ORIGINAL ARTICLE

Determination of the Parameters of Six Multiple Choice Tests of Mashhad University of Medical Sciences (1389-90) based on Item-Response Theory (IRT)

Background: In this study, general and technical tests which were held in 1389 and 90 in Mashhad University of Medical Sciences were studied based on Item Response Theory (IRT) for determining the quality of multiple choice tests and the effective factors in their results.

Methods: The study was conducted on 251 answer sheets of 6 tests (2 tests of atagers,1 of internship ,and 3 tests of residency) based on IRT.

Findings: Based on IRT showed that in the first test which was related to residency questions number 1,5,and 27 had better omitted because their line slope was out of the interval of 0.5 to 205.Questions of 3 and 26 because of being bigger than the degree of freedom had meaningful qui squares and should be omitted from the test. Questions number 14 and 9 should be omitted their scores ignored because of low.

Conclusion: In a 31 item test which was related to residency 7 items were omitted because of being invalid ,therefore it is worth to plan such questions by which examinees are assessed properly and the intervention of other factors in getting their real score is prevented.

Key Words: Multiple choice Tests, Item Line Slope, Difficulty Level,Latent Trait , Item Response Theory Model

Mahmood Ekrami[1],Sahar Amirian[1],Somayeh Rajab zadeh[1]

[1] Department of Educational Science , Payame noor University, Tehran,IRAN

* Department of Educational Science , Payame noor University, PO BOX 19395-3697 Tehran,IRAN

Tel:+98-9121362970
Fax:021-22009234
Email: m.ekrami@pnu.ac.ir

تعيين معايير ستة امتحانات ذو اربعه اجوبه فى جامعه مشهد للعلوم الطبيه خلال سنوات ۹۰-۱۳۸۹ على اساس اقتراح IRT نظرية سؤال – جواب item response theory.

المقدمه: نظرا الى إن الدول اخذت طابعا صناعيا و نموالحضارات فى البقاع المختلفه يجب ايجاد آلية عمل عند المؤسسات والمصانع حين استقطاب الموظفين حتى يتم هذا الامر على احسن وجه. إن الطرق المتداولة فى اجراء مراوده كلاميه، اختبارات كتبيه بشكل اجوبه مشروحه و ايضا اسئله ذو اربعه اجوبه. من بين هذه الاساليب، المتداول هوا ختبارذواربعه اجوبه و لاجل سهوله العمل به يتم استخدامه.

هدف هذه الدراسه هو تقييم مستوى اختبارات ذو اربعه اجوبه و العوامل الموثره فيها عند طلاب الطب العام و التخصص فى جامعه مشهد الطبيه خلال عامى ۹۰-۸۹ هـ.ش على اسلوب IRT.

أسلوب العمل: فى هذه الدراسه العمليه التى تمت على عدد من الطلاب الأناث و الذكورفى قسمى الطب العام و تخصص الإسعافات الاوليه و المسالك البوليه اجريت الدراسه على ۲۵۱ استماره تتعلق بستة امتحانات امتحانين لمرحله الإستاجر، امتحان لمرحله اينترن و ۳ امتحانات لمرحله التخصص.

النتائج: إن مقارنه مستوى الصعوبه و قدرة التشخيص و انحدارالخط فى كل سؤال على أساس IRTكانت فى اختبارالتخصص على الشكل التالى: الأسئله ۲۷ ، ۵ ، ۱ كان انحدار الخط بهم خارج نطاق ۰.۵ ، ۲.۵ لذا كان من الافضل أن يتم حذفهم او لم تحسب علامتهم، اسئله ۳ ، ۲۶ كان اكبر بثلاثه اضعاف من درجه الحريه فقط كان مجذور كاى ذو معنى لديهم يجب حذفهم من الاسئله، اسئله ۱۴ ، ۹ لأجل القدرة التشخيصيه المنخفظه يجب حذفهم و تجاهل علامتهم.

