ORIGINAL RESEARCH

Comparisons of item difficulty and passing scores by test equating in a basic medical education curriculum

Jung Eun Hwang[1], Na Jin Kim[2] and Su Young Kim[1,2,3]

[1]Department of Pathology and [2]Master Center for Medical Education Support, College of Medicine, The Catholic University of Korea and [3]Department of Biomedicine and Health Sciences, The Catholic University of Korea, Seoul, Korea

Purpose: Test equating studies in medical education have been conducted only for high-stake exams or to compare two tests given in a single course. Based on item response theory, we equated computer-based test (CBT) results from the basic medical education curriculum at the College of Medicine, the Catholic University of Korea and evaluated the validity of using fixed passing scores.

Methods: We collected 232 CBTs (28,636 items) for 40 courses administered over a study period of 9 years. The final data used for test equating included 12 pairs of tests. After test equating, Wilcoxon rank-sum tests were utilized to identify changes in item difficulty between previous tests and subsequent tests. Then, we identified gaps between equated passing scores and actual passing scores in subsequent tests through an observed-score equating method.

Results: The results of Wilcoxon rank-sum tests indicated that there were no significant differences in item difficulty distribution by year for seven pairs. In the other five pairs, however, the items were significantly more difficult in subsequent years than in previous years. Concerning the gaps between equated passing scores and actual passing scores, equated passing scores in 10 pairs were found to be lower than actual passing scores. In the other two pairs, equated passing scores were higher than actual passing scores.

Conclusion: Our results suggest that the item difficulty distributions of tests taught in the same course during successive terms can differ significantly. It may therefore be problematic to use fixed passing scores without considering this possibility.

***Key Words*:** Universities, Educational measurement, Academic performance

Introduction

Basic medical education curricula often utilize cut-off scores in multiple-choice examinations to gauge whether students have achieved their learning goals [1-3]. Although it is recommended not to use arbitrary passing scores in education, medical schools still use arbitrary cutoffs to determine passing scores [3]. Medical students at our institution must score 70% or more on tests to pass courses. For practical reasons, it is costly and infeasible for content experts to examine all items and to determine minimum levels of achievement [4], since courses are taught by teams of experts and assessments are made via several tests administered throughout each course. Therefore, when students score 70% or more on tests in

Received: April 15, 2019 • Revised: May 16, 2019 • Accepted: May 17, 2019

Corresponding Author: Su Young Kim (https://orcid.org/0000-0003-3066-3246)
Department of Pathology, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Korea
Tel: +82.2.2258.7315 Fax: +82.2.537.6586 email: suyoung.dr@gmail.com

Korean J Med Educ 2019 Jun; 31(2): 147-157.
https://doi.org/10.3946/kjme.2019.126
eISSN: 2005-7288

any course, which is generally equivalent to a grade of C, they are considered to have achieved their learning goals in the course.

In addition to setting standards arbitrarily, what are considered passing scores are fixed over the years at some medical education institutions [2,5]. In a previous study of a licensing exam, the use of fixed passing scores was found to potentially cause problems associated with pass/fail decisions [6]. If a test in a given year was significantly easier than in past years, examinees could earn high scores and pass the test even though they do not achieve the expected level. Another study found that test difficulty was related to variation in examinees' failure rates when pre-fixed cut-off scores were used [4]. The disadvantages of fixed passing scores indicate a need to adjust passing scores to ensure that they are equivalent to those of previous terms through the process of test equating.

In this study, we applied test equating based on item response theory (IRT) [7], in which a unique item characteristic curve (ICC) is determined for each item to indicate the probability that an examinee with specific ability will respond correctly to the item. Mathematical models of IRT may be used to express item difficulty, discrimination, and guessing parameters on ICCs. Item difficulty ('b') is a corresponding ability parameter defined when the probability of a correct response is 0.5. Item discrimination ('a') is defined as the ICC slope when the probability of a correct response to the item is 0.5. Item guessing parameter ('c') is the probability that an examinee with no ability will answer the item correctly. One-parameter models, or Rasch models, calculate only item difficulty. Two-parameter models add item discrimination to determinations of item difficulty. Three-parameter models add an item guessing parameter to item difficulty and discrimination.

