ORIGINAL ARTICLE

## Validation of Final Residency Tests Based on the Classical Model in Medical University of Mashhad (A Case Study Improving Dermatology, Ophthalmology, Obstetrics and Gynecology)

Mahmoud Ekrami[1]; Najme Moghadami[1,*]
[1] Department of Educational Science, Payame Noor University, Tehran, IRAN

Address: Department of Educational science, Payame Noor University, PO BOX 19395-3697

Fax: +982122009234
Tel: +98 9155095389
Tel: +98 9121362970
Email: n.moghadami@yahoo.com
Email: m.ekrami@pnu.ac.ir
Received: July 10, 2014
Accepted: November 15, 2014

**Background:** Evaluation of students' educational progress is one of the main goals of universities. One of the most important means to this end is the final test. In this study, the results of the final residency tests for dermatology, ophthalmology and Gynecology and Obstetrics of Medical University of Mashhad in (Tir 1391) have been examined.

**Methods:** in this applied study, the analysis of 113 answer sheets was performed through the classic model, our results which include difficulty Index and the coefficient of question determination.

**Results:** the questions at either extremes were introduced. The vague questions or the ones lacking the discriminant index as well as other incompatible questions were excluded; the rest of the questions were regarded as the basis for decision-making and ranking of the test takers.

**Conclusion:** 1) the ranking of test takers, being influenced by group changes was studied 2) All the questions of each test are compatible regarding both form and content so that all of them share a common feature 3) The questions were screened out through certain formulae and the final test questions were determined 4) The questions are chosen so as to make the test taker produce an answer rather than to pick an alternative from among others 5) Those questions which were left in the final test reflect a general scientific progress rather than a certain course belonging to that particular science

**Key words:** Educational Evaluation, Difficulty Index, discriminant index, validity index, Residency Test

## اعطاء قيمه لإختبارات طلاب التخصص فی جامعه مشهد للعلوم لطبيه علی اساس النموذج الكلاسيكی اجراء دراسه علی اختبارات قسم الجلد و العيون و النسائی و التوليد ).

**التمهيد و الهدف:** إن تقييم التطور التعليمی عند الطلاب يعتبر من اهم اهداف الجامعات . احدی اهم وسائل التقييم هو اختبار نهايه المرحله. اهتمت هذه الدراسه فی نتائج اختبار الارتقاء فی اقسام الجلد و العيون و الطب النسائی فی جامعه مشهد للعلوم الطبيه فی عام ۱۳۹۱ه .ش .

**اسلوب العمل:** فی هذه الدراسه الاستعمالية تم اجراء البحث علی ۱۱۳ ورقه اجابه اختبار ارتقاء مع استعمال نموذج الكلاسيكی ان المحاسبات كانت تشمل مستوی الصعوبه و ضريب اختيار الاسّله فی الاختبار

**النتائج:** تم تعيين الاسئله الی كانت ذو افراط او تفريط من جهه الصعوبه و السهوله. تم حذف الاسئله المبهمه و الغير المتناسقه و تم اشراك الاسئله المتبقيه فی الدراسه.

**الإستنتاج:** ۱- ترتيب المعلومات المختبره تتأثر بتغييرات المجموعه الخاضعه للدراسه. ۲- هناك خصوصيه مشتركه بين جميع اسئله الاختبار.۳- تم غربله الاسئله بمعاييردقيقه و تم تعيين الاسئله الهادفه ٤- الاسئله المتبقيه كانت تكتسب الجواب من ذهن المجيب و ليس من الاجابات المطروحه ٥- الاسئله المتبقيه كانت تعكس المعلومات العلميه الكليه عند المجيب و ليل معلومات علم معين.

**الكلمات الرئيسيه:** التقييم التعليمی. مستوی صعوبه السؤال. قدرة تشخيص السوال . مستوی صحه السئوال. اختبار التخصص.

## کلاسیک ماڈل کی اساس پر مشہد یونیورسٹی آف میڈیکل سائنسس میں ریزیڈینٹ ڈاکٹروں کے فائنل امتحان کے پرچوں کا معیار معین کرنا۔ اس تحقیق میں جلد کے امراض، امراض چشم، امراض النساء کےشعبوں کا

**بیک گراونڈ:** میڈیکل طلباء کی تعلیم میں بہتری لانا ہر میڈیکل یونیورسٹی کا ہدف ہوتا ہے۔ طلباء کی تعلیم میں پیشرفت کا جائزہ لینے کا ایک بہترین وسیلہ فائنل امتحان ہے۔ مشہد یونیورسٹی آف میڈیکل سائنسس میں جلد،آنکھوں اور خواتین کی بیماریوں کے شعبوں میں کام کرنے والے ریذیڈنٹ ڈاکٹروں کے فائنل امتحانوں کے نتائج کا جائزہ لیا گیا ہے۔

