Item analysis and evaluation in the examinations in the faculty of medicine at Ondokuz Mayis University

I Tomak, Y Bek

Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ondokuz Mayis University, Samsun, Turkey

Abstract

**Background:** Item analysis is an effective method in the evaluation of multiple-choice achievement tests. This study aimed to compare the classical and the latent class models used in item analysis, as well as their efficacy in the evaluation of the examinations of the medical faculty.

**Materials and Methods:** The achievement tests in the medical faculty were evaluated using different methods. The two methods used were the classical and the latent class models. Among the classical methods, Cronbach’s alpha, split half methods, item discrimination, and item difficulty was investigated. On the other hand, various models of item response theory (IRT) and their statistics were compared in the group of latent class methods.

**Results:** Reliability statistics had values above 0.87. Item no. 7 was found easy, item no. 45 difficult and item no. 64 fairly difficult according to the evaluations done by classical and item response theories. In terms of item discrimination, item no. 45 had lower, item no. 7 had middle and item no. 64 had high discrimination levels. The distribution graph shows that personal abilities are good enough to tick the correct choice.

**Conclusion:** In this study, similar results were obtained by classical and latent methods. IRT can be considered perfect at a mathematical level, and if its assumptions are satisfied, it can easily perform assessments and measurements for most types of complex problems. Classical theory is easy to understand and to apply, while IRT is, on the contrary, sometimes rather difficult to understand and to implement.

**Key words:** Classical test theory, item analysis, item difficulty, item discrimination, item response theory, reliability

**Date of Acceptance:** 05-Nov-2014

Introduction

Item analysis is a general term that refers to the specific methods used in education to evaluate test items, typically for the purpose of test construction and revision.\(^1\)-\(^3\)

Item analysis methods include the classical and the latent class models. The classical test theory (CTT) is the most commonly used method in item analysis. In the classical analysis, the score of any test is calculated from the sum of the true value and a random error.\(^4\)-\(^6\) Latent class model is the probability of answering an item correctly or of attaining a particular response level. This model includes different methods regarding the item response theory (IRT) and the Rasch analysis.\(^7\)

The aim of this study was to comparatively evaluate the CTT and IRT models used in the item analysis and to determine the efficacy of these models for the evaluation of medical faculty examinations. Furthermore, the study also attempted to identify the most suitable evaluation criteria, evaluation method, and computer programs within the context of the relevant models.

Materials and Methods

**Participants**
In this study, both the CTT and the IRT methods were...
applied to the multiple choice examinations used to assess
the education provided in the Medical Faculty at the
Ondokuz Mayys University. The study data obtained from
the multiple-choice progress tests taken by 207 5th-year
students in the medical faculty. The test included 87
questions.

Data analysis was performed using the Iteman\(^8\) package
program for the CTT model, and the NCSS\(^9\) and
RUMM\(^10\) package programs for the IRT model.

Methods in the classical test theory
The methods of the CTT are used for three purposes.
These include the statistics related to the reliability of
the test, the item difficulty, and the item discrimination.\(^{11-13}\)
The reliability statistics of the CTT include a scale with
multiple-choice items (Likert-type scale), and serve to
evaluate the correlation between the items of the scale.\(^{14,15}\)
In this study, the Cronbach’s alpha coefficient and the
Spearman–Brown correlation were used to ensure internal
consistency.\(^{16-18}\)

Item difficulty and item discrimination are part of the item
analyses performed within the context of the CTT.\(^{19,21}\) For
a particular item, the item difficulty can be defined as the
ratio of those who provide correct answers.\(^{22-24}\)

The biserial point correlation (\(r_{\text{bis}}\)) is an advanced
measurement method that indicates the selectivity or
discrimination, of an item.\(^{25-27}\) The value for the biserial
point correlation is calculated as shown in Formula 1.

\[
r_{\text{bis}} = \frac{\overline{x}_1 - \overline{x}_0}{S_x} \cdot \sqrt{\frac{n_1n_0}{n(n-1)}}
\]  

(1)