الإستنتاج: بعد تعيين النتائج الثلاثه، الإنحدار a_g، الصعوبه b_gو التوفيق على مستوى الخط c_g، و لكل الـ ۲۸٤ ماده، و حذف المواد الغيرمؤيده و تم تعيين مستوى المقدره و على هذا الاساس تم تعيين المواد المؤيده و المعتبره، نرى من خلال متابعه ۲۵۱ سؤال على حسب اركان انحدار الخط و صعوبه السؤال و الحدس فى ستة اختبارات، ۷ اسئله خارج محدودة ۰.۵ تا ۲.۵ ، ۱۰ سؤال اكبر من ثلاث درجات للحريه و لهم مجذور كاى ذو قيمه، و اسئله ذو قدر تشخيصيه متدنيه بين ۰.۱ -۰.۲ و ۱۹ اسئله ذوحدس اعلى من ۰.۲۵ فى فحص ذو ۳۱ التى تم اجراوه لمرحله التخصص علينا حذف ۷ مواد لعدم كفائتهم لذا يجب ان نعمل على وضع الامتحانات بشكل ادق لمعرفه مستوى الطالب.

الكلمات الرئيسيه: طلاب مرحله التخصص فى الطب، الرؤيه، الاداء.

مشہد یونیورسٹی آف میڈیکل سائنسس میں آئي آرٹی ماڈل پرمبنی آبجیکٹیو سوالات سے چھے امتحانوں کے معیارات کا تعین

بیک گراونڈ: دنیا میں صنعتی ترقی اور ملکوں کی پیشرفت کے پیش نظر ضرورت اس بات کی ہے کہ صنعتوں اور اداروں میں ایسے سوالی پرچے بنانے کی ضرورت ہے جو بہترین افراد کے انتخابات میں مفید واقع ہوں۔ انٹرویو، کتبی امتحانات، ٹسٹ معمول کی روشیں ہیں جن سے عام طور سے استفادہ کیا جاتا ہے۔ ان روشوں میں چار سوالات پرمبنی آبجیکٹیو طریقہ جو کہ طلباء کے امتحانی نتائج میں بھی مفید واقع ہوتا ہے انتخاب کیا گیا اور دوہزار دس اور گیارہ میں آئیٹم رسپانس تھیوری کی روش سے اس کا تجزیہ کیا گیا۔ اس تحقیق میں جنرل اور اسپیشیالائزیشن لیول کے طلباء شامل تھے۔

روش: اس تحقیق میں جیسا کہ بتایا گیا جنرل میڈیسنس اور اسپیشیالائزیشن کے طلبا وطالبات نے شرکت کی۔ اسپیشیالائزیشن کے طلباء کا تعلق ایمرجنسی میڈیسن اور یورولوجی سے تھا۔ چونکہ تعلیمی تجزیوں میں نمونوں پر اکتفا نہیں کی جاسکتی لھذا جوابات کا مکمل طرح سے جائزہ لیا گیا۔

نتیجے : آئي آر ٹی کے طریقے سے تینوں پیرامیٹر کی تشخیص کے بعد جو کہ اے جی، بی جی اور سی جی سے عبارت ہیں یہ معلوم ہوتا ہے کہ سات سوالات صفر اعشاریہ پانچ سے دواعشاریہ پانچ کی حد سے خارج تھے اسی طرح دس سوالات نہایت وسیع تھے اور نو سوالات کی تشخیص دینا طلباء کے لئے مشکل تھا لھذا ان سوالات کو نکال دیا جانا چاہیے۔

سفارش : متعد سوالوں کے مناسب نہ ہونے کی وجہ سےیہ سفارش کی جاتی ہے کہ کوئي ایسا طریقہ اپنایا جائے جس سے طلباء کی صلاحیت میں بھی اضافہ ہوا اور وہ بہتر طریقے سے سوالات کا جواب دے سکیں۔

کلیدی الفاظ: آئي آرٹی، آبجیکٹیو ، سوالات ۔ توانائي ، صلاحیت ۔

تعیین پارامترهای شش آزمون چهار گزینه ای دانشگاه علوم پزشکی مشهد در سالهای ۹۰-۱۳۸۹ بر پایه مدل تئوری سوال-پاسخ