The principal advantage of IRT is the invariance of item and ability parameters [8,9]. In other words, item parameters are not dependent on examinees' abilities, and ability parameters are not dependent on item parameters if the IRT model shows good model-data fit and meets the assumptions of unidimensionality and local independence. The assumption of invariance is valid when items and ability parameters are represented by the same ability scale [7]. Thus, the ability parameter of a student may be different if tests they took are calibrated using different ability scales.

Test equating locates two tests on a common ability scale so that the scores of the tests can be interchangeably compared. Test equating is required to compare different forms of tests. Putting two tests (X and Y) on the same ability scale requires linear transformation of ability (θ) as shown below [10].

Equation (1): $\theta_x=A\theta_y+B$

A and B are equating coefficients in a linear equation (1). Relationships of item parameters of two tests are as follows:

$a_x=a_y/A$

$b_x=Ab_y+B$

$c_x=c_y$

Methods used for calculating equating coefficients in IRT include mean/mean, mean/sigma, and characteristic curve methods [6,10].

Until now, test equating based on IRT has been used in medical education mostly for high-stakes examinations or to equalize two tests given in the same course [6,11,12]. Yim and Huh [6] equated the 2004 Medical Licensing Examination in Korea to the 2003 exam, identified changes of item and ability parameters between the tests, and judged the equity of passing scores between them. Two other studies performed test equating for two tests given in the same course over 2 years [11,12]. However, to the best of our knowledge, few researchers have attempted to equate all of the tests given in courses that are part of a basic medical education curriculum.

Using fixed cutoffs to determine passing scores requires that the item and ability parameters of tests remain constant year after year. In order to investigate the validity of using fixed passing scores in this manner, we examined whether there were significant differences in items according to year after performing test equating. We also identified gaps between scores equated to previous passing scores and actual passing scores in subsequent tests. The aim of this study was to equate in-house computer-based test (CBT) results accumulated in our basic medical education curriculum, to test change of item difficulty distributions, and to verify that passing scores of previous and subsequent years were equivalent.

Methods

The equating design used in this research was non-equivalent groups with common items, a method that is generally used in test equating [10], because our examinee groups were not equivalent by year, and common items were used in the tests. The common items included in this study were the same for clinical vignettes, lead-in questions, keys, and distractors.

1. Objects of study

The data used for test equating were selected from CBT results from the basic medical education curriculum from 2009 to 2017 at the College of Medicine, the Catholic University of Korea. We included only multiple-choice tests. The CBT results were presented as a matrix consisting of examinees and items. Test results for 28,636 items were coded as 0s or 1s. We divided the matrix into courses, and then split the data in each course by year. The final data used for test equating were 12 pairs of tests that were administered in eight courses (Table 1). We describe how the final data were selected in the ‘procedure’ section (Fig. 1).

2. Procedure

We decided to fit our whole dataset to a Rasch model because of the small sample size [8]. First, assumptions were checked for 232 Rasch models for the 40 courses. CBT results for a year in a course became one Rasch model, which was referred to as ‘a test’ in this paper. We confirmed unidimensionality, using the unidimTest function in R package ltm, which performs a procedure that analyzes the latent dimensionality of dichotomous responses [13]. The assumption of local independence indicates that a response to an item is independent of any response to other items [8]. According to a previous study, when the criterion of unidimensionality was met, the assumption of local independence was considered to be satisfied [14]. Next, we tested model-data fit. After excluding 10 tests that did not meet the criterion of unidimensionality, we assessed goodness-of-fit for 222 tests, using the GoF function in ltm, which was based on Pearson’s chi- square [13]. A total of 78 tests showed goodness-of-fit. Out of these, 12 pairs of tests with common items (18 tests) were finally analyzed for test equating (Table 1, Fig. 1).