In this formula, \(\overline{x}_1\) represents the mean score of those who
answered the item correctly; \(\overline{x}_0\), the mean score of those who
answered the item incorrectly; \(S_x\), the standard deviation
of the mean score; \(n_1\), the number of individuals who
answered the item correctly; \(n_0\), the number of individuals
who answered the item incorrectly; and \(n\) represents the
total number of individuals who answered the item, both
correctly or incorrectly.\(^{26}\)

Methods in the item response theory
With the IRT, it is possible to identify on a graph the ratio for
answering questions of individuals with different skill levels,
and also to determine the item difficulty and discrimination
values.\(^{28-30}\)

One-parameter logistic model and Rasch model
The one-parametered logistic (1 PL) model and the
Rasch model are IRT models. They represent the
simplest of the IRT models.\(^{31}\) In these models, \(\theta\)
defines the skill of the individual, while \(b\) indicates the
item difficulty. An individual’s probability of answering
an item correctly is defined as the function of the
ratio between an individual’s skill level and the item
difficulty.\(^{32,33}\)

In a 1 PL model of the test, the probability of correctly
answering the item “i” is given in Formula 2:

\[
P_i(\theta) = \frac{e^{(0-b_i)}}{1 + e^{(0-b_i)}} = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}
\]

(2)

Two-parameter logistic model
In most tests, the items differ not only according to their
level of difficulty, but also according to their discriminative
power. The two-item logistic model is abbreviated as
Two-parameter logistic (2 PL).\(^{6}\) In this model, a new
parameter is introduced into the model, in addition
to the item difficulty. This second parameter is item
discrimination.\(^{34,35}\) The model for 2 PL is provided in
Formula 3:

\[
P_i(\theta) = \frac{\exp[\alpha_i(\theta - b_i)]}{1 + \exp[\alpha_i(\theta - b_i)]}
\]

(3)

\(\alpha_i\) is the discrimination parameter for item \(i\).

Results

Evaluation with the classical method
According to the study data obtained from 207 students for
84 question items, the mean score was 47.729 ± 11.786,
the lowest score was 15, the highest score was 73, the mean
difficulty value was 0.570, and the mean discrimination
value was 0.277.

The reliability assessment regarding the 84-item test is
provided below in Table 1. The alpha coefficient and the
Spearman–Brown value on this table were obtained by using
the “dividing into two” method. All the obtained results and
reliability statistics had values above 0.87.

The items that constitute a test can have different
characteristics. The answering ratio of these items, the
group in which they are answered correctly at a higher
rate, and their difficulty and discrimination level can all be

---

**Table 1: Reliability analyses regarding the test**

<table>
<thead>
<tr>
<th>Alpha</th>
<th>SE*</th>
<th>Spearman–Brown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Random First-last Single-double</td>
</tr>
<tr>
<td>0.889</td>
<td>3.927</td>
<td>0.886</td>
</tr>
</tbody>
</table>

*Standard error
identified through evaluations performed at an item-level. The results of the evaluations regarding the difficulty and discrimination of the question items are provided in Table 2.

Three items that could be considered as easy, difficult, and moderately difficult with regards to item difficulty were evaluated in further detail. These easy, difficult, and moderately difficult items were Item 7, Item 45, and Item 64, respectively. For Items 7, 45, and 64, the item difficulty values were 0.937, 0.180, and 0.515, while the item discrimination values were 0.246, 0.111, and 0.481, respectively.

Information regarding the ratios at which these questions were answered correctly by each skill group is provided in Table 3. For Item 7, the correct answer “E” was selected by >80% of individuals from different skill groups. In Item 45, on the other hand, the correct answer “E” was selected only by 37 individuals, while the distracting answer “D” was selected by 101 individuals. For Item 64, the correct answer “B” was selected at significantly higher rates by 106 individuals with higher skill levels, with the ratio of correct answers increasing in parallel to the skill level.

For Item 64, the correct answer “B” was selected at significantly higher rates by individuals with higher skill levels, with the ratio of correct answers increasing in parallel to the skill level.

The graph regarding the ratios of answers/choices selected by the study participants for Items 7, 45, and 64 is provided in Figure 1. This graph shows that Item 7 was correctly answered by the large majority of individuals in all skill groups, while Item 64 was correctly answered mostly by individuals in the high skill group. In Item 45, on the other hand, it was observed that the incorrect answers were selected more frequently than the correct answer.

