219
	6	6.164	6.144	6.206	0.014	0.219
	7	7.066	7.126	7.054	0.019	0.219
	8	8.023	8.130	7.971	0.037	0.219
	9	9.020	9.139	8.945	0.044	0.219
	10	10.035	10.145	9.953	0.043	0.219
	11	11.053	11.142	10.973	0.037	0.219
	12	12.059	12.123	11.992	0.029	0.219
	13	13.030	13.071	12.989	0.018	0.219
	14	13.935	13.963	13.938	0.009	0.219
	15	14.868	14.879	14.907	0.013	0.219

Table 7 Estimates of RMSD for the PS section

Forms linked	Scaled Score	Transformed score			RMSD	SDTM
		Total	Male	Female		
35 to 34	1	1.712	1.527	1.785	0.058	0.206
	2	2.772	2.691	2.820	0.027	0.206
	3	3.700	3.662	3.745	0.017	0.206
	4	4.642	4.643	4.676	0.010	0.206
	5	5.596	5.646	5.607	0.015	0.206
	6	6.555	6.662	6.534	0.032	0.206
	7	7.515	7.681	7.455	0.051	0.206
	8	8.471	8.679	8.389	0.065	0.206
	9	9.418	9.631	9.326	0.068	0.206
	10	10.344	10.527	10.253	0.059	0.206
	11	11.256	11.393	11.173	0.046	0.206
	12	12.172	12.267	12.093	0.036	0.206
	13	13.110	13.180	13.021	0.033	0.206
	14	14.077	14.142	13.946	0.043	0.206
	15	15.067	15.151	14.820	0.076	0.206
36 to 35	1	0.629	0.832	0.541	0.061	0.195
	2	1.394	1.777	1.019	0.148	0.195
	3	2.556	2.848	2.258	0.115	0.195
	4	3.626	3.818	3.510	0.062	0.195
	5	4.627	4.724	4.552	0.034	0.195
	6	5.571	5.594	5.541	0.010	0.195
	7	6.469	6.412	6.487	0.016	0.195
	8	7.360	7.243	7.419	0.036	0.195
	9	8.293	8.147	8.374	0.046	0.195
	10	9.290	9.149	9.370	0.045	0.195
	11	10.363	10.269	10.413	0.029	0.195
	12	11.478	11.448	11.481	0.008	0.195
	13	12.591	12.586	12.551	0.011	0.195
	14	13.639	13.619	13.575	0.018	0.195
	15	14.578	14.495	14.500	0.031	0.195
36 to 34	1	0.946	1.200	0.696	0.104	0.206
	2	2.109	2.538	1.804	0.153	0.206
	3	3.235	3.566	3.010	0.117	0.206
	4	4.226	4.517	4.097	0.093	0.206
	5	5.156	5.348	5.091	0.059	0.206
	6	6.066	6.181	6.038	0.034	0.206
	7	6.967	7.041	6.958	0.022	0.206
	8	7.862	7.919	7.860	0.017	0.206
	9	8.765	8.823	8.762	0.017	0.206
	10	9.701	9.771	9.685	0.021	0.206
	11	10.688	10.772	10.646	0.027	0.206
	12	11.717	11.806	11.633	0.036	0.206
	13	12.753	12.837	12.608	0.049	0.206
	14	13.758	13.836	13.500	0.078	0.206
	15	14.702	14.772	14.470	0.070	0.206