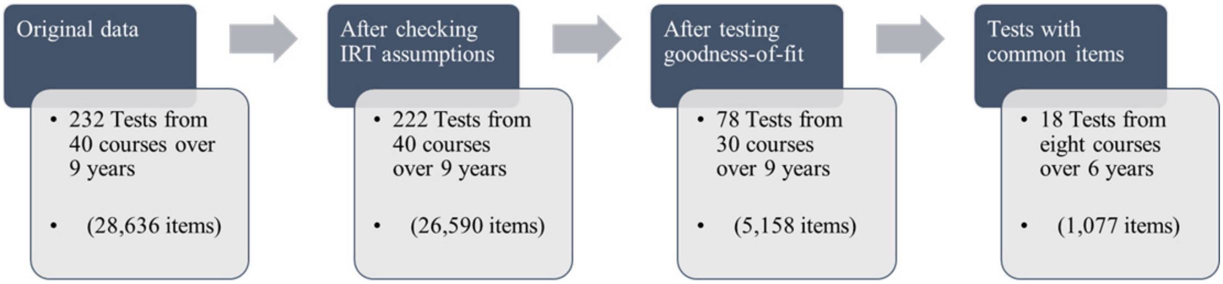
3. Analysis

For every 12 pairs, we estimated equating coefficients for re-scaling using Stocking and Lord’s characteristic curve method which is known to give stable results [15]. After transforming a scale of a subsequent test to that of a previous test in each pair, item difficulty parameters

Table 1. Pairs of Tests Used for Test Equating

No.	Course	Year	No. of examinees	No. of items	No. of common items
1	Anatomy II	2014	97	69	29
		2015	97	87	
2	Medical terminology	2010	106	21	6
		2011	97	16	
3	Medical terminology	2010	106	21	7
		2012	96	33	
4	Medical terminology	2011	97	16	6
		2012	96	33	
5	Physiology	2011	100	25	4
		2015	94	21	
6	Epidemiology	2013	97	30	4
		2014	95	22	
7	Hematology & oncology	2013	98	111	4
		2014	95	102	
8	Gastroenterology I	2013	95	206	26
		2014	97	227	
9	Evidence-based medicine II	2012	115	17	2
		2013	96	14	
10	Personalized medicine	2012	93	27	2
		2014	94	20	
11	Personalized medicine	2012	93	27	3
		2015	99	29	
12	Personalized medicine	2014	94	20	5
		2015	99	29	

Fig. 1. Tests That Passed Each Step When Selecting Final Data


Original data
• 232 Tests from 40 courses over 9 years
• (28,636 items)
After checking IRT assumptions
• 222 Tests from 40 courses over 9 years
• (26,590 items)
After testing goodness-of-fit
• 78 Tests from 30 courses over 9 years
• (5,158 items)
Tests with common items
• 18 Tests from eight courses over 6 years
• (1,077 items)

IRT: Item response theory.

were extracted from the two tests. We extracted both original item difficulty parameters and converted ones using the itm function in the R package equateIRT [16]. To examine whether difficulty parameters of common items, the other items, and all items significantly changed by year, we conducted Wilcoxon rank-sum tests.

Then, observed-scores in each pair were equated using an observed-score equating method. We used the score function in the package equateIRT [16]. We identified gaps between scores that were equated to previous passing scores and actual passing scores (70%) in subsequent tests. In this paper, we used the term 'equated passing score' to refer to a score that was equated to a previous passing score.

We used the R software package ver. 3.5.0 (https://cran.r-project.org/bin/windows/base) to perform test equating and estimate item difficulty. The level of significance for all analyses was set at p<0.05. The Institutional Review Board of the College of Medicine, the Catholic University of Korea approved the study protocol (IRB approval no., MC15EISI0121).