![Graph of Item 7, 45, and 64 answers](image)

**Table 2: Evaluation of item difficulty and item discrimination**

<table>
<thead>
<tr>
<th>Item difficulty</th>
<th>$r_{pbis}$</th>
<th>Number</th>
<th>Item difficulty</th>
<th>$r_{pbis}$</th>
<th>Number</th>
<th>Item difficulty</th>
<th>$r_{pbis}$</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.704</td>
<td>0.197</td>
<td>22</td>
<td>0.681</td>
<td>0.482</td>
<td>43</td>
<td>0.797</td>
<td>0.353</td>
<td>64</td>
</tr>
<tr>
<td>0.686</td>
<td>0.333</td>
<td>23</td>
<td>0.493</td>
<td>0.331</td>
<td>44</td>
<td>0.485</td>
<td>0.274</td>
<td>65</td>
</tr>
<tr>
<td>0.743</td>
<td>0.298</td>
<td>24</td>
<td>0.324</td>
<td>0.096</td>
<td>45</td>
<td>0.180</td>
<td>0.111</td>
<td>66</td>
</tr>
<tr>
<td>0.826</td>
<td>0.374</td>
<td>25</td>
<td>0.401</td>
<td>0.137</td>
<td>46</td>
<td>0.623</td>
<td>0.290</td>
<td>67</td>
</tr>
<tr>
<td>0.609</td>
<td>0.457</td>
<td>26</td>
<td>0.449</td>
<td>0.255</td>
<td>47</td>
<td>0.319</td>
<td>-0.017</td>
<td>68</td>
</tr>
<tr>
<td>0.884</td>
<td>0.167</td>
<td>27</td>
<td>0.922</td>
<td>0.373</td>
<td>48</td>
<td>0.295</td>
<td>0.129</td>
<td>69</td>
</tr>
<tr>
<td>0.937</td>
<td>0.246</td>
<td>28</td>
<td>0.657</td>
<td>0.215</td>
<td>49</td>
<td>0.315</td>
<td>0.209</td>
<td>70</td>
</tr>
<tr>
<td>0.696</td>
<td>0.391</td>
<td>29</td>
<td>0.386</td>
<td>0.300</td>
<td>50</td>
<td>0.903</td>
<td>0.419</td>
<td>71</td>
</tr>
<tr>
<td>0.824</td>
<td>0.354</td>
<td>30</td>
<td>0.835</td>
<td>0.271</td>
<td>51</td>
<td>0.665</td>
<td>0.359</td>
<td>72</td>
</tr>
<tr>
<td>0.193</td>
<td>0.011</td>
<td>31</td>
<td>0.541</td>
<td>0.313</td>
<td>52</td>
<td>0.188</td>
<td>0.181</td>
<td>73</td>
</tr>
<tr>
<td>0.739</td>
<td>0.338</td>
<td>32</td>
<td>0.792</td>
<td>0.420</td>
<td>53</td>
<td>0.280</td>
<td>-0.069</td>
<td>74</td>
</tr>
<tr>
<td>0.382</td>
<td>0.100</td>
<td>33</td>
<td>0.629</td>
<td>0.343</td>
<td>54</td>
<td>0.563</td>
<td>0.374</td>
<td>75</td>
</tr>
<tr>
<td>0.748</td>
<td>0.277</td>
<td>34</td>
<td>0.333</td>
<td>0.263</td>
<td>55</td>
<td>0.768</td>
<td>0.431</td>
<td>76</td>
</tr>
<tr>
<td>0.228</td>
<td>0.196</td>
<td>35</td>
<td>0.539</td>
<td>0.415</td>
<td>56</td>
<td>0.643</td>
<td>0.372</td>
<td>77</td>
</tr>
<tr>
<td>0.286</td>
<td>0.244</td>
<td>36</td>
<td>0.126</td>
<td>0.021</td>
<td>57</td>
<td>0.734</td>
<td>0.281</td>
<td>78</td>
</tr>
<tr>
<td>0.623</td>
<td>0.196</td>
<td>37</td>
<td>0.652</td>
<td>0.236</td>
<td>58</td>
<td>0.430</td>
<td>0.247</td>
<td>79</td>
</tr>
<tr>
<td>0.498</td>
<td>0.342</td>
<td>38</td>
<td>0.312</td>
<td>0.319</td>
<td>59</td>
<td>0.870</td>
<td>0.382</td>
<td>80</td>
</tr>
<tr>
<td>0.556</td>
<td>0.253</td>
<td>39</td>
<td>0.400</td>
<td>0.341</td>
<td>60</td>
<td>0.752</td>
<td>0.399</td>
<td>81</td>
</tr>
<tr>
<td>0.386</td>
<td>0.320</td>
<td>40</td>
<td>0.870</td>
<td>0.295</td>
<td>61</td>
<td>0.594</td>
<td>0.354</td>
<td>82</td>
</tr>
<tr>
<td>0.937</td>
<td>0.349</td>
<td>41</td>
<td>0.728</td>
<td>0.241</td>
<td>62</td>
<td>0.338</td>
<td>0.143</td>
<td>83</td>
</tr>
<tr>
<td>0.464</td>
<td>0.293</td>
<td>42</td>
<td>0.778</td>
<td>0.211</td>
<td>63</td>
<td>0.688</td>
<td>0.300</td>
<td>84</td>
</tr>
</tbody>
</table>