**روش:** اس تحقیق میں مذکورہ بالا شعبوں میں ایک سو تیرہ پرچوں کا کلاسیک ماڈل کے مطابق جائزہ لیا گیا۔

**نتیجے:** اس تحقیق میں ایسے سوالات جو یا تو بہت آسان تھے یا بہت مشکل، ان سوالات کو نکال دیا گیا تھا اور جو سوال بچے تھے وہ امتحان میں شامل کئے گئے تھے۔

**سفارشات:** اس تحقیق سے پتہ چلتا ہے کہ سوالات میں شرکاء تحقیق کے لحاظ سے تبدیلیاں آتي ہیں۔ ایک معین فارمولے کے ذریعے سوالات کاجائزہ لیا گیا تھا جس کے بعد ہر ٹسٹ کے لئے مناسب سوال باقی بچتے تھے۔ یہ آبجیکٹیو ٹائپ سوالات نہیں تھے بلکہ ان کے جوابات سوچ سمجھ کر لکھنا پڑتا ہے۔ سوالات ایک علمی موضوع سے متعلق ہیں۔

**کلیدی الفاظ:** سوالات،علمی جواب، جائزہ ۔

## رواسازی آزمون های نهایی دستیاری دردانشگاه علوم پزشکی مشهد بر پایه مدل کلاسیک (مطالعه موردی آزمون های ارتقای پوست،چشم پزشکی، زنان وزایمان)

**زمینه و هدف:** ارزشیابی پیشرفت تحصیلی دانشجویان یکی از اهداف عمده دانشگاه ها است. یکی از مهم ترین وسیله اندازه گیری پیشرفت تحصیلی، آزمون نهایی است. در این بررسی نتایج آزمون ارتقای دستیاری پوست، چشم پزشکی، و زنان و زایمان دانشگاه علوم پزشکی مشهد در تیرماه سال ۱۳۹۱ مورد ارزیابی قرار گرفت.

**روش کار:** در این مطالعه کاربردی، پژوهش بر روی ۱۱۳ پاسخنامه آزمون ارتقا با استفاده از مدل کلاسیک انجام شد. محاسبات، درجه دشواری وضریب تشخیص سؤال های هر آزمون را دربر می گیرد.

**یافته ها:** سؤال هایی که درجه دشواری آنها دردوحد افراط و تفریط است معرفی گردید. سؤال هایی که مبهم یافاقد قدرت تشخیص بوده و باسایر سؤال ها به هر دلیل هماهنگی نداشته است ازمجموعه های مربوط حذف ، سؤال های باقیمانده ، مبنای تصمیم گیری ودرجه بندی نهایی شرکت کنندگان در آزمون گردید.

**نتیجه گیری:** ۱- درجه بندی آزمودنی ها تحت تاثیر تغییرات گروه مورد مطالعه است، ۲- همه سؤال های هرآزمون ازلحاظ شکل ومحتوا همگون است، به قسمی که تمام آنها بایک نوع خصیصه مشترک ارتباط دارد، ۳- سؤال ها بااستفاده از فرمول های معین غربال گردیده، وسؤال های نهایی وسازنده هرتست مشخص گردید، ۴- سؤال های باقیمانده به صورتی است که آزمودنی برای پاسخ دادن به آنها، موظف به تولید پاسخ است(ونه آنکه پاسخ درست را ازمیان گزینه های معین برگزیند)، ۵- سؤال های باقیمانده منعکس کننده رشدکلی یک نوع دانش علمی است، نه یک درس بخصوص از آن دانش علمی.

**واژه های کلیدی:** ارزشیابی آموزشی، درجه دشواری سؤال، قدرت تشخیص سؤال۱، روایی سوال، آزمون دستیاری

## INTRODUCTION

Evaluation of students' educational progress is one of the main goals of universities as this can help assess the rate of students' learning and finally, the rate to which educational goals have been fulfilled (1). Assessment is one of the major aspects of the educational process to figure out the strengths and weaknesses of education and to increase the strengths, correct deficiencies and take some further positive steps in the modification and correction of the educational system (2). Educational evaluation is meant to determine quality and in achievement evaluation, quality is defined as the rate of students' achieving knowledge, skills, and abilities (4). The most essential means to his end is test (3).In general; two types of exams are used in educational evaluationto overlap. 1) Formative tests which are usually used for step-by-step assessment. The main goal is to improve learning, modify teaching methods, and correct educational deficiencies; this is usually given by the teacher or professor on one or some sections of the educational content. 2) Summative test which is usually given at the end of a course by the specialized department in order to distinguish and rank the test takers, and issue certificates (5). In this study, the results of the final residency test in Mashhad University of Medical Sciences are considered as a summative test.