Results

1. Equating coefficients

In each pair, scales of subsequent tests were converted to those of previous tests. Because item discrimination parameters are all set to one in Rasch models [7], only equating coefficient Bs were reported. The equating coefficient Bs and standard errors for each pair were as follows. The equating coefficient B for anatomy II was 0.8658 (standard error [SE]=0.2033). Equating coefficient Bs for medical terminology in 2010 and 2011, in 2010 and 2012, and in 2011 and 2012 were 0.3102 (SE=0.2871), −0.1468 (SE=0.2686), and −0.2784 (SE=0.2457), respectively. The equating coefficient B for physiology was 1.3568 (SE=0.2628), for epidemiology 2.6354 (SE=0.3146), for hematology & oncology 2.3077 (SE=0.2978), for gastroenterology I 1.4301 (SE=0.1027), and for evidence-based medicine II 1.4814 (SE=0.3418), respectively. Finally, equating coefficient Bs for personalized medicine in 2012 and 2014, in 2012 and 2015, and in 2014 and 2015 were 2.0605 (SE=0.3646), 1.6216 (SE=0.2623), and 0.2147 (SE=0.2331), respectively.

2. Differences in item difficulty by year

Wilcoxon rank-sum tests revealed that two tests out of seven pairs did not have significantly different distributions for item difficulty (Table 2). The seven pairs were

Table 2. Wilcoxon Rank-Sum Test for Differences in Item Difficulty Distribution by Year

Course	Items	Year[a)]	Median	Min.	Max.	W
Anatomy II	All	2014	-2.1164	-3.8016	0.3504	1,504*
		2015	-1.0625	-2.8848	4.0288	
	Common	2014	-1.4714	-3.8016	0.3504	454
		2015	-1.967	-2.885	1.240	
	The other	2014	-2.3248	-3.8016	-0.1719	254*
		2015	-0.3023	-2.5801	4.0288	
Medical terminology	All	2010	-3.2548	-4.0900	-0.1661	120
		2011	-2.5978	-3.5193	0.5255	
	Common	2010	-3.189	-4.090	-1.848	17
		2011	-2.965	-3.519	-2.295	
	The other	2010	-3.2548	-4.0900	-0.1661	46
		2011	-2.2947	-3.5193	0.5255	
Medical terminology	All	2010	-3.2548	-4.0900	-0.1661	253
		2012	-2.628	-3.827	1.485	
	Common	2010	-2.6222	-3.4763	-0.1661	29
		2012	-2.913	-3.283	1.485	
	The other	2010	-3.611	-4.090	-0.978	99*
		2012	-2.5054	-3.8274	0.9593	
Medical terminology	All	2011	-2.9080	-3.8295	0.2154	227
		2012	-2.759	-3.959	1.354	
	Common	2011	-2.908	-3.277	-1.065	12
		2012	-2.785	-3.045	-1.508	
	The other	2011	-3.0625	-3.8295	0.2154	118
		2012	-2.759	-3.959	1.354	

(Continued to the next page)

Table 2. (Continued)