$r_{pbis}$: The biserial point correlation
Evaluation with the item response theory
The item characteristics curve (ICC) for Item 7 is provided in Figure 2. According to this ICC, the item difficulty and item discrimination for Item 7 were −1.975 and 1.523, respectively.

For Item 45, the item difficulty and item discrimination were −3.505 and 0.504, respectively. The ICC for this item is provided in Figure 3.

For Item 64, the item difficulty and item discrimination were −0.003 and 1.604, respectively. The ICC for Item 64 is provided in Figure 4.

Another important method used in assessing multiple-choice tests is the distraction analysis. This method allows the evaluation of the answer-selection processes. The distraction analysis for Item 7 is shown in Figure 5.

The distraction analysis for Item 45 is shown in Figure 6.

The distraction analysis for Item 64 is shown in Figure 7.

The distribution graph illustrating item difficulty together with the skill level of the individuals is shown in Figure 8.
Measurements and assessments play an important role in the evaluation of the education provided by medical faculties. In this study, the CTT and IRT methods were employed to evaluate the multiple-choice test used in medical faculty examinations.

The reliability analysis values for the test (exam) were above 0.85. The Cronbach’s alpha coefficient of the test was 0.89, while the Spearman–Brown coefficient was 0.87 and above. These values indicate that the test items were well designed, and that they together fulfilled the same purpose. The fact that these statistics used for reliability analysis had values above 0.80 served to demonstrate the overall reliability of the test.

Based on the evaluation performed with the CTT, the item difficulty and the item discrimination values for the entire test consisting of 84 questions were determined as 0.570 and 0.277, respectively. When the results were considered together with the item difficulty, it was determined that the test had moderate level strength. The item difficulty index varies between 0 and 1. A difficulty value close to 1 indicates an easy item, while a value near 0.50 indicates a moderately difficult item, and a value near 0 indicates a difficult item.

Very difficult and very easy items are not sufficient in themselves for distinguishing well-performing and poorly-performing students. Most of the items that constitute a test should be moderately difficult, and the test items should cover a wide range of difficulty levels that can effectively assess individuals from all levels of skill.

The items were also evaluated with respect to discrimination, in order to distinguish those knowledgeable about the test subject and those who were not. With respect to item discrimination, 3 (3.57%) of the items were within the −0.1–0.0 range, 6 (7.14%) were within the 0.0–0.1 range, 11 (13.10%) were within the 0.1–0.2 range, 25 (29.76%) were within the 0.2–0.3 range, 28 (33.33%) were within the 0.3–0.4 range, and 11 (13.10%) of the items were within the 0.4–0.5 range. With respect to the biserial point correlation, a value below 0 indicated a negative discrimination power, a value between 0 and 0.14 indicated a weak discrimination power, a value between 0.15 and 0.25 indicated a moderate discrimination power, a value between 0.26 and 0.35 indicated a good discrimination power, and a value above 0.35 indicated very good discrimination power. A higher discrimination index is more effective in distinguishing individuals with low and high skill levels.