In the field of medical education, educational progress is of greater significance; thus, it not only does it pursue the goals to be achieved by other common tests in other fields, but it also must investigate further details like the decision-making competence, the competence to retrieve memorized information and to use them well, and open-mindedness in dealing with the patient's problem. Another important aspect of these tests in medicine is the investigation of students' practical skills since excellence in medicine means obtaining great knowledge along with skillful usage; therefore, those tests which can assess the above-mentioned skills are the most indispensible (6).Medical universities are responsible for training skillful professionals; that makes medical education to be repeatedly investigated and correct deficiencies to improve itself (7).

The goal of this study is to investigate the validation of residency test questions

## METHODS

This research is classified as a descriptive evaluative study and an applied research.

The measuring tool, used in this research, was the students' answer sheets in the final residency test (totally 113) of Mashhad Medical University in Tir 1391, which were handed in by obtaining the university and the department of education sanction. Moreover, due to moral reasons, the names and specifications of students are not indicated; in other words, the results were merely used for implementing the study. In this study, the annual final residency test of students of Mashhad medical University in three fields, namely Dermatology, Ophthalmology, and Gynecology and Obstetrics, were taken as the sample, in which 150 multiple-choice questions, which were equally the same for all residents, but the passing score of which depended on the study level, respectively 65, 75, 85, and 95 for four successive years, were the basis of this research. The total score was 150.

The questions with as low level of difficulty as 0.1 or less were considered as very hard, and the ones with as high level of difficulty as 0.9 or more were regarded as very easy. Besides, the questions with $0<R_{pbis}<0.1$, lacked the discriminant index and the ones with $R_{pbis}<-0.1$, represented a negative correlation between the question and the whole test, which meant that weak students had answered the question right more than the strong ones; such questions lacked the discriminant index. In both of the former cases the questions were considered as invalid and required modification or removal, after which the final score is determined. The raw score before and after are not comparable as the total number of questions is lower than 150. Therefore, the achievement scores of the test takers before and after elimination of invalid questions have been reportedas a percentage.

In answering the question "to what extent can a question with two values (namely zero and one) contribute to the total score?" a two-point correlation coefficient will serve as an appropriate model. This coefficient is the index of the correlation of two variables, one of which has only two values while the others score distribution is continuous; it can be calculated from the following formula:

$$r_{Pbis} = \frac{M_p - M_t}{S_t}\sqrt{\frac{p}{q}}$$

p= the ratio of the participants who answered the question correctly

q= the ratio of the participants who answered the question incorrectly

Mp= the average number of all the participants who answered the question correctly

Mt= the average number of all the participants

St= the standard deviation of all the scores

The distribution of the scores of 113 participants in the final residency test of Dermatology, Ophthalmology and Gynecology and Obstetrics of Tir 1391, announced by the department of education reveals that:

17 participants sat the final residency test of Dermatology, including 7 men and 10 women; 41 participants sat the final residency test of Ophthalmology, including 31 men and 10 women; and finally, 55 participants sat the final residency test of Gynecology and Obstetrics, all women. Given the slight discrepancy between the final scores calculated in this study and the ones announced by the department of education, the scores obtained in this research were taken as the basis for further analysis.

## RESULTS

As mentioned above, if $p<0.1$, then the question is very hard and if $p>0.9$, then the question is very easy. If $r_{Pbis}<0.1$, the question is unable to discriminate, hence invalid, and must be modified or omitted. As a result, the

very easy, very hard, and invalid questions were specified. The final results for the residency test of these 3 fields are presented below:

9 questions from the Dermatology residency test were very easy but still valid, the level of difficulty of which was illustrated as p=0.941 and the discriminant index as $0.111<R_{pbis}<0.433$, all of which is valid. On the other hand, there was only one very hard, but valid question with the level of difficulty of 0.059 and discriminant index of 0.107. However, 41 items with the level of difficulty, ranging from 0.118 (items: 44 and 79) to 0.882 (items: 16, 97, 99, and 100), and discriminant index, ranging from -0.358 (item: 135) to 0.097 (items: 1 and 63) were regarded as invalid. 11 items with the level of difficulty, from 0.941 (items: 47, 134) to 1.000, and the discriminant index, from -0.273 (item: 134) to 0.093 (item: 47) were very easy but invalid. Totally, 52 items from Dermatology test were taken as invalid (indiscriminant), and there remained 98 perfectly valid items.

