ORIGINAL ARTICLE

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

L 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]

Address for correspondence:
Dr. Leman Tomak,
Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ondokuz Mayis University, Samsun, Turkey.
E-mail: lemantomak55@gmail.com

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

Access this article online	
Quick Response Code:	**Website**: www.njcponline.com
	DOI: 10.4103/1119-3077.151720
	PMID: 25772924

applied to the multiple choice examinations used to assess the education provided in the Medical Faculty at the Ondokuz Mayıs 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_{pbis}) 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.

$$r_{pbis} = \frac{\overline{x}_1 - \overline{x}_0}{S_x} \cdot \sqrt{\frac{n_1 n_0}{n(n-1)}} \quad (1)$$

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, θ 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:

$$P_i(\theta) = P(x_i = 1/\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} \quad (2)$$

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:

$$P_i(\theta) = \frac{\exp[\alpha_i(\theta - b_i)]}{1 + \exp[\alpha_i(\theta - b_i)]} \quad (3)$$

α_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

Alpha	SE*	Spearman-Brown		
		Random	First-last	Single-double
0.889	3.927	0.886	0.872	0.889

*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.

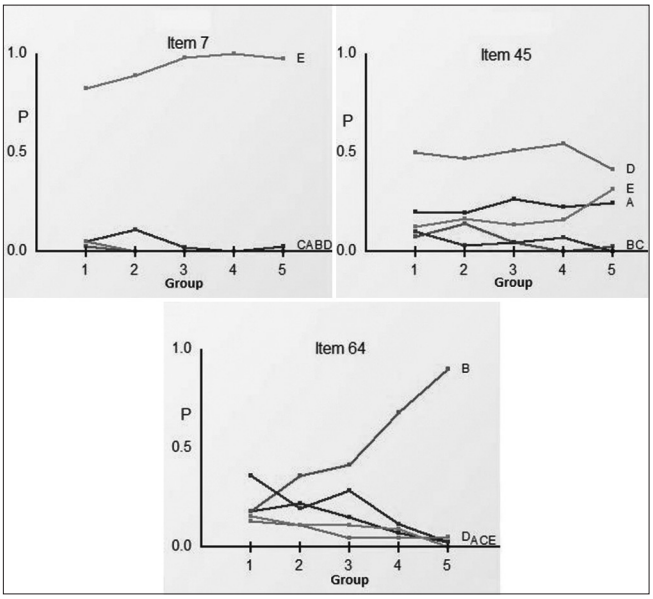

Figure 1: The answer ratios of items 7, 45, and 64

Table 2: Evaluation of item difficulty and item discrimination

Item difficulty	r_{pbis}	Number	Item difficulty	r_{pbis}	Number	Item difficulty	r_{pbis}	Number	Item difficulty	r_{pbis}
0.704	0.197	22	0.681	0.482	43	0.797	0.353	64	0.515	0.481
0.686	0.333	23	0.493	0.331	44	0.485	0.274	65	0.676	0.409
0.743	0.298	24	0.324	0.096	45	0.180	0.111	66	0.417	0.345
0.826	0.374	25	0.401	0.137	46	0.623	0.290	67	0.580	0.369
0.609	0.457	26	0.449	0.255	47	0.319	−0.017	68	0.461	0.205
0.884	0.167	27	0.922	0.373	48	0.295	0.129	69	0.536	0.311
0.937	0.246	28	0.657	0.215	49	0.515	0.209	70	0.430	0.368
0.696	0.391	29	0.386	0.300	50	0.903	0.419	71	0.783	0.386
0.824	0.354	30	0.835	0.271	51	0.665	0.359	72	0.510	0.357
0.193	0.011	31	0.541	0.313	52	0.188	0.181	73	0.382	0.080
0.739	0.338	32	0.792	0.420	53	0.280	−0.069	74	0.820	0.393
0.382	0.100	33	0.629	0.343	54	0.563	0.374	75	0.432	0.144
0.748	0.277	34	0.333	0.263	55	0.768	0.431	76	0.718	0.421
0.228	0.196	35	0.539	0.415	56	0.643	0.372	77	0.210	0.085
0.286	0.244	36	0.126	0.021	57	0.734	0.281	78	0.671	0.406
0.623	0.196	37	0.652	0.236	58	0.430	0.247	79	0.369	0.132
0.498	0.342	38	0.312	0.319	59	0.870	0.382	80	0.733	0.271
0.556	0.253	39	0.400	0.341	60	0.752	0.399	81	0.659	0.205
0.386	0.320	40	0.870	0.295	61	0.594	0.354	82	0.185	−0.016
0.937	0.349	41	0.728	0.241	62	0.338	0.143	83	0.873	0.464
0.464	0.293	42	0.778	0.211	63	0.688	0.300	84	0.482	0.283

*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.