مقدمه: مصاحبه، آزمونهای کتبی تشریحی،...و آزمون های تستی برای بکار گیری افراد در صنایع و سازمانها روشهای معمول می باشند. از بین این روشها، آزمونهای تستی چهار گزینه ای بدلیل سهولت برآورد نتایج، ساده ترین روش می باشد. در این تحقیق، جهت برآورد کیفیت آزمون های چهارگزینه ای و عوامل تاثیر گذار در نتایج آنها، آزمونهایی که برای دانشجویان دانشگاه علوم پزشکی مشهد دردو سطح عمومی و تخصصی طی سالهای ۸۹ و۹۰برگزار شده است، از طریق مدل item response theory(IRT) مورد بررسی و مطالعه قرار گرفته است.

روش کار: در این مطالعه کاربردی، جامعه آماری شامل دانشجویان خانم و آقای مشغول به تحصیل در رشته پزشکی در سطح عمومی و تخصص در رشته های طب اورژانس و اورولوژی دانشگاه علوم پزشکی بود. از آن جایی که در ارزشیابی آموزشی نمونه برداری جایز نیست لذا کلیه اطلاعات در دسترس مورد قضاوت قرار گرفت . بنا براین پژوهش برروی ۲۵۱ پاسخنامه مربوط به ۶ آزمون (۲ آزمون کارآموزی، یک آزمون کارورزی و سه آزمون دستیاری) با استفاده از مدلIRT انجام شد

یافته ها: بر اساس مدلIRT در آزمون یکم که مربوط به گروه دستیاری بود سؤالهای ۱، ۵ و۲۷ به دلیل اینکه شیب خم آنها خارج از بازه ۰/۵ تا ۲/۵ بود بهتر است حذف و یا نمره های آنها نادیده گرفته شود سؤالهای ۳ و۲۶ به دلیل آنکه از سه برابر درجه آزادی بزرگتر است دارای مجذور کای معنا دار است و باید از مجموعه سؤالات حذف گردد. سؤال های ۱۴و ۹ به دلیل قدرت تشخیص پایین بین باید از مجموعه حذف و یا نمره های آنها نادیده گرفته شود.

نتیجه گیری: در یک آزمون ۳۱ سوالی که مربوط به دوره دستیاری است تعداد ۷ ماده آزمون به دلیل روا نبودن آزمون باید حذف گردند لذا شایسته است که در تهیه آزمونها به روشی عمل شود تا آزمون شوندگان به درستی سنجش شده و از دخالت سایر عوامل در کسب نمره واقعی آنان جلوگیری شود.

INTRODUCTION

One of the important goals of education is the determination of capabilities, aptitudes, and limitations of members of the society .So that chances of growth and improvement are prepared by social,economical ,and professional leadership. In this way accurate tools and methods of testing and assessment are necessary and no educational system can reach its goals without applying them(1).

Different kinds of tests are used in a wide variety in schools ,psychotherapy centers ,industry, army, and governmental organizations for giving advice, identification of mental issues, making a choice, leadership and job selection. psycho -educational tests are generally used in assessing individual differences in the decision making process .Test, measure individual differences from the point of talents and personality traits(2).

Because of great improvements in the case of individual identification in behavioral science, different theories have been invented in the field of tests and testing.Normally psychotherapists use mathematic models for planning the items and analyzing the scores in order to express their theories.A mathematic model includes some theories about the data which explains the specific relation among observable constructions and unobservable constructions of the model(3).

If we consider a multiple choice test as a tool for assessing a kind of trait or specification of a person ,the most basic question about the test will be that what traits or specifications the mentioned test measures?and how well does it measure that trait or specification?(4)

The Classic Model

The Spearman model is based on this theory that tells us that each score can be considered as a combination of two additive components,which are the true score and random error score .In other words, by taking a test of a person or a group of people ,some scores are achieved which are although favored by the tester to consider them as the real measurement of the trait or capability but because of some factors this score does not express the real amount of that trait or the level of the person b itself and it is just an observed score and nothing much.(5)