14 items from the Ophthalmology residency test were very easy but still valid, the level of difficulty of which was illustrated as $0.902 <p< 0.976$ and the discriminant index as 0.106 (item: 93) $<R_{pbis}<$ 0.424 (item:121), all of which were valid. On the other hand, there were no very hard items; consequently discriminating the much stronger participants from the rest was not possible.In addition, 28 items with the level of difficulty, ranging from 0.244 (item: 9) to 0.878 (items: 43, 70, 90), and discriminant index, ranging from -0.260 (item: 35) to 0.097 (items: 63 and 94) were regarded as invalid. 3 items with the level of difficulty of 0.976 (items: 85, 110, 142), and the discriminant index, from -0.158 (item: 110) to 0.037 (item: 142) were very easy but invalid. In general, 31invalid items from Ophthalmology test wereremoved, hence there remained 119 perfectly valid items.

11 items from the Gynecology and Obstetrics residency test were very easy but still valid, of which the level of difficulty was ranging from 0.909 (items 66 and 41) $<p<$ 0.964 (Items: 26, 37, 76) and the discriminant index was between 0.143 (item: 131) $<R_{pbis}<$ 0.434 (item:41).In contrast, there were 2 very hard but valid items (25 and 68) with the level of difficulty between 0.073 and 0.091, and the discriminant index from 0.132 to 0.369. 19questions with the level of difficulty, ranging from 0.145 (item: 137) to 0.891 (item: 118), and discriminant index, ranging from-0.260 (item: 35) to 0.097 (items: 63 and 94) were regarded as invalid. 3 items with the level of difficulty of 0.976 (items: 85, 110, 142), and the discriminant index, from -0.311 (item: 51) to 0.093 (item: 89) were invalid. Another 11 questions with the level of difficulty between 0.909 (items: 21, 101, 106, q43) and 1.000 (items: 9, 50, 56, 67) and the discriminant index between -0.064 (item: 21) and 0.094 (item: 40) were very easy but invalid; hence lacking the discriminant index. This test included a hard bbut invalid item. In general, 31 invalid items from Ophthalmology residency test were omitted, leaving 119 perfectly valid items behind.

After removing the invalid items, the final evaluation was conducted. In order for making a better comparison of scores before and after the elimination of invalid questions, the percentage of correct answers, before and after omission, are illustrated for each group in a separate table.

As can be seen in table 1, the test takers (No. 7 and 17) passed the test before removing the invalid items but failed afterwards.Although, the participant in the 12$^{th}$ row failed, both before and after the modification, and the rest of them passed under both circumstances, the ranking of top students might as well be different.

As can be seen in table 2, the test taker (No. 39) was the only one who passed the test before removing the invalid items but failed afterwards. Although, the participant in the 23$^{rd}$ row failed, both before and after the modification, and the rest of them passed under both circumstances, the ranking of top students may as well change.

As can be seen in table 3, the test takers in the 1$^{st}$, 3$^{rd}$, 8$^{th}$, 34$^{th}$, and 54$^{th}$ rows passed before removing the invalid items but failed afterwards. The ranking of top students may as well change under either circumstance, so that although the participants in the 11$^{th}$ and 36$^{th}$ row were jointly ranked as the top students (1/5) and the ones in the 18$^{th}$ and 25$^{th}$ row held the 3$^{rd}$ rank (3/5), the 25$^{th}$ test taker was announced as the only top student, followed by the 36$^{th}$ and the 11$^{th}$ as the second-best and the third.

**Table 1. The percentage of initial and ultimate success in Dermatology residency test (n=17)**

| ultimate result | percentage of ultimate result | initial result | percentage of initial result | No. | ultimate result | percentage of ultimate result | initial result | percentage of initial result | No. |
|---|---|---|---|---|---|---|---|---|---|
| pass | 76.53 | pass | 72.67 | 10 | pass | 58.16 | pass | 60.67 | 1 |
| pass | 75.51 | pass | 69.33 | 11 | pass | 74.49 | pass | 69.33 | 2 |
| fail | 43.88 | fail | 54 | 12 | pass | 61.22 | pass | 62 | 3 |
| pass | 69.39 | pass | 68 | 13 | pass | 53.06 | pass | 56.67 | 4 |
| pass | 65.31 | pass | 64 | 14 | pass | 70.41 | pass | 67.33 | 5 |
| pass | 58.16 | pass | 60.67 | 15 | pass | 88.77 | pass | 77.33 | 6 |
| pass | 58.16 | pass | 60 | 16 | fail | 36.73 | pass | 48 | 7 |
| fail | 39.80 | pass | 48 | 17 | pass | 93.88 | pass | 80.67 | 8 |
| | | | | | pass | 73.47 | pass | 71.33 | 9 |