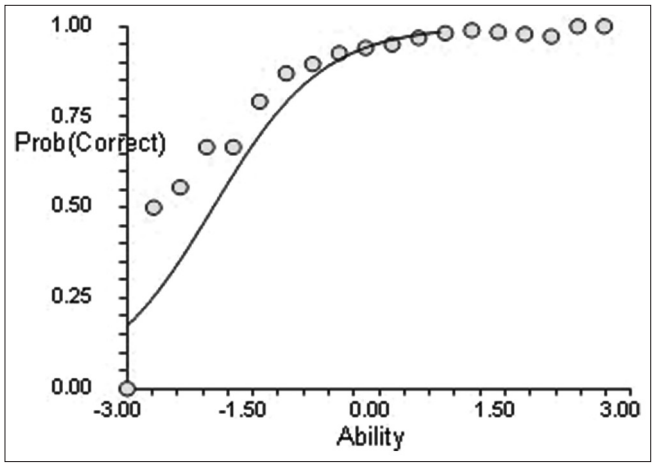

Figure 2: Item characteristics curve of item 7

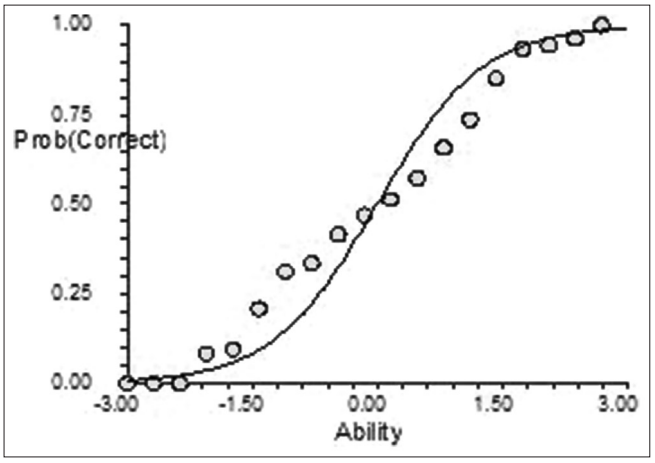

Figure 4: Item characteristics curve of item 64

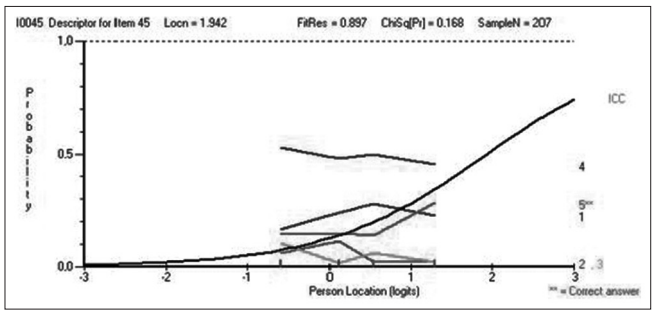

Figure 6: Distractor analysis for item 45

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.