The items were also evaluated with respect to discrimination, in order to distinguish those knowledgeable about the test subject and those who were not. With respect to item discrimination, 3 (3.57%) of the items were within the −0.1–0.0 range, 6 (7.14%) were within the 0.0–0.1 range, 11 (13.10%) were within the 0.1–0.2 range, 25 (29.76%) were within the 0.2–0.3 range, 28 (33.33%) were within the 0.3–0.4 range, and 11 (13.10%) of the items were within the 0.4–0.5 range. With respect to the biserial point correlation, a value below 0 indicated a negative discrimination power, a value between 0 and 0.14 indicated a weak discrimination power, a value between 0.15 and 0.25 indicated a moderate discrimination power, a value between 0.26 and 0.35 indicated a good discrimination power, and a value above 0.35 indicated very good discrimination power. A higher discrimination index is more effective in distinguishing individuals with low and high skill levels.

In light of this information, it is possible to predict that the test items had good discrimination, and that they could effectively distinguish those who with adequate and inadequate knowledge on the test subject.

The answer choices can be easily evaluated based on detailed information regarding the items. Within the
context of this study, Item 7, Item 45, and Item 64 were evaluated in detail. For Item 7, the item difficulty and item discrimination values were 0.937 and 0.246, respectively. This easy item was correctly answered by 194 of the students; however, the item’s level of discrimination was not good. Even in the group with the lowest level of skill, 83% of the individuals were able to correctly answer this question. For Item 45, the item difficulty and item discrimination values were 0.180 and 0.111, respectively. In this very difficult item with very low discrimination, the distraction choice was often selected instead of the correct choice. It can hence be described as an item that did not fulfill its function very well. For Item 64, the item difficulty and item discrimination values were 0.515 and 0.481, respectively. For this item with very high discrimination, the ratio of individuals who provided correct answers increased significantly, parallel to the increase in level of skill. It can hence be described as a very well designed item. The graphs used in the CTT method clearly illustrate the relevant information regarding the items.

The IRT model was also used to evaluate the multiple-choice test. The ratios of individuals who correctly answered Item 7, Item 45, and Item 64 were 93.12%, 16.93%, and 55.03%, respectively.

The difficulty value for Item 7 was determined as −1.976, which indicated that the item was fairly easy. The difficulty value for Item 45, on the other hand, was determined as 3.506, which indicated that the item was difficult. The difficulty value for Item 64 was determined as −0.003, which showed that the item was moderately difficult. The item discrimination values for Item 7, Item 45, and Item 64 were 0.524, 0.504, and 1.604, respectively. While Item 45 had low discrimination, Item 7 had a moderate level discrimination, and Item 64 had the highest level of discrimination.

With the distraction analysis, it is possible to evaluate a question or item in a multiple choice test in further detail within the context of the IRT method. For Item 7, it was observed that the correct answer was the most frequently selected answer choice, with a frequency ratio of nearly 1. Item 45 was a very difficult item with very low discrimination, which also included distraction choices. In Item 45, the distraction choices were selected by 49.74% of the individuals taking the test. It is often necessary to revise items in which the distraction choice is selected more often than the correct choice. In Item 64, the correct answer was selected more frequently than the other choices, and an increase in skill level was associated with a higher frequency of correct answers.

The IRT model includes a breakpoint graph, or individual-item graph, that allows the evaluation of the individuals’ skill levels and the items’ difficulty values on the same axis. On this graph, the skill of the individuals coincides with the item difficulty levels, thus indicating the level of items they can answer correctly. Depending on the skills of an individual, it is possible to see on this graph the items that are very easy for him/her, and the items that are beyond his/her skills.

The individual-item map functions in a similar way to the graph, although it is somewhat more detailed. This map lists the items that correspond to the skill level of the individuals, which renders the whole evaluation process much easier. For example, the difficulty value of Item 45 is between 1.60 and 2.40, the difficulty value of Item 64 is between 0 and 0.80, and the difficulty value of Item 7 is between −2.40 and −3.20. Based on these values, it is possible to categorize Item 45 as a difficult question, Item 64 as an easy question, and Item 30 as a very easy question. Based on this graph, it is also possible to observe that questions 7, 20, and 27 were very easy in comparison to the level of skill of the individuals.