Table 8 Estimates of RMSD for the VR section

Forms linked	Scaled Score	Transformed score			RMSD	SDTM
		Total	Male	Female		
35 to 34	1	1.212	0.972	1.503	0.107	0.200
	2	2.149	1.976	2.313	0.068	0.200
	3	3.047	2.974	3.114	0.028	0.200
	4	3.959	3.952	3.982	0.007	0.200
	5	4.913	4.920	4.936	0.007	0.200
	6	5.919	5.902	5.970	0.015	0.200
	7	6.959	6.911	7.041	0.027	0.200
	8	8.015	7.950	8.108	0.032	0.200
	9	9.075	9.016	9.159	0.029	0.200
	10	10.147	10.111	10.207	0.020	0.200
	11	11.251	11.250	11.279	0.008	0.200
	12	12.419	12.457	12.413	0.011	0.200
	13	13.896	13.992	13.860	0.029	0.200
	14	15.014	15.005	15.072	0.017	0.200
	15	15.406	15.388	15.439	0.011	0.200
36 to 35	1	1.474	1.571	1.313	0.056	0.212
	2	2.830	2.890	2.778	0.024	0.212
	3	4.229	4.271	4.191	0.017	0.212
	4	5.391	5.442	5.335	0.023	0.212
	5	6.263	6.307	6.207	0.021	0.212
	6	6.992	7.010	6.958	0.011	0.212
	7	7.696	7.692	7.687	0.003	0.212
	8	8.378	8.349	8.394	0.010	0.212
	9	9.079	9.030	9.122	0.020	0.212
	10	9.816	9.759	9.870	0.024	0.212
	11	10.497	10.466	10.537	0.015	0.212
	12	11.269	11.272	11.254	0.005	0.212
	13	11.988	12.084	11.847	0.051	0.212
	14	12.540	12.885	12.351	0.118	0.212
	15	13.482	14.048	13.191	0.191	0.212
36 to 34	1	1.619	1.550	1.673	0.025	0.200
	2	2.890	2.867	2.919	0.010	0.200
	3	4.167	4.210	4.153	0.013	0.200
	4	5.282	5.335	5.261	0.016	0.200
	5	6.180	6.193	6.188	0.004	0.200
	6	6.951	6.921	6.996	0.015	0.200
	7	7.692	7.629	7.769	0.028	0.200
	8	8.416	8.313	8.532	0.044	0.200
	9	9.159	9.048	9.286	0.048	0.200
	10	9.955	9.849	10.079	0.046	0.200
	11	10.758	10.682	10.857	0.035	0.200
	12	11.527	11.558	11.523	0.009	0.200
	13	12.408	12.570	12.291	0.057	0.200
	14	13.373	13.834	13.017	0.165	0.200
	15	14.676	15.031	14.162	0.177	0.200

Table 9 shows the estimates of REMSD and their corresponding SDTM criteria. Again, a REMSD is a single index summarizing values of RMSD at each scaled score. Table 9 presents the estimates of REMSD for BS, PS and VR, respectively. It was evidenced that all the estimated REMSD were also less than the SDTM criteria, thus overall when the three different test forms of computerized MCAT were linked to each other, the equating functions were population invariant across gender groups.

Table 9 Estimates of REMSD and SDTM

Test section	Test Forms linked	REMSD	SDTM
BS	Form 35 to 34	0.169	0.219
	Form 36 to 34	0.044	0.219
	Form 36 to 35	0.124	0.187
PS	Form 35 to 34	0.047	0.206
	Form 36 to 34	0.072	0.206
	Form 36 to 35	0.059	0.195
VR	Form 35 to 34	0.038	0.200
	Form 36 to 34	0.069	0.200
	Form 36 to 35	0.063	0.212

Across the analyses presented above, overall the equating functions obtained by equating the computerized MCAT to the paper & pencil MCAT were invariant across gender groups. Also, the equating functions obtained by equating the three different test forms of the computerized MCAT were also invariant across gender groups.

Conclusions

Examination for the invariance of linking function using population invariance statistics is one way to assess score equity of the MCAT test. This study investigated the invariance of the linking functions across gender groups in the linkage of the computerized MCAT to the paper & pencil-based MCAT and in the linkage of all possible pairs of the three different test forms of the computerized MCAT. The major research questions of this study were (1) whether changes from paper & pencil-based

MCAT to computer-based MCAT would result in differential linking functions on gender groups, and (2) whether different forms of the computerized MCAT would result in differential linking functions on gender groups. If the relationship between two test forms being equated depends on whether examinees are male or female, then the tests are probably not measuring the same thing with comparable degrees of reliability (Liu, Cahn, & Dorans, 2006).