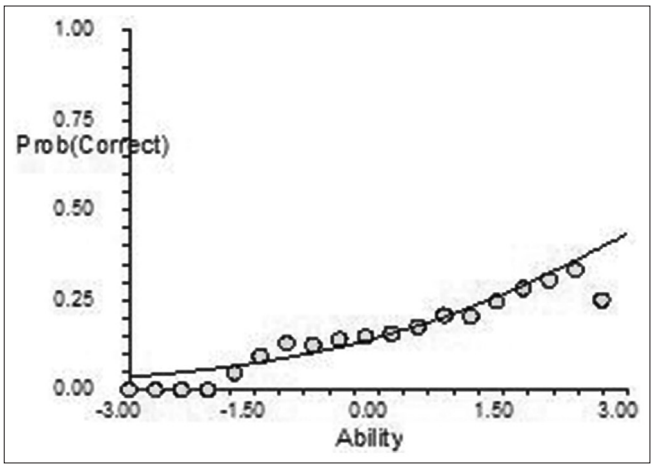

Figure 3: Item characteristics curve of item 45

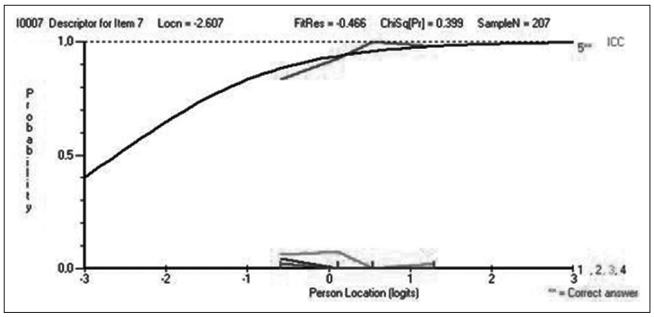

Figure 5: Distractor analysis for item 7

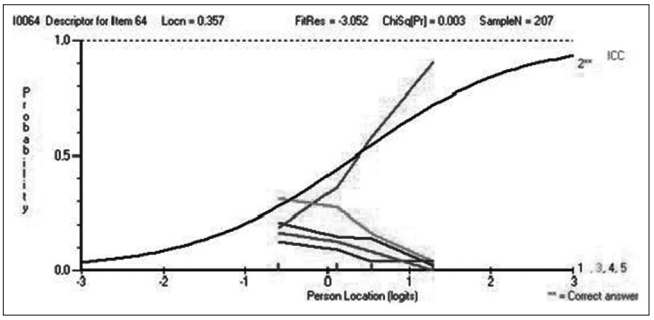

Figure 7: Distractor analysis for item 64

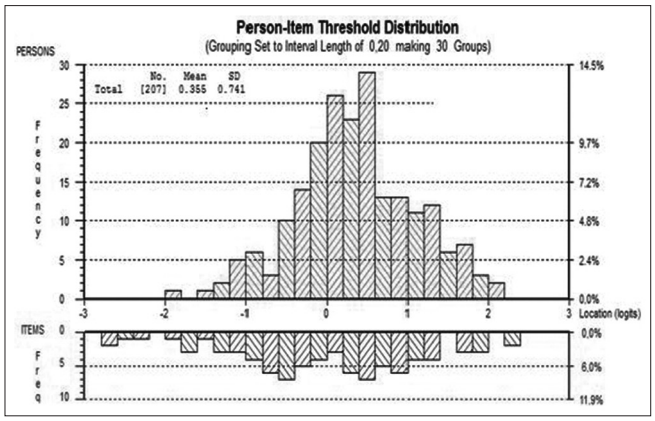

Figure 8: Individual and item distribution

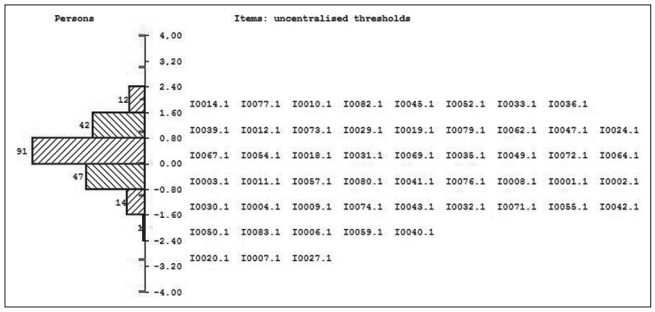

Figure 9: Individual and item map

Table 3: The answer ratios of items 7, 45 and 64

	n	Skill level (%)				
		0-20	20-40	40-60	60-80	80-100
Item 7						
A	2	0.050	0.000	0.000	0.000	0.000
B	1	0.025	0.000	0.000	0.000	0.000
C	8	0.050	0.111	0.022	0.000	0.024
D	2	0.050	0.000	0.000	0.000	0.000
E*	194	0.825	0.889	0.978	1.000	0.976
Item 45						
A	47	0.200	0.194	0.267	0.227	0.244
B	11	0.075	0.139	0.044	0.000	0.024
C	10	0.100	0.028	0.044	0.068	0.000
D	101	0.500	0.472	0.511	0.545	0.415
E*	37	0.125	0.167	0.133	0.159	0.317
Item 64						
A	26	0.179	0.222	0.152	0.068	0.024
B*	106	0.179	0.361	0.413	0.682	0.902
C	40	0.359	0.194	0.283	0.114	0.024
D	15	0.128	0.111	0.043	0.045	0.049
E	19	0.154	0.111	0.109	0.091	0.000

*Correct answer

The map illustrating item difficulty together with the skill level of the individuals is shown in Figure 9. These figures show the skill intervals to which the difficulty level of each item corresponded.

Discussion

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.[16-18] 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.[24,25] 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.[36,37]

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.[24,25] 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 1.524, 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.[38,39] 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.[40-42] CTT is easy to understand and to apply, while IRT is, on the contrary, sometimes rather difficult to understand and to implement.[13,43]

Item response theory was initially developed in order to resolve the problems associated with the use of the CTT.[14] 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 a test level.[44]

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.[45] 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.[2]

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.[6]

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

1. DeVellis RF. Classical test theory. Med Care 2006;44:S50-9.