Course	Items	Year[a]	Median	Min.	Max.	W
Physiology	All	2011	-1.1733	-3.4304	3.7471	178
		2015	-0.1465	-2.4166	3.1674	
	Common	2011	-1.4965	-1.7618	-0.0926	8
		2015	-1.1320	-1.8600	-0.6343	
	The other	2011	-1.0068	-3.4304	3.7471	105*
		2015	0.4832	-2.4166	3.1674	
Epidemiology	All	2013	-1.7453	-3.7616	2.8059	118*
		2014	0.6047	-1.0922	3.9999	
	Common	2013	-0.7810	-1.8259	0.2687	8
		2014	-0.8178	-1.0921	-0.1693	
	The other	2013	-1.7856	-3.7616	2.8059	68*
		2014	1.1633	-1.0922	3.9999	
Hematology & oncology	All	2013	-1.3778	-3.4603	2.2240	1,529*
		2014	0.5515	-1.3708	5.8282	
	Common	2013	-0.9781	-3.4598	1.1702	6
		2014	-0.7275	-1.3708	-0.4586	
	The other	2013	-1.3784	-3.4603	2.2240	1,298*
		2014	0.6271	-1.3685	5.8282	
Gastroenterology I	All	2013	-1.5687	-3.5908	3.0284	10,814*
		2014	-0.1521	-2.3147	4.3765	
	Common	2013	-1.6052	-3.5901	1.0941	318
		2014	-1.2583	-2.0124	0.5305	
	The other	2013	-1.5687	-3.5908	3.0284	7,384*
		2014	0.0384	-2.3147	4.3765	
Evidence-based medicine II	All	2012	-1.2956	-3.9626	1.1447	106
		2013	-1.2695	-2.3601	1.8794	
	Common	2012	-1.2235	-2.2299	-0.2172	2
		2013	-1.1853	-2.0500	-0.3206	
	The other	2012	-1.2956	-3.9626	1.1447	79
		2013	-1.2695	-2.3601	1.8794	
Personalized medicine	All	2012	-1.0863	-3.5852	1.8627	135*
		2014	0.1151	-1.3174	4.0077	
	Common	2012	-0.0327	-1.3340	1.2687	2
		2014	-0.0325	-1.3174	1.2524	
	The other	2012	-1.086	-3.585	1.863	103*
		2014	0.1151	-1.0768	4.0077	
Personalized medicine	All	2012	-1.0863	-3.5852	1.8627	290
		2015	-0.5634	-2.1078	3.2235	
	Common	2012	-1.3340	-2.5216	1.8627	3
		2015	-0.9050	-1.1905	0.2855	
	The other	2012	-1.0572	-3.5852	1.3336	225
		2015	-0.5150	-2.1078	3.2235	
Personalized medicine	All	2014	-1.9454	-3.3779	1.9472	297
		2015	-1.9703	-3.5147	1.8166	
	Common	2014	-0.8081	-3.1372	1.6929	11
		2015	-1.1214	-2.9677	1.8166	
	The other	2014	-2.036	-3.378	1.947	188
		2015	-2.0790	-3.5147	1.6794	

[a] Subsequent tests were converted to scales of previous tests. *$p<0.05$.

medical terminology (2010 and 2011; 2010 and 2012; 2011 and 2012), physiology, evidence-based medicine II, and personalized medicine (2012 and 2015; 2014 and 2015).

In contrast, two tests among the other five pairs showed significantly different distributions for item difficulty. Wilcoxon rank-sum tests showed that items in test 2015 in anatomy II were more difficult than those in test 2014 (median: −1.0625 versus −2.1164; W=1,504; $p<0.05$). The same pattern held true for epidemiology; test items in 2014 were more difficult than those in 2013 (median: 0.6047 versus −1.7453; W=118; $p<0.05$). In hematology & oncology, items in test 2014 were more difficult than those in test 2013 (median: 0.5515 versus −1.3778; W=1,529; $p<0.05$). In gastroenterology I, test items in 2014 were harder than those in 2013 (median: −0.1521 versus −1.5687; W=10,814; $p<0.05$). In personalized medicine, items in test 2014 were more difficult than those in test 2012 (median: 0.1151 versus −1.0863; W=135; $p<0.05$). For five pairs, items in subsequent tests were significantly more difficult than those in previous tests. In addition, the other items' difficulties became harder in subsequent tests than in previous tests at $p<0.05$ (Table 2).

