تأمین مولفیه نتایج امتحانات در اجرای امرهای تصمیم‌گیری IRT

مقدمه: این تحقیق سلسله‌ی چهارگانه‌ی آزمون‌های گروهی علوم پزشکی در دانشگاه علوم پزشکی مشهد در مدت سال‌های 1389 – 1390 بر پایه‌ی اصول اصولات IRT انجام گرفت.

پایه‌های ژاکویر و اسکندرزی از مدل IRT به‌عنوان مدل تصمیم‌گیری کیفیتی در آزمون‌های گروهی قابل استفاده‌ی نموداری می‌دانند و گزارش با نتایجی مثبت که در علوم پزشکی و آموزش پزشکی بدست آمده‌اند.

روشن کار: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.

روش: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.

نتایج: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.

مقدمه: در این تحقیق، سلسله‌ی چهارگانه‌ی آزمون‌های گروهی علوم پزشکی در دانشگاه علوم پزشکی مشهد در مدت سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT انجام گرفت.

پایه‌های ژاکویر و اسکندرزی از مدل IRT به‌عنوان مدل تصمیم‌گیری کیفیتی در آزمون‌های گروهی قابل استفاده‌ی نموداری می‌دانند و گزارش با نتایجی مثبت که در علوم پزشکی و آموزش پزشکی بدست آمده‌اند.

روشن کار: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.

نتایج: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.

مقدمه: در این تحقیق، سلسله‌ی چهارگانه‌ی آزمون‌های گروهی علوم پزشکی در دانشگاه علوم پزشکی مشهد در مدت سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT انجام گرفت.

پایه‌های ژاکویر و اسکندرزی از مدل IRT به‌عنوان مدل تصمیم‌گیری کیفیتی در آزمون‌های گروهی قابل استفاده‌ی نموداری می‌دانند و گزارش با نتایجی مثبت که در علوم پزشکی و آموزش پزشکی بدست آمده‌اند.

روشن کار: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.

نتایج: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.

مقدمه: در این تحقیق، سلسله‌ی چهارگانه‌ی آزمون‌های گروهی علوم پزشکی در دانشگاه علوم پزشکی مشهد در مدت سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT انجام گرفت.

پایه‌های ژاکویر و اسکندرزی از مدل IRT به‌عنوان مدل تصمیم‌گیری کیفیتی در آزمون‌های گروهی قابل استفاده‌ی نموداری می‌دانند و گزارش با نتایجی مثبت که در علوم پزشکی و آموزش پزشکی بدست آمده‌اند.

روشن کار: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.

نتایج: این مطالعه کاملاً پژوهشی چهارگانه‌ی جامعه‌ای شامل شاخصات نمونه‌ای مرتبط به دانشکده علوم پزشکی مشهد در سال‌های 1389 – 1390 بر پایه‌ی اصولات IRT نیست.
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 because of some factors this score does not express the real amount of that trait or the level of the person 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 determination of the item (the difference of the relation of correct answers of criteria group-item 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 of latent trait and adaptive assessment with computer in late 1970s and early 1980. Latent traits 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). Psychologists 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 dimension which is a fundamental specification with which the test is measured. ICCs specific 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 of specific slope and 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 \( p_i \) as a function of different trait levels (8) 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 endocrinology 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 endocrinology in 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
determination of six multiple choice tests based on IRT

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 stagiership, 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 the slope line of each item?
2. What is the item difficulty?
3. What is the probability of choosing the correct answer by chance?
4. How much is the capability \( \theta \) 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 1.

<table>
<thead>
<tr>
<th>Test name</th>
<th>Number of examinees</th>
<th>Number of females</th>
<th>Number of males</th>
<th>Number of questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_Residency test</td>
<td>19</td>
<td>6</td>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td>2_Residency test</td>
<td>10</td>
<td>3</td>
<td>7</td>
<td>34</td>
</tr>
<tr>
<td>3_Internship test</td>
<td>24</td>
<td>14</td>
<td>10</td>
<td>28</td>
</tr>
<tr>
<td>4_Residency test</td>
<td>33</td>
<td>30</td>
<td>3</td>
<td>72</td>
</tr>
<tr>
<td>5_Stagiership test</td>
<td>91</td>
<td>51</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>6_Stagiership test</td>
<td>74</td>
<td>45</td>
<td>29</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td>251</td>
<td>149</td>
<td>102</td>
<td>284</td>
</tr>
</tbody>
</table>