In the classic theory of the test the difference in standard variables such as item difficulty (the ratio of correct answers) cannot be used for the evaluation of these side-takings.Sometimes the criteria of determinationof the item (the difference of the relation of correct answers of criteria groupsgroups to test items) can not delete these side takings.These criteria can not find the side taking of in group difference in a specification which is measured by the test. (6)

Accessibility to computer which started in 1960s ,provided the invention of the testing theory oflatent traitand adaptive assessment with computer in late 1970s and early 1980. Latent traitis an unobservable trait which determines a specific collection of stability and coordination among individuals along with the differences between them all at once.(7)As computer was used in analyzing the data of psychotherapy,the theory of which has been invented previously but could not be used was used.

The practicality of latent trait theory caused a lot of changes in the performance of psychotherapy tests which affected all the job which was done in 60 years in classic theory (8).

Psychotherapists and testing experts turned in to such theories with more interest and gradually scientific texts and computer software spread among them for the purpose of psychotherapy. It is not the matter of advantages or disadvantages of new or classic theories anymore and studies are conducted in the case of choosing the appropriate model, selection of faster and more accurate methods for calculating the parameters of the models, performing stronger tests for the determination of the appropriateness of data models.(6)

The new theories of psychotherapy first was strongly connected to latent trait

In a way that can be seen a lot in the review of literature and background of new theories but nowadays IRT with the theory of specific slope has become more popular and seems more appropriate for testing and data analysis.(9)

Item-Response Theory and Specific Slope of the Item

The Item-Response theory is usually shown by IRT.(10)

Tests in which the items are harmonious from the point of content, it is logical to consider a united dimension of trait which is the fundamental function in all items of the test. This trait dimension which does not necessarily have to be psychologically simple, is statistically considered a united structure which acts as a determining factor of success in all items of the test and based on the imagination of the trait dimensionwhich is a fundamental specification with which the test is measured. ICCorspecific slope can be examined.(7)

Specific slope of the item is a function which relates success probability of the triable place the item to the testable position in the measured fundamental trait. (Torkashvand)

Two terms ofspecific slopeand IRT are used interchangeably a lot.

Specific slope of the item is a slope which shows the possibility of giving a correct answer to an item(p_g) as a function of different trait levels(θ) which leads to success in answering the item.IRT considers both the role of the items of the test and answers to them.(5)

METHODS

In this applied study ,the participants included male and female students of general and specialized medicine. Specialty majors included emergency medicine and euro logy in Mashhad University of Medical Sciences. As sampling is not allowed in educational assessment,therefore all available data was measured.

In this study, according to the sensitivity of the university in some cases and Lack of archives of the multiple choice questions of the previous years, just the questions of two fields of emergency medicine and euro logy of 2010 and 2011 were accessible and it was not possible to get access to other questions.Thus the questionnaires, answer keys, and answer sheets were received and the study was

done on the residents of the two mentioned specialties. Finally this study was conducted on 251 answer sheets collected from taking 6 tests (2 tests of stagership,1 test of internship, and 3 tests of residency) by using IRT model.

For each file a separate table was planned in which the first row included the number of items,the second row the answer key ,and the following rows related to the answers students had given.Data of each table was analyzed.

For data analysis IRT model was used as follows.

1_What is slope line of each item?

2_Ehat is the item difficulty?

3_What is the probability of choosing the correct answer by chance?

4_How much is the capability (θ) of students?

5_What are the valid items in each test?

RESULTS

The study was conducted on 251 answer sheets related to 6 tests of Mashhad University of Medical Sciences based on IRT model.The results are as following.

Before data analysis ,it is necessary to show the type of the test, number of examinees, and number of items of each test in table no.1,then the tertiary parameters of each test along with students' capabilities (θ) are determined,and finally the progress score of the students based on the standard score ,average of 80,and standard deviation of 10 are given. Success in an individual item ,even more than success in the whole test, depends on the examinee's status of latent trait ,systematic, and various random factors. For a multiple choice item to be acceptable ,it is necessary that the success chance of the examinee ,in the mentioned item increases continuously along with the promotion of his rank in latent trait.