3. Gaps between equated passing scores and actual passing scores in subsequent tests

After test equating, scores of subsequent tests using scales of previous tests were transformed. Then, we compared gaps between actual passing scores of subsequent tests and the equated passing scores (Table 3). Among 10 pairs, the equated passing scores were lower than the actual passing scores. In anatomy II, the gap between an equated score for the 2015 tests and an actual passing score in 2015 was −20.3; the equated passing score for the 2015 test was 46.7% of the total score. The gaps for medical terminology in 2010 and 2011, in 2010 and 2012, and in 2011 and 2012 were −1.5, −2.5 and −0.2, respectively. The equated passing scores were 60.6%, 62.4%, and 69.4% of the total scores. In physiology, the gap was −4 (51.0%); in epidemiology, the gap was −7.3 (36.8%); in hematology & oncology, the gap was −37.5 (33.2%); and in gastroenterology I, the gap was −57.8 (44.5%). Gaps for personalized medicine in 2012 and 2014, and in 2012 and 2015 were −4.7 (46.5%) and −3 (59.7%), respectively. The range of gaps for the 10 pairs was −57.8 to −0.2.

In contrast, the equated passing scores were higher

Table 3. Gaps between Equated Passing Scores and Actual Passing Scores in Subsequent Tests

Course	A	B	B′	B-B′	B%
Anatomy II	48.3	40.6	60.9	-20.3	46.7
Medical terminology (2010, 2011)	14.7	9.7	11.2	-1.5	60.6
Medical terminology (2010, 2012)	14.7	20.6	23.1	-2.5	62.4
Medical terminology (2011, 2012)	11.2	22.9	23.1	-0.2	69.4
Physiology	17.5	10.7	14.7	-4	51
Epidemiology	21	8.1	15.4	-7.3	36.8
Hematology & oncology	77.7	33.9	71.4	-37.5	33.2
Gastroenterology I	144.2	101.1	158.9	-57.8	44.5
Evidence-based medicine II	11.9	10	9.8	0.2	71.4
Personalized medicine (2012, 2014)	18.9	9.3	14	-4.7	46.5
Personalized medicine (2012, 2015)	18.9	17.3	20.3	-3	59.7
Personalized medicine (2014, 2015)	14	21.4	20.3	1.1	73.8

A: A passing score on a previous test, B: An equated passing score on a subsequent test, B′: An actual passing score on a subsequent test, B−B′: A gap between an equated passing score and an actual passing score, B%: (An equated passing score/a total score on a subsequent test)×100.

than actual passing scores in two pairs. In evidence-based medicine II, the gap was +0.2 (71.4%), and the gap for personalized medicine in 2014 and 2015 was +1.1 (73.8%).

Discussion

In this study, we performed test equating using CBT results for a basic medical education curriculum. We examined whether item difficulty distributions in 12 pairs of tests varied significantly by year, and investigated gaps between equated passing scores and actual passing scores. The implications of our findings are as follows.

First, it was found that items for subsequent tests were significantly more difficult than those of previous tests for five pairs. Changes in item difficulty distribution over the years is consistent with previous research results for a national examination [6] and a course's two tests [11]. Since we investigated only 12 pairs of tests, it is cautious to generalize our results, but they demonstrate the possibility that significant differences may occur between item difficulties according to year. The significant change in item difficulty can confuse students as they prepare for tests. It implies that examiners need to set comparable items each year.

One thing notable is that the other items in the five pairs also became harder in subsequent years. These results differed from a previous study that found no significant difference between the other items over two years [12]. It is difficult to present the reason for the

Table 4. Changes in Difficulty Distribution of Common Items[a)] and Relationships with Other Items