**Table 2. The percentage of initial and ultimate success in Ophthalmology residency test (n=41)**

| ultimate result | percentage of ultimate result | initial result | percentage of initial result | No. | ultimate result | percentage of ultimate result | initial result | percentage of initial result | No. |
|---|---|---|---|---|---|---|---|---|---|
| pass | 79.832 | pass | 76.667 | 22 | Pass | 57.983 | Pass | 60.667 | 1 |
| fail | 26.891 | fail | 37.333 | 23 | Pass | 79.832 | Pass | 72.667 | 2 |
| Pass | 83.193 | Pass | 77.333 | 24 | Pass | 79.832 | Pass | 75.333 | 3 |
| Pass | 89.076 | Pass | 84.667 | 25 | Pass | 59.664 | Pass | 60.667 | 4 |
| Pass | 78.992 | Pass | 76.667 | 26 | pass | 65.546 | Pass | 66.000 | 5 |
| Pass | 66.387 | Pass | 63.333 | 27 | Pass | 76.471 | Pass | 72.667 | 6 |
| Pass | 88.235 | Pass | 84.667 | 28 | Pass | 60.504 | Pass | 62.000 | 7 |
| Pass | 72.269 | Pass | 69.333 | 29 | Pass | 66.387 | Pass | 66.667 | 8 |
| Pass | 46.218 | Pass | 50.667 | 30 | Pass | 67.227 | Pass | 66.000 | 9 |
| Pass | 57.983 | Pass | 57.333 | 31 | pass | 80.672 | Pass | 76.667 | 10 |
| Pass | 69.748 | Pass | 70.667 | 32 | Pass | 83.193 | Pass | 78.000 | 11 |
| Pass | 82.353 | Pass | 79.333 | 33 | Pass | 61.345 | pass | 62.667 | 12 |
| Pass | 59.664 | Pass | 60.667 | 34 | Pass | 46.218 | Pass | 47.333 | 13 |
| Pass | 70.588 | Pass | 70.000 | 35 | Pass | 86.555 | Pass | 80.667 | 14 |
| Pass | 58.824 | Pass | 58.667 | 36 | Pass | 84.034 | Pass | 80.000 | 15 |
| Pass | 88.235 | Pass | 83.333 | 37 | Pass | 73.950 | Pass | 71.333 | 16 |
| pass | 75.630 | Pass | 72.000 | 38 | Pass | 68.908 | Pass | 67.333 | 17 |
| fail | 38.655 | Pass | 45.333 | 39 | pass | 48.739 | Pass | 52.667 | 18 |
| Pass | 54.622 | Pass | 54.667 | 40 | Pass | 47.899 | Pass | 50.667 | 19 |
| pass | 67.227 | Pass | 64.667 | 41 | Pass | 66.387 | Pass | 66.000 | 20 |
| | | | | | pass | 60.504 | pass | 60.000 | 21 |

The column representing the percentage of the ultimate success is based on the items which truly discriminate students' achievements and progress. These summative tests include 150 multiple-choice questions, which are the same for all the residents but the minimum pass level varies due to the educational level. Simply put, these minimum scores are 65, 75, 85, and 95 for first- to fourth-level residents. For instance, the university does not let a third- to fourth-level student, who has not yet received the score of 85 reach a higher level.In the column before the last, the percentage of residents' ultimate success is illustrated (after the omission of invalid questions).

In table 1, the 7th test taker managed to pass the test when it contained 150 items, but after the removal of invalid ones and calculating the percentage of ultimate success was banned from going on to the next level. On the other hand, the percentage of success increased effectively for the 2nd, 5th, 6th, 8th, and 11thparticipants in Dermatology so that they are placed in better and more realistic professional positions.

Table 2 presents the results of the Ophthalmology residency test, in which the 4th column on the left represents the initial percentage of success in a 150-item test. The figures in the 2nd column on the left, however, illustrate the ultimate percentage of success in a 119-item test. After initial calculations, the 39th test taker managed to pass the test when it contained 150 items, but after the removal of invalid ones and calculating the percentage of ultimate success was banned from going on to the next level. In contrast, the 25th and 28th participants scored the same based on the initial percentage of success (84.667); though after the ultimate calculations, the 25th test taker received the highest rank in the group.

Table 3, representing the Gynecology and Obstetrics residency test results, depicts the changes which took place after the removal of 31 items. According to the 150-item test scores, all the residents managed to pass and go on to the next level; however, the 1st, 3rd, 8th, 34th,and 54th participants failed after the removal of invalid ones and could not get a promotion to the next level. Furthermore, the 11th and 36th participants scored equally the highest, based on the initial percentage of success (86.000); though after the ultimate calculations, the 25th test taker received the highest rank in the group (89.831).