A function which shows the success probability in the item is called Latent trait ,or awareness function ,or item specific slope .(11)

This can be described by three parameters:

1_The parameter which shows the bend slope ,it means when the relation of change in success probability moves toward above latent trait. This parameter is shown through the sign of a ,and its usual amount is from 0.5 to 2.5.(12)

2_The parameter that shows specific slope of the item is placed where in

latent trait ,which is the parameter which expresses the item difficulty and is shown by b. Its usual amount is from _2.5 to 2.5,item difficulty for amounts less than -2.5 is considered very easy, and for amounts more than 2.5 very difficult.(12)

3_The parameter which shows the base line curve for very low levels of the, it means that the level of success by chance for multiple choice questions is expressed with C and in multiple choice items its usual amount is 0.25.(12)

The tertiary parameters of the first residency test were calculated and their amount are shown in table 2 along with the number of examinees, amount of qui square, and degree of freedom.

As it was mentioned the amounts of a in table 2 are determined from 0.5 to 2.5 and amounts other than this interval have been determined as invalid items by the computer and shown by Item Deleted . Therefore items 1,5,and 27 which are mentioned below are invalid and deleted from the rest of the calculations.

On the other hand,experimentally qui square (the second column on the right) need to be less than three times of degree of freedom (the first column on the right),otherwise the amount of qui square is meaningful and the reliability of the item gets in risk.(11)The amount of qui square in items number three and twenty six are respectively 72.057 and 61.439are larger than the three times of degree of freedom (48) and are meaningful ,thus items three and twenty six are invalid and need to be deleted.

In the first residency test ,parameter a is in the interval of 0.618(item 2) to 1.817(item 13) .It is normal that the more the amount of a,the steeper the bend slope is ,which means that in item 13 with increase in the level of latent trait ,the probability of success in the question increases rapidly .In other words this item just relies on latent trait and nothing else much, in a way that a person with enough latent trait is sure about his success and a person with not enough latent trait is almost sure about his failure.In contrast ,the item slope of item 2 is almost flat (a=0.618) and in the case of item 2 with increase in the level of latent trait ,the probability of success slowly increases.

Table 1. Data of conducting 6 tests on 251 students of Mashhad University of Medical Sciences

Test name	Number of examinees	Number of females	Number of males	Number of questions
1_Residency test	19	6	13	30
2_Residency test	10	3	7	34
3_Internship test	24	14	10	28
4_Residency test	33	30	3	72
5_Stagership test	91	51	40	60
6_Stagership test	74	45	29	60
Total	251	149	102	284

Table 2. The Determination of Parameters of the First Residency Test(n=19)

Item	a	b	c	N	χ^2	df
1	------Item Deleted------					
2	0.618	0.342	0.290	19	30.116	16
3	0.832	-2.294	0.260	19	61.439	16
4	1.574	2.008	0.290	19	20.289	16
5	------Item Deleted------					
6	1.466	1.762	0.310	19	20.573	16
7	1.503	3.000	0.330	19	21.816	16
8	1.494	3.000	0.280	19	20.592	16
9	1.712	0.661	0.190	19	13.237	16
10	1.635	0.562	0.310	19	18.025	16
11	0.958	0.006	0.250	19	22.833	16
12	1.605	2.005	0.240	19	17.850	16
13	1.817	0.192	0.160	19	8.374	16
14	1.743	0.729	0.170	19	11.047	16
15	1.460	3.000	0.310	19	20.984	16
16	1.186	-0.133	0.270	19	22.739	16
17	1.528	3.000	0.200	19	16.166	16
18	1.524	3.000	0.230	19	18.037	16
19	1.431	-0.577	0.260	19	12.913	16
20	1.558	-1.006	0.230	19	8.311	16
21	1.362	-0.064	0.280	19	18.223	16
22	1.148	-0.188	0.210	19	37.083	16
23	1.121	0.695	0.310	19	21.203	16
24	0.954	-1.080	0.260	19	27.211	16
25	1.154	-1.532	0.250	19	37.834	16
26	0.793	-0.660	0.280	19	72.057	16
27	------Item Deleted------					
28	0.942	0.064	0.260	19	21.912	16
29	1.000	0.359	0.270	19	21.106	16
30	1.448	0.484	0.320	19	18.977	16