While CTT is a commonly used method, IRT represents a newly-developing system. The most important advantage of CTT is that is based on relatively weak theoretical assumptions, and that it can be easily used for most tests. On the other hand, assumptions in IRT are very strong. IRT can be considered perfect at a mathematical level, and if its assumptions are satisfied, it can easily perform assessments and measurements for most types of complex problems. CTT is easy to understand and to apply, while IRT is, on the contrary, sometimes rather difficult to understand and to implement.

Item response theory was initially developed in order to resolve the problems associated with the use of the CTT. As can be understood from its name, IRT primarily focuses on information at an item level, while CTT, on the other hand, primarily focuses on information at a test level.

Classical test theory has two main limitations. The first of these is that the listing of the items in the CTT does not involve the formation of a set. The second limitation is that the data in the listing scale are not collectible. IRT provides a solution for these limitations. Unlike CTT, IRT involves modeling, and these models are used for the prediction of model parameters (individual and item parameters). These models are mainly used for the evaluation of collected data (i.e. of the answers provided by the individuals). As the evaluated characteristic of an individual is calculated from the answers he/she provides to each item, the predicted score of the IRT provides a more accurate prediction than the total score obtained by using the CTT.

Item response theory has numerous advantages in comparison to CTT. While CTT only provides a standard error for its measurements and a single estimation regarding reliability, the IRT model effectively demonstrates the sensitivity of the scale for all the underlying latent
variable. Another disadvantage of CTT is that during the analysis of the participants’ scores, it is dependent on the test items/questions. IRT, on the other hand, is independent of the test items/questions used to assess the level of skill of individuals.[41] As the expected score for the subjects/participants are calculated based on the subjects’ answers to each item, the predicted scores of the IRT method are more sensitive to differences between individual answer patterns. Furthermore, these predictions have a better probability of being correct than the predictions made by using scores obtained with the CTT method.[40]

Conclusion

An examination consisting of multiple-choice items was evaluated with the CTT and the IRT by using computer programs. The test items were examined through evaluations performed at both test-and item-levels. Reliability values for the test as a whole were obtained only with the CTT method, while reliability statistics at an item-level were obtained with both CTT and IRT. For both CTT and IRT, the item-related statistics were determined by using the similar approaches. Although IRT is superior as a method compared with CTT, novel computer programs have facilitated researchers to obtain graphs and results from the CTT approach that were nearly as informative as the results obtained with the IRT approach. However, due to the certain negative characteristics of CTT, it is ultimately up to the researcher to decide which method should be used.

References


How to cite this article: Tomak L, Bek Y. Item analysis and evaluation in the examinations in the faculty of medicine at Ondokuz Mayis University. Niger J Clin Pract 2015;18:387-94.

Source of Support: This research was supported by the Ondokuz Mayis University Scientific Research Project Coordination Unit (project no: PYO. TIP.1904.12.017).

Conflict of Interest: None declared.

---

New features on the journal’s website

Optimized content for mobile and hand-held devices
HTML pages have been optimized for mobile and other hand-held devices (such as iPad, Kindle, iPod) for faster browsing speed.
Click on [Mobile Full text] from Table of Contents page.
This is simple HTML version for faster download on mobiles (if viewed on desktop, it will be automatically redirected to full HTML version)

E-Pub for hand-held devices
EPUB is an open e-book standard recommended by The International Digital Publishing Forum which is designed for reflowable content i.e. the text display can be optimized for a particular display device.
Click on [EPub] from Table of Contents page.
There are various e-Pub readers such as for Windows: Digital Editions, OS X: Calibre/Bookworm, iPhone/iPod Touch/iPad: Stanza, and Linux: Calibre/Bookworm.

E-Book for desktop
One can also see the entire issue as printed here in a ‘flip book’ version on desktops.
Links are available from Current Issue as well as Archives pages.
Click on View as eBook