## DISCUSSION

The purpose of this study was to further the quality of achievement tests, applied for students' evaluation in upgrade residency tests, through recent theories in analyzing the questions, such as the classical test model as well as applying technology to increase the accuracy of classical model-based calculations for residency students. The analysis was carried out to reduce the sources of errors, determine proper items, identify the questions in need of

**Table 3. The percentage of initial and ultimate success in Gynecology and Obstetrics residency test (n=55)**

| ultimate result | percentage of ultimate result | initial result | percentage of initial result | No. | ultimate result | percentage of ultimate result | initial result | percentage of initial result | No. |
|---|---|---|---|---|---|---|---|---|---|
| pass | 62.712 | Pass | 63.333 | 29 | fail | 36.441 | Pass | 44.000 | 1 |
| Pass | 71.186 | Pass | 71.333 | 30 | Pass | 81.356 | Pass | 79.333 | 2 |
| Pass | 84.746 | Pass | 82.000 | 31 | fail | 42.373 | Pass | 49.333 | 3 |
| Pass | 51.695 | Pass | 55.333 | 32 | Pass | 77.119 | Pass | 76.667 | 4 |
| Pass | 54.237 | Pass | 60.667 | 33 | Pass | 64.407 | Pass | 64.667 | 5 |
| fail | 33.898 | Pass | 44.000 | 34 | Pass | 79.661 | Pass | 78.000 | 6 |
| Pass | 50.847 | Pass | 55.333 | 35 | Pass | 54.237 | Pass | 58.000 | 7 |
| Pass | 88.983 | Pass | 86.000 | 36 | fail | 49.153 | Pass | 55.333 | 8 |
| Pass | 74.576 | Pass | 72.667 | 37 | Pass | 70.339 | Pass | 71.333 | 9 |
| Pass | 71.186 | Pass | 70.000 | 38 | Pass | 82.203 | Pass | 82.000 | 10 |
| Pass | 50.000 | Pass | 55.333 | 39 | Pass | 87.288 | Pass | 86.000 | 11 |
| Pass | 85.593 | Pass | 82.667 | 40 | Pass | 73.729 | Pass | 73.333 | 12 |
| Pass | 80.508 | Pass | 78.667 | 41 | Pass | 72.881 | Pass | 72.000 | 13 |
| Pass | 80.508 | Pass | 78.667 | 42 | Pass | 81.356 | Pass | 78.667 | 14 |
| Pass | 65.254 | Pass | 66.000 | 43 | Pass | 75.424 | Pass | 72.667 | 15 |
| Pass | 63.559 | Pass | 65.333 | 44 | Pass | 63.559 | Pass | 65.333 | 16 |
| Pass | 67.797 | Pass | 70.000 | 45 | Pass | 49.153 | Pass | 55.333 | 17 |
| Pass | 72.034 | Pass | 73.333 | 46 | Pass | 86.441 | Pass | 84.667 | 18 |
| Pass | 63.559 | Pass | 66.000 | 47 | Pass | 57.627 | Pass | 60.667 | 19 |
| Pass | 59.322 | Pass | 64.667 | 48 | Pass | 61.017 | Pass | 64.000 | 20 |
| Pass | 69.492 | Pass | 70.000 | 49 | Pass | 83.898 | Pass | 82.000 | 21 |
| Pass | 58.475 | Pass | 60.667 | 50 | Pass | 62.712 | Pass | 64.667 | 22 |
| Pass | 40.678 | Pass | 47.333 | 51 | Pass | 81.356 | Pass | 78.667 | 23 |
| Pass | 81.356 | Pass | 78.667 | 52 | Pass | 55.085 | Pass | 59.333 | 24 |
| Pass | 83.051 | Pass | 81.333 | 53 | Pass | 89.831 | Pass | 84.667 | 25 |
| fail | 42.373 | Pass | 50.000 | 54 | Pass | 72.881 | Pass | 73.333 | 26 |
| Pass | 68.644 | Pass | 69.333 | 55 | Pass | 66.102 | Pass | 67.333 | 27 |
| | | | | | Pass | 61.864 | Pass | 62.667 | 28 |

modification and discriminant index, and finally, to validate items and assess the relicompetence of the test.