In other words success in replying item 2 depends on factors other than latent trait to a great extent and the level of havinglatent trait hardly affects success in replying item 2. Experimentally amounts less than 0.5 for a ,causes the item becomes inappropriate. In the first test ofresidency item difficulty or the amounts of b varies from -2.294 (item 3) to 3.0 (items 7,8,15,17,18).The amounts b<-2.5 show that item difficulty is very low and b>2.5 shows that item difficulty is very high or the item is very difficult. Therefore just item 3 is very easy, items 25(b=-1.532), 24(b=-1.080), 20(b=-1.006), 26(b=-0.660), 19(b=-0.577), 22 (b=-0.188), 16 (b=-0.133), and 21 (b=-0.064) are the easiest questions respectively ,and on the other hand , items 7, 8, 15, 17, 18 (b=3.0) arevery difficult. Items 4(b=2.008). 12(b=2.005), 6(b=1.762) 14(b=0.729), 23(b=0.695), 9(b=0.661), 10(b=0.562), 30(b=0.484), 29(b=0.359), 2

(b=0.342), 13 (b=0.192), 28 (b=0.064), and 11 (b=0.006) were difficult to easy items respectively. In the first test of residency ,parameter C expresses the probability of selecting the correct choice by chance.The amounts of C lie between 0.160(item 13) and 0.330(item 7).In this test ,C=0.25 shows that selecting the right choice among the others is not possible through chance, larger amounts like C=0.330 in item 7 shows that selecting the right choice is possible by little chance and is related less to.

On the other hand, C<0.25 can put the item among difficult and often invalid items.

The scale of ability (θ) of participants of the first test of residency based on standard score (mean of 0 and standard deviation of 1) and also based on progress score (mean of 80 and standard deviation of 10) is shown in table 3 .

As the data of table 3 shows the standard score of ability (θ) of the examinees lie between -1.59 to 1.47,minus scores express being less than the mean and positive scores show latent trait in more than average in residents.Corresponding with θ the progress score of students lie in the interval of 64.10 to 94.65.

In the second test of residency the amounts of a in items 5 and 33 are determined as invalid because of standing out of the interval and are deleted.

Qui square in item 16 was 21.607 which was bigger than three times of degree of freedom (21) and is meaningful,therefore item 16 is invalid and needs to be deleted.

In the second test of residency item difficulty or b lies between _1.608 (item 32) to 3.0(items 6,17,18,23,28). b<_2.5 shows that item difficulty is very low or in other words the item is very easy,in contrast the amounts of b>2.4 express that item difficulty is very high or the item is very difficult .Therefore item 32 is the easiest and items 6,17,18,23,28 (b=3.0) are very difficult .In the second test of residency ,parameter c of item 1,26,and 31 show that choosing the right answer among others is possible with a little chance and by guessing and is less related to latent trait.

The standard ability scores of the participants lie between _1.45 to 1.45,the negative scores are less than the average ,and the positive scores are more than the average. The progressscores of the participants lie between 65.48 to 95.36.

In the third test if internship as the allowed amounts for are between 0.5 to 2.5 and according to the fact that all the numbers in column a placed in this interval ,therefore all the items of the tests are valid and none are deleted .Qui square in items 17 and 19 are 84.245 and 77.347 respectively and are bigger than the three times ofdegree of freedomand meaningful, thus items 17 and 19 are invalid and must be deleted.

In internship test, item difficulty or the amounts of b lie between _2.552 (item 17) to 3.0(items 3, 4, 5, 8, 9, 11, 18, 22, 24, 25, 26). b<_2.5 shows that item difficulty is very low and in other words the item is very easy and in contrast b>2.5 shows that the item is very difficult or high item difficulty. Therefore just item 17 is very easy and items 13 (b=_1.558), 23(b=_1.399), 12 (b=_1.227), 19 (b=_1.011), 14 (b=_0.849), 15 (b=_0.778), 6(b=_0.755), 21 (b=_0.655), 10 (b=_0.500),27 (b=_0.434), 20 (b=_0.092), and 1 (b=_0.077) are the easiest items respectively .On the other hand, items 3, 4, 5, 8, 9, 11, 18, 22, 24, 25, 26 (3.0) are very difficult and items 2 (b=1.916), 28 (b=1.786), 7 (b=1.202) and (b=0.750) are difficult to easy items respectively.