Course	Year	Median	Min.	Max.	W	The other[b)]	All[c)]
Anatomy II	2014	-1.4714	-3.8016	0.3504	616*	More difficult	More difficult
	2015	-2.8327	-3.7505	0.3739			
Medical terminology	2010	-3.189	-4.09	-1.848	20	-	-
	2011	-3.276	-3.829	-2.605			
Medical terminology	2010	-2.6222	-3.4763	-0.1661	24	More difficult	-
	2012	-2.767	-3.136	1.632			
Medical terminology	2011	-2.908	-3.277	-1.065	12	-	-
	2012	-2.506	-2.767	-1.23			
Physiology	2011	-1.4965	-1.7618	-0.0926	16*	More difficult	-
	2015	-2.489	-3.217	-1.991			
Epidemiology	2013	-0.781	-1.8259	0.2687	16*	More difficult	More difficult
	2014	-3.453	-3.728	-2.805			
Hematology & oncology	2013	-0.9781	-3.4598	1.1702	13	More difficult	More difficult
	2014	-3.035	-3.679	-2.766			
Gastroenterology I	2013	-1.6052	-3.5901	1.0941	546*	More difficult	More difficult
	2014	-2.6884	-3.4425	-0.8996			
Evidence-based medicine II	2012	-1.2235	-2.2299	-0.2172	3	-	-
	2013	-2.667	-3.531	-1.802			
Personalized medicine	2012	-0.0327	-1.334	1.2687	3	More difficult	More difficult
	2014	-2.093	-3.3779	-0.8081			
Personalized medicine	2012	-1.334	-2.5216	1.8627	8	-	-
	2015	-2.527	-2.812	-1.336			
Personalized medicine	2014	-0.8081	-3.1372	1.6929	14	-	-
	2015	-1.336	-3.1824	1.6019			

[a)]Difficulty distribution of common items before test equating. [b)]Difficulty distribution of the other items on subsequent tests. [c)]Difficulty distribution of all items on subsequent tests. *$p<0.05$.

increase in the other items' difficulty in this study. Given the possibility that information about previous test items is passed on to junior students by senior students in Korea [17], examiners may write more difficult items for in subsequent tests [18]. Alternatively, common items that are made easier by item disclosure may have caused biases in item difficulties [11,18]. Before test equating, the difficulties of common items between two tests are expected to be similar [19]. In this study, however, common items among the four pairs of tests were significantly easier in subsequent tests than in previous tests (Table 4). The easier common items in subsequent tests could have biased the difficulty of the other items by equating. Indeed, the difficulties of the other items in subsequent tests became significantly harder than before in the four pairs of tests (Table 4). The effect of item disclosure on test equating should be investigated in future studies.

Second, the equated passing scores were found to be lower than actual passing scores in 10 pairs. This may be because items on subsequent tests were harder than those on previous tests. For the five courses in which items became harder, the equated passing scores were less than 70% of the total score. The equated passing scores ranged from 33.2% to 69.4% of the total scores. Our results imply that examinees who took later tests may not have passed, even if they had scores equal to or higher than passing scores in previous years. Taking anatomy II as an example, the passing score for 2015 should be lower because test items in 2015 were more difficult than in 2014. That is, the passing score would have to be 40.6 to be equivalent to the previous passing score (Table 3). However, the passing score for 2015 was 60.9, which was 70% of the total score. This indicates that if some examinees scored higher than 40.6 but less than 60.9 in 2015, they would have passed in 2014, but did not in 2015. Gaps were also found in pairs where item difficulty distributions did not show significant differences by year: medical terminology in 2010 and 2011, in 2010 and 2012, in 2011 and 2012; physiology, evidence-based medicine II; personalized medicine in 2012 and 2015, in 2014 and 2015 (Tables 2, 3). Apart from the remaining item difficulty being constant across tests, our results demonstrate the need for equating passing scores, and confirm previous findings that gaps affect pass/fail decisions [6]. Notably, equating passing scores is a practical task for medical students because passing or failing tests has significant effects on grade promotion.

A limitation of this study is that we examined only a small number of common items for some pairs. In five pairs (hematology & oncology, gastroenterology I, evidence-based medicine II, and personalized medicine in 2012 and 2014, and in 2012 and 2015), common items were less than 20% of the entire test (Table 1). The small percentage of common items could have caused increased random equating errors. This may be due to items or examinees being excluded for computational reasons, or many common items being avoided in the courses. The results for these five pairs thus need to be interpreted with caution.