This study was meant to evaluate 3 multiple-choice residency tests of Mashhad Medical University held in 1391. The evaluation included Dermatology, Ophthalmology, and Gynecology and Obstetrics residency tests of 113 participants. The results indicated that in Dermatology test, 9 items were very easy but valid, which means 6% of the items assessed the lowest educational goal (Bloom), i.e. awareness. Only one item, on the other hand was very hard but valid (item:12), capable of assessing the highest educational goal (Bloom), i.e. evaluation and judgment. However, this is not sufficient for discriminating the very strong students from the rest. Indeed it is necessary for the very hard questions ($p<0.1$) to be as many as the very easy ones. Besides, 41 items were invalid, 11 of which were very easy and not able to discriminate. For instance in item 134: 134. A person suffering from dental Periapical abscess has consequently developed headache, dysphasia, fever, and adenopathy. Which complication is not observed in this phenomenon?

a. Ludwig Angina

b. Sepsis

c. Cavernous sinus thrombosis

d.Gingivitis descamativa

the level of difficulty (0.941), that is over 94% of the test takers have picked d and answered correctly, but the discriminant index (-0.273) with a minus reveals that the weaker participants have answered this item (134) correctly more in comparison to the stronger ones; thus, this item is incapable of discriminating the strong ones from the rest and has to be reviewed or removed. The most serious licompetence to items such as 134 is that the negative verb is not noticed. In specialized sources for multiple-choice test development, avoiding them is strongly recommended.

Likewise, in the Ophthalmology test, 14 items were

considered to be very easy but valid so that around 9% of the items assessed the lowest level of educational goal, i.e. awareness. Surprisingly, there were no items for the assessment of the highest level. Moreover, 28 items were invalid, 3 of which were very easy; despite which, these items are incapable of discriminating the strong test takers from the rest. For instance, item 110:

110. In distinguishing toxoplasma retinochroiditis in HIV patients and immunocompetent ones, all the following options are true, except for :

a. The toxoplasma retinochroiditic wastes are bigger in size in HIV patients

b. The toxoplasma retinochroiditic waste pattern in HIV patients looks milliary and multifocal

c. the ocular toxoplasmosis can be accompanied by cerebral toxoplasmosis in HIV patients

d. Generally speaking, Choroidal vitreous and retin inflammatory reaction is stronger in HIV patients

The level of difficulty (0.976) indicates that over 97% of the test takers have chosen d and answered the questioncorrectly. However, the discriminant index (-0.158) with a minus reveals that the weaker participants have answered this item (110) correctly more in comparison to the stronger ones; thus, this item is incapable of discriminating the strong ones from the rest and has to be reviewed or removed from the Ophthalmology test. The most serious drawback of items such as 110 is that the word "except" is not noticed.

Also, in the Gynecology and Obstetrics test, 11 items were considered to be very easy but valid so that about 7% of the items assessed the lowest level of educational goal, i.e. awareness. On the other hand, only 2 items (25, 68) were very hard and valid to assess the highest level. Moreover, 19 items were invalid, 11 of which were very easy and only one was very hard; despite easiness, these items are incapable of discriminating the strong test takers from the rest. For instance, item 21:

21. According to College of Obstetriciansand Gynaecologists ofAmerica, the kiwi uterine systole is:

A. a kind of increase in uterine activities leading to fetal acidosis

B. the extension of a single uterine contraction for more than 2 minutes

C. the occurrence of 6 or more contractions in 10 minutes

D. the extension of a single uterine contraction for more than 1 minute

The level of difficulty (0.909) indicates that over 91% of the test takers have chosen c and answered the question correctly. However, the discriminant index (-0.064) with a minus reveals that the weaker participants have answered this item (21) correctly slightly more in comparison to the stronger ones; as a result, this item is incapable of discriminating the strong ones from the rest and has to be reviewed or removed from the test.

In the residency tests of Dermatology, Ophthalmology, and Gynecology and Obstetrics, 52, 31, and 31 items need to be removed as they are invalid; hence, it is advisable to design tests so as to assess the test takers properly and avoid any interference in their actual scores.

In a research by Macdonad &Paunonen for investigating the question parameters and competence, the Monte Carlo technique was applied. The results suggested that the anticipation of variable parameters, based on IRT, was mostly exact while the same anticipation, based on CTT, was in some cases less accurate.

In a study by Stage, results indicated that modern and classical models of a measurement are equally capable of predicting the test data systematically. Nevertheless, due to the significant resistance existing between the two, in theory and in practice, and hence, the more exact evaluations of question and competence parameters through IRT in comparison to CTT, the IRT theory is remarkably superior.

Another research by Mam sharifi 1391 was performed to investigate the psychological features of the theoretical driving test items. To analyze the test, the classical model and question-answer were used. The results indicated its one dimensionality and its local independence. Moreover, the two-parameter model is more in line with the collected data. The analysis of item parameters and the test takers revealed its simplicity and capcompetence in segregating the participants' competence, based on which the items showed a higher accuracy and coordination for less capable test takers. The estimated competence in the IRT theory was closer to the actual results than the one estimated through the classical model. Using the appropriate items for evaluating the competence of test takers can result in a question bank.