In internship test ,parameter c for the items 2 and 22 shows that the right choice can be guessed and is little related to latent trait.

The standard score of ability (θ) of the participants is between _1.76 to 1.33, the negative scores are less than the average and the positive ones are higher than the average .The progressscores of the participants lie between 62.38 to 93.27.

In the fourth test of residency the amount of a in item 72 is

Table 3. Gender and Ability of Examinees in the First Test of Residency based on θ and progress scale (n=19)

NO	sex	θ	X1	NO	sex	θ	X1
01	male	-1.59	64.10	11	male	.96	89.61
02	female	1.47	94.65	12	male	-.91	70.95
03	male	-.15	78.53	13	male	.19	81.90
04	male	-.53	74.72	14	female	.13	81.26
05	male	-.20	78.00	15	female	.22	82.17
06	female	-.62	73.77	16	female	.14	81.44
07	male	.36	83.63	17	male	.30	83.03
08	male	.30	82.95	18	male	-.26	77.37
09		.64	86.43	19	male	.55	85.49
10		-.05	79.55				

out of the interval ,therefore it is invalid and must be deleted.

The amounts of qui square in item 21 and 59 are 110.682 and 58.971 respectively which are bigger than the three times of degree of freedom and are meaningful .Thus items 21 and 59 are also invalid and need to be deleted.

In the fourth test of residency item difficulty (b) ranges from _2.388 (item 48) to 3.0 (items 12, 4,1, 18, 20, 23, 27, 29, 30, 31, 37, 40, 41, 43,47, 51, 55, 56, 57, 60, 61, 65, 66, 67, 68, 69, 70, 71).b< _2.5 shows high item difficulty or a very hard question. In this way just item 48 is easy.

In the fourth test of residency ,parameter c introduces the probability of choosing the right answer by chance .Its amounts range from 0.130 (items 43 and 59) to 0.410 (item 27).

The standard scores of ability (θ) of the participants range from _1.49 to 1.39, negative scores are less than the average and positive ones are higher than the average. Progress scores of students vary from 65.11 to 93.94.

In the fifth test of internship as the amounts of a lie between 0.5 to 2.5 therefore all the item are valid and none are deleted .Qui square of item 36 is 55.185 which is larger than the three times of degree of freedom (51) and so is meaningful, thus item 36 is invalid and must be deleted.

In the fifth test of internship item difficulty (b) lie between _2.920 (item 36) to 3.0 (items 28, 43, 46, 47, 48, 49, 51, 53, 54, 57, 58, 60). b< _2.5 shows that item difficulty is very low and in other words the item is very easy and need to be deleted ,on the other hand b>2.5 expresses high item difficulty or a very difficult item.

Therefore just item 36 is very easy and items 28, 43, 46, 47, 48, 49, 51, 53, 54, 57, 58, 60 are very difficult.

In the fifth test of internship ,the amount of c range from 0.020 (item 49) to 0.440(item 45).

θ lies between _1.36 to 1.70,negative scores are higher than average and positive scores are higher than average. Progress scores of students are placed between 66.40 to 97.04.

In the sixth test of internship as it is mentioned in table 12 the amounts of a lie between 0.5 to 2.5 and any amount except this interval is considered as invalid by the computer and marked as Item Deleted. Thus item 30 is invalid and is deleted.

Qui squares of items 7,31,35 are 85.178,53.120,54.635 respectively and are bigger than the three times of degree of freedom and are meaningful, therefore items 7,31,and 35 are invalid and must be deleted .In the sixth test of internship item difficulty or b ranges from _3.0 (item 35) to 3.0 (items 48,52,56,60) ._2.5 shows that item difficulty is very low and in other words the item is very easy ,on the other hand the amounts of b<2.5 show that item difficulty is very high or the item is very difficult. Therefore just item 35 is very easy and in contrast items 48,52,56,60 (b=3.0) are very difficult.