The results of this study indicated that overall population invariance requirement holds across gender groups, even though there are some gender dependencies of linking functions at some points of the MCAT scaled score, especially at the low end of the MCAT scaled score for the BS section, but there are only a few points on the MCAT scaled score that have such dependencies. The linking functions for the BS section of the computerized MCAT that differ between male and female groups at the low end of the MCAT scaled score suggest that males and females who are considered relatively low achievement test takers might respond to items presented in the BS section differently and this might be because of a guessing factor of test takers or another factor associated with the computerized MCAT itself.

Based on the invariance of equating functions found in this study, this result provides another empirical evidence to corroborate that different forms of computerized MCAT and that paper & pencil MCAT test and computerized MCAT test measure the same construct with comparable reliability. This study also highlights that computerized MCAT scores and paper & pencil MCAT scores can be used interchangeably. Also scores obtained from different forms of the computerized MCAT can be used interchangeably. In addition, the results indicate that the degree of dis/advantages between male and female subgroups by the use of total group equating is unnoticeable, suggesting that test fairness related to gender holds for the MCAT examination.

There are possible criticisms of these analyses that need to be pointed out. First, this study used the uniform weights in the RMSD and REMSD equations, meaning that the proportion of males and the proportion of females in the population of MCAT test takers are thought to be equivalent. It would be interesting to conduct further analyses by varying the weights in the RMSD and REMSD equations in order

to investigate whether such changes would result in similar results. Second, this study used the equipercentile equating and assumed that test takers were equivalent. The descriptive statistics shows that the males perform slightly better than the females in terms the means. So, the assumption made in this study might not be perfectly satisfied.

Some future research related to population invariance of equating functions across some subgroups of MCAT test takers should be undertaken. First, we can use the MCAT test data to examine the invariance in other subgroups of MCAT test takers, such as ethnic groups, and language groups. Second, using other equating methods through item response theory (IRT) to assess population invariance requirement is an interesting area for future MCAT research.

Given that the SEA assessment has a premise for evaluating comparability of scores on different test forms, it will be useful for testing programs that use different test forms used for multiple testing occasions. The Ordinary National Education (ONET) of Thailand is an example of a testing program offering several testing occasions annually to students. It is necessary for ONET test developers to collect different types of evidences through validation processes to show ONET users that scores on ONET are comparable across occasions and that it is fair to students taking different test forms administered at different occasions, especially when educators and policy makers urge students to take the test only once. In addition, because educational inequality between urban and rural schools of Thailand is evidenced, it is interesting to assess if ONET scores will be fairly and validly used for college admission across groups of students from urban and rural schools. To answer such inquiries, SEA is one of methods that will be useful in this regard. However, SEA is conducted through a score equating procedure and; hence, it will not be relevant for testing programs not performing a score transformation.

References

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

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, M. Feryok, A. P. Schmitt, & N. K. Wright, *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS RM-94-10). Princeton, NJ: Educational Testing Service.

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

Dorans. N. J. (2004). Using population invariance to assess test score equity. *Journal of educational measurement, 41*(1), 43-68.

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales (pp. 89-106)*. New York: Springer.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement (4th ed., pp. 181-220)*. New York: American Council on Education and Praeger.

Liu, J., Cahn, M. F., & Dorans, N. J. (2006). An application of score equity assessment: Invariance of linkage of the new SAT to old SAT across gender groups. *Journal of educational measurement, 43*(2), 113-129.

Ngudgratoke, S. Manalo, J. R., von Davier, A. A. (2007). *The empirical standard errors for two population invariance measures for the equipercentile equating case.* Paper accepted for presentation at the Annual Meeting of American Education Research Association, April 9-13, Chicago.

von Davier, A. A., Holland, P W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer.

von Davier, A.A., Manalo, J. R. (2006). *Theoretical and Empirical Standard Errors for Two Population Invariance Measures in the Linear Equating Case.* Paper presented at the National Council on Measurement in Education (NCME) Annual Meeting, San Francisco, CA.

von Davier, A. (2007). Potential solutions to practical equating issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales (pp. 89-106).* New York: Springer.