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. The amount of qui square in items number three and twenty six are respectively 72.057 and 61.439 are 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.
In other words success in replying item 2 depends on factors other than latent trait to a great extent and the level of having latent 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 of residency 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 are very 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

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

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>N</th>
<th>$\chi^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>----</td>
</tr>
<tr>
<td>2</td>
<td>0.618</td>
<td>0.342</td>
<td>0.290</td>
<td>19</td>
<td>30.116</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>0.832</td>
<td>-2.294</td>
<td>0.260</td>
<td>19</td>
<td>61.439</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>1.574</td>
<td>2.008</td>
<td>0.290</td>
<td>19</td>
<td>20.289</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>----</td>
</tr>
<tr>
<td>6</td>
<td>1.466</td>
<td>1.762</td>
<td>0.310</td>
<td>19</td>
<td>20.573</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>1.503</td>
<td>3.000</td>
<td>0.330</td>
<td>19</td>
<td>21.816</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>1.494</td>
<td>3.000</td>
<td>0.280</td>
<td>19</td>
<td>20.592</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>1.712</td>
<td>0.661</td>
<td>0.190</td>
<td>19</td>
<td>13.237</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>1.635</td>
<td>0.562</td>
<td>0.310</td>
<td>19</td>
<td>18.025</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>0.958</td>
<td>0.006</td>
<td>0.250</td>
<td>19</td>
<td>22.833</td>
<td>16</td>
</tr>
<tr>
<td>12</td>
<td>1.605</td>
<td>2.005</td>
<td>0.240</td>
<td>19</td>
<td>17.850</td>
<td>16</td>
</tr>
<tr>
<td>13</td>
<td>1.817</td>
<td>0.192</td>
<td>0.160</td>
<td>19</td>
<td>8.374</td>
<td>16</td>
</tr>
<tr>
<td>14</td>
<td>1.743</td>
<td>0.729</td>
<td>0.170</td>
<td>19</td>
<td>11.047</td>
<td>16</td>
</tr>
<tr>
<td>15</td>
<td>1.460</td>
<td>3.000</td>
<td>0.310</td>
<td>19</td>
<td>20.984</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>1.186</td>
<td>-0.133</td>
<td>0.270</td>
<td>19</td>
<td>22.739</td>
<td>16</td>
</tr>
<tr>
<td>17</td>
<td>1.528</td>
<td>3.000</td>
<td>0.200</td>
<td>19</td>
<td>16.166</td>
<td>16</td>
</tr>
<tr>
<td>18</td>
<td>1.524</td>
<td>3.000</td>
<td>0.230</td>
<td>19</td>
<td>18.037</td>
<td>16</td>
</tr>
<tr>
<td>19</td>
<td>1.431</td>
<td>-0.577</td>
<td>0.260</td>
<td>19</td>
<td>12.913</td>
<td>16</td>
</tr>
<tr>
<td>20</td>
<td>1.558</td>
<td>-1.006</td>
<td>0.230</td>
<td>19</td>
<td>8.311</td>
<td>16</td>
</tr>
<tr>
<td>21</td>
<td>1.362</td>
<td>-0.064</td>
<td>0.280</td>
<td>19</td>
<td>18.223</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>1.148</td>
<td>-0.188</td>
<td>0.210</td>
<td>19</td>
<td>37.083</td>
<td>16</td>
</tr>
<tr>
<td>23</td>
<td>1.121</td>
<td>0.695</td>
<td>0.310</td>
<td>19</td>
<td>21.203</td>
<td>16</td>
</tr>
<tr>
<td>24</td>
<td>0.954</td>
<td>-1.080</td>
<td>0.260</td>
<td>19</td>
<td>27.211</td>
<td>16</td>
</tr>
<tr>
<td>25</td>
<td>1.154</td>
<td>-1.532</td>
<td>0.250</td>
<td>19</td>
<td>37.834</td>
<td>16</td>
</tr>
<tr>
<td>26</td>
<td>0.793</td>
<td>-0.660</td>
<td>0.280</td>
<td>19</td>
<td>72.057</td>
<td>16</td>
</tr>
<tr>
<td>27</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>----</td>
</tr>
<tr>
<td>28</td>
<td>0.942</td>
<td>0.064</td>
<td>0.260</td>
<td>19</td>
<td>21.912</td>
<td>16</td>
</tr>
<tr>
<td>29</td>
<td>1.000</td>
<td>0.359</td>
<td>0.270</td>
<td>19</td>
<td>21.106</td>
<td>16</td>
</tr>
<tr>
<td>30</td>
<td>1.448</td>
<td>0.484</td>
<td>0.320</td>
<td>19</td>
<td>18.977</td>
<td>16</td>
</tr>
</tbody>
</table>
choosing the right answer among others is possible with a of residency, parameter c of item 6, 17, 18, 23, 28 words the item is very easy, in contrast deleted. meaningful, therefore the amounts of C lie between 0.160 (item 13) and 0.530 (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=0.350 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 1, 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 progress scores 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 of degree of freedom and 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), 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 a 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 progress scores 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)**