In the sixth testof internship,parameter c shows the probability of selecting the right choice by chance. Itsamounts vary between 0.150 (item 48)to 0.410 (item 21) ,which shows that in item 21 selecting the correct choice is possible by guessing and is less related to latent trait.The standard ability score of the participants lie between _1.61 to 1.71,the negative scores are lower than the average andpositive scores are higher than average .Progress scores scores of the students lie between 46.94 to 90.46.

DISCUSSION

The main purpose of this study was torefinethe items and determination valid and invalid items and recognition of parameters of each item .The advantage of this study over the classic one is that it is free of testing and in other words it does not have a dependent test and each item is analyzed independently.Therefore the organization which need them can supply their resources and provide tests of the highest standard and get real results.

Zolfaghari (2007)conducted a study on 457 participants which were selectedrandomlyfrom the first ,second, and third level of secondary school of Ferdows .For studying the parameters of the test factor analysis and tertiary parameter model were used which was determined thatthis test has more than one dimension.For the determination of the properness of this test with the examineesawareness functionand specific slope functionof the item showed that this test is proper and along with the ability of the examinees.As a result the first hypothesis was rejected and the second one was accepted.

The results of holding the test show that the comparison of item difficulty, judgementcapability,and slope lineof each item expressed that in file 1 which related to residents items 1,5,and 27 had better be omitted becausetheir line slopeswas out of the interval of0.5 to 2.5.Items 3 and 26 must also be deleted as they have a meaningful qui square .Items 9 and 14 should also be deleted because oflow.

In a 31 item test which related to the residency level ,7 items had to be deletedbecause of not being valid thus it is worth paying the essential attention in planning tests so that the examinees are assessed appropriately and other factors do not interfere.

The studied tests are far from standard conditions .For planning tests more attention must be paid so that their results are reliable.

REFERENCES

1. Ebrahimifar T. Studying the second scale 2 of cattle's intelligence test that is against the culture by using the classical method and IRT among the girls and boys of Tehran province. MS. Dissertation. Islamic Azad University of Tehran, 1999. [In Persian].

2. Pashasharifi H. The fundamentals of psychometry and the assessment of psychology. Tehran: NourHekmat; 1995.

3. Thorndike R. Applied psychometrics. Tehran: Tehran University; 1996.

4. Joker B. Assessment of the scale 2 of the intelligence test that is according to the Cattle culture. MS. Dissertation. Shiraz University; 1993. [In Persian].

5. Zolfaqhari QH. Discussing about the dimensions of the content of Riven test and assessing its parameters according to the question and answer hypothesis. MS. Dissertation. Islamic Azad University of Tehran; 2007. [In Persian].

6. Sepasi H. The classical test view and its limitation. The journal of educational

sciences of Chamran University 1995; 2:133 [Persian].
7. Abbassi F. Studying the second scale 2 of cattle's intelligence test that is against the culture by using the classical method and IRT among the girls and boys of Tehran province. MS. Dissertation. Islamic Azad University of Tehran, 1999. [Persian].
8. Alimohamadi F. Using the IRT model in balancing the test of self-respect of Izneck And Kooper for the girl students in the guidance school and the first grade of high school in Tehran according to the three oarameter IRT model. Islamic Azad University of Tehran, 1997.[Persian].
9. Hambleton RK. Item response theory: Principle and application. Boston: Kluwer-Nijhoff; 1985.
10. Human H. Statistical presumption in behavioral research. Tehran: PeikFarhang; 1994. [In Persian].
11. Human H. Psychological and educational assessing and the way of providing test. Tehran: Parsa; 1994.[Persian].
12. HumanH. Comparing the Rash one parameter method with a two parameter method with the Tehran-Bines data for assessing the intelligence of the students from11 to 15. MS. Dissertation. Allame Tabatabaee University of Tehran,1994.