The reason for the reduced data included in our analyses is that we excluded tests that did not meet criteria for applying IRT (Fig. 1). In particular, many tests were not included because of poor fits between the Rasch models and data. Some common items decreased or disappeared after excluding rows or columns coded as only 0s or 1s. For data reduction, it is important to take into account the context of this study, which evaluated several tests from a medical school, rather than a national exam. This source of data reduction may lead to difficulty in applying in-house tests to IRT and performing test equating in future studies.

In this research, we tried to equate in-house tests from a basic medical education curriculum. Based on these findings, we propose methods to ensure that tests are implemented consistently every year. First, items should have similar difficulty indices in different terms. Selecting items with similar difficulty from an item bank could make this possible. Second, passing scores could be equated to previous scores each year.

ORCID:

Jung Eun Hwang: https://orcid.org/0000-0002-7670-7577; Na Jin Kim: https://orcid.org/0000-0001-8278-2541; Su Young Kim: https://orcid.org/0000-0003-3066-3246

Acknowledgements: Not applicable.

Funding: Not applicable.

Conflicts of interest: No potential conflict of interest relevant to this article was reported.

Author contributions: Conception or design of the work: SYK; data analysis and interpretation: JEH, SYK; drafting the article: JEH; critical revision of the article: NJK; completion of the article: SYK; and final approval of the version to be published: all authors.

References

1. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE guide no. 37. Med Teach. 2008;30(9-10):836-845.
2. MacDougall M. Variation in assessment and standard setting practices across UK undergraduate medicine and the need for a benchmark. Int J Med Educ. 2015; 6:125-135.
3. Yousef MK, Alshawwa L, Tekian A, Park YS. Challenging the arbitrary cutoff score of 60%: standard setting evidence from preclinical operative dentistry course. Med Teach. 2017;39(sup1):S75-S79.
4. Cohen-Schotanus J, van der Vleuten CP. A standard setting method with the best performing students as point of reference: practical and affordable. Med Teach. 2010;32(2):154-160.
5. Tekian A, Norcini J. Overcome the 60% passing score and improve the quality of assessment. GMS Z Med Ausbild. 2015;32(4):Doc43.
6. Yim MK, Huh S. Test equating of the medical licensing examination in 2003 and 2004 based on the item response theory. J Educ Eval Health Prof. 2006;3:2.
7. Baker FB. The basics of item response theory. Washington DC, USA: ERIC Clearinghouse on Assessment and Evaluation; 2001.
8. Downing SM. Item response theory: applications of modern test theory in medical education. Med Educ. 2003;37(8):739-745.
9. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44(1):109-117.
10. Guilera G, Gómez J. Item response theory test equating in health sciences education. Adv Health Sci Educ Theory Pract. 2008;13(1):3-10.
11. Huh S. Test equating of a medical school lecture examination based on item response theory: a case study. Korean J Med Educ. 2005;17(1):15-28.
12. Liao SW, Chang KY, Ting CK, et al. Comparison of proficiency in an anesthesiology course across distinct medical student cohorts: psychometric approaches to test equating. J Chin Med Assoc. 2014;77(3):150-154.
13. Rizopoulos D. ltm: an R package for latent variable modeling and item response theory analyses. J Stat Softw. 2006;17(5):1-25.
14. Teker GT, Kelecioğlu H, Eroğlu MG. An investigation of goodness of model data fit. Procedia Soc Behav Sci. 2013;106:394-400.
15. Stocking ML, Lord FM. Developing a common metric in item response theory. Appl Psychol Meas. 1983;7(2): 201-210.

16. Battauz M. equateIRT: an R package for IRT test equating. J Stat Softw. 2015;68(7):1-22.

17. Kim KS. What kind of physicians should we foster? Korean J Med Educ. 2011;23(1):3-5.

18. Park YS, Yang EB. Three controversies over item disclosure in medical licensure examinations. Med Educ Online. 2015;20:28821.

19. Michaelides MP. A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. Front Psychol. 2010; 1:167.