Amiriyan 1391 applied the IRT model to determine the 3 parameters, namely a,b,c as well as the achievement scores (θ) in the residency tests in Mashhad Medical University. The findings depicted that in the 1st residency test (10%), the 2nd residency test (6.06%), the 3rdinternship test (0%), the 4th residency test (1.39%), the 5thclerkship test (0%), and the 6thclerkship test (1.67%) of all the items were considered as invalid and must be removed.

Taghizadeh1390 studied 6 variables (in-service test) among the staff of education and research of Tehran County. The questionnaire included 50 multiple-choice items, the level of difficulty and discriminant index of each of which were identified. Regarding the personal manners variable, 24 items were considered as invalid and another 17 items on Quranic upbringing and its methods were not valid enough. With regards to access to information 12 invalid items, social communication 6 items, and family relations 16 items were not valid. The last variable, a survey on the political views of the West and Islam, 32 invalid items were identified and needed to be removed. The results indicated that the test on the survey on the political views of the West and Islam was invalid.

The test items parameters in this study were defined by the classical model; therefore the probability of random correct answers is not calculated. This limitation means that this study is only applicable to the sample group, and the valid/ invalid items do not separately make sense; consequently, the classical model cannot be employed for the question bank. The limitation is caused by the number of subjects in the sample group, which is not generalizeable. Provided

that the 150 items of residency tests in the above-mentioned 3 departments do not simply evaluate the level of residency, they will be content invalid.

The investigated tests are far from the standard condition. Final test evaluations should be based on discriminating questions and without extremely hard or easy items so that the evaluations can be reliable. Needless to say, using the classical model was valid and reliable enough in the exclusive case of the three studied fields. However, in order toidentify the valid and reliable items for the question bank, considering the particular twist in every item and excluding the test takers and total reliability, the IRT model should be applied.

The authors would like to thank all personnel of the student affair and academic staff affair offices of the Tehran Payame noor university (TPNU).

**Conflict of interest:** The authors declare no conflict of interest.

## REFERENCES

1. Guilbert JJ. Educational handbook for health personnel. 6th ed. Geneva: World Health Organization; 1987. 53-57.

2. Smith-Strøm H, Nortvedt MW. Evaluation of evidence-based methods used to teach nursing students to critically appraise evidence. J NursEduc 2008; 47(8): 372-5.

3. Garakyaraghi M, Avijegan M, Ebrahimi A, Esfandiari E, Esmaeili A, ShayanSh,et al. Assessment of qualitative and quantitative indexes of Clerkship Tests in general medicine. Iranian journalof medical education 2011; 10(5): 533-42. [In Persian].

4. Seif A. Measurement, assessment and evaluation of training. 4th ed. Tehran: Dowran; 2008. [In Persian]. 124-126.

5. Hooman H.A. Educational and psychological measurements.Tehran: PeykFarhang; 2011. [In Persian]. 50-55.

6. Abrayshmkar S, Sabouri M, Shayan SH, Eshraghi N, Maleki L. Analyzing and comparing the results of Objective Structured Clinical Examination (OSCE), in-group evaluation and final improvement examination of neurosurgical assistants of Isfahan University of MedicalSciences in 2009-2010. Iranian journalof medical education 2011; 10(5): 634-42. [In Persian].

6. Avizhgan M, Omid A, Dehghani M, Esmaeili A, Asilian A, Akhlaghi M,et al. Determining minimum skill achievements in advanced clinical Clerkship(externship) in school of medicine using logbooks. Iranian journalof medical education 2011; 10(5): 543-51. [In Persian].

McDonaldP,PaunonenSV. A Monte Carlo comparison of item and person statisticsbased on Item Response theory versus classical test theory. EducPsycholMeas2002; 62(6): 921-43.

7. StageC.A comparison between item analogismbased onitem response theory and classical test theory.A study of the SweSAT subset ERC,2000.

9. Taghizadeh T. Evaluation of short term course held in education and research department province of Tehran; 2011. [In Persian].

10. Mam sharifi, E. Delavar,A. Bloki, A. Shaabani, S. The evaluation of driving theory test is based on Item Response Theoryand comparison with test classical theory. 1391, 25-30

11. Amiriyan, S. determination of Triple parameters of the multiple choice tests, medical university of Mashhad on the IRT. Master thesis of the department of education and PNU Tehran. Department of educational sciences. 1391

12. Determination of the Parameters of Six Multiple Choice Tests of Mashhad University of Medical Sciences (1389-90) based on Item-Response Theory (IRT). FMEJ 3; 2 mums.ac.ir/j-fmej JUNE 21, 2013.