<table>
<thead>
<tr>
<th>NO</th>
<th>sex</th>
<th>θ</th>
<th>X1</th>
<th>NO</th>
<th>sex</th>
<th>θ</th>
<th>X1</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>male</td>
<td>-1.59</td>
<td>64.10</td>
<td>11</td>
<td>male</td>
<td>.96</td>
<td>89.61</td>
</tr>
<tr>
<td>02</td>
<td>female</td>
<td>1.47</td>
<td>94.65</td>
<td>12</td>
<td>male</td>
<td>-.91</td>
<td>70.95</td>
</tr>
<tr>
<td>03</td>
<td>male</td>
<td>-.15</td>
<td>78.53</td>
<td>13</td>
<td>male</td>
<td>.19</td>
<td>81.90</td>
</tr>
<tr>
<td>04</td>
<td>male</td>
<td>-.53</td>
<td>74.72</td>
<td>14</td>
<td>female</td>
<td>.13</td>
<td>81.26</td>
</tr>
<tr>
<td>05</td>
<td>male</td>
<td>-.20</td>
<td>78.00</td>
<td>15</td>
<td>female</td>
<td>.22</td>
<td>82.17</td>
</tr>
<tr>
<td>06</td>
<td>female</td>
<td>-.62</td>
<td>73.77</td>
<td>16</td>
<td>female</td>
<td>.14</td>
<td>81.44</td>
</tr>
<tr>
<td>07</td>
<td>male</td>
<td>.36</td>
<td>83.63</td>
<td>17</td>
<td>male</td>
<td>.30</td>
<td>83.03</td>
</tr>
<tr>
<td>08</td>
<td>male</td>
<td>.30</td>
<td>82.95</td>
<td>18</td>
<td>male</td>
<td>-.26</td>
<td>77.37</td>
</tr>
<tr>
<td>09</td>
<td>.64</td>
<td>86.43</td>
<td>19</td>
<td>male</td>
<td>.55</td>
<td>85.49</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-.05</td>
<td>79.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 5.0 (items 12, 47, 21, 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
certain 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 items 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). b< 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 test of internship, parameter c shows the
probability of selecting the right choice by chance. Its
amounts vary between 0.150 (item 48) to 0.410 (item 21).
Thus 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.
Progress scores scores of the students lie between 46.94 to 90.46.

DISCUSSION

The main purpose of this study was to refine the 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 selected randomly from 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 that this test has
more than one dimension. For the determination of the
propenseness of this test with the examinee-awareness
function and specific slope function of the test 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, judgement capability, and slope line of each
item expressed that in file 1 which related to residents items
1, 5, and 27 had better be omitted because their line
slopes was out of the interval 0.85 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 of low.
In a 31 item test which related to the residency level, 7
items had to be deleted because 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 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. Pashashari H. The fundamentals of


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. Zolfaghari 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
7. Abbassi F. Studying the second scale 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 parameter IRT model. Islamic Azad University of Tehran, 1997. (Persian).