2. De Grutijter DN, Van der Kamp LJ. Statistical Test Theory for the Behavioral Sciences. London: Chapman and Hall; 2008.
3. Allen DD. Validity and reliability of the movement ability measure: A self-report instrument proposed for assessing movement across diagnoses and ability levels. Phys Ther 2007;87:899-916.
4. Baker FB. The Basics of Item Response Theory. 2nd ed. USA: ERIC; 2001.
5. Traub RE. Classical test theory in historical perspective. Educ Meas Issues Pract 1997;16:8-14.
6. Reeve BB. An introduction to modern measurement theory. Bethesda, MD: National Cancer Inst.; 2002.
7. Recklase MD. Multidimensional Item Response Theory. New York: Springer; 2009.
8. Iteman. Classical Item Analiysis. Iteman Software version 4.2.1.2. St. Paul, Minnesota: Assessment Systems Corporation; 2012.
9. Hintze J. NCSS and GESS. Kaysville, Utah; 2007.
10. Andrich D, Sheridan B, Luo G. RUMM 2030 Version 5.4 for windows. RUMM Laboratory Pty Ltd.; 2012. p. 21.
11. Benson IG. A Comparison of Three Types of Item Analysis in Test Development Using Classical and Latent Trait Methods. USA: University of Florida; 1977.
12. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. USA: Cengage Learning; 2008.
13. Brennan RL. Generalizability theory and classical test theory. Appl Meas Educ 2011;24:1-21.
14. Dimitrov DM. Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. Educ Psychol Meas 2002;62:783-801.
15. Haertel EH. Reliability. In: Brennan RL, editor. Educational Measurement. 5th ed. Westport, CT: American Council on Education/Praeger; 2005.
16. Cronbach LJ. Essentials of Psyhological Testing. 4th ed. New York: Harper and Row; 1984.
17. McDonald RP. Test Theory: A Unified Treatment. Mabwah NJ: Lawrance Erlbaum Associates; 1999.
18. Bechger T, Gunter M, Huub H, Beguin A. Using classical test theory in combination with item response theory. Appl Psychol Meas 2003;27:319-34.
19. MacDonald P, Paunonen SV. A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. Educ Psychol Meas 2002;62:921-43.
20. Ural A, Kilic I. Scientific Research Process and Data Analysis with SPSS. 1st ed. Ankara: Detay; 2005.
21. Zaman A, Niwaz A, Faize FA, Dahar MA. Analysis of multiple choice items and the effect of items sequencing on difficulty level in the test of mathematics. Eur J Soc Sci 2010;17:61-7.
22. Lewis RF, Ortiz KK. All of the Above. A Guide to Classroom Testing and Evaluation. 2nd ed. Costa Mesa: CA: Economics Research; 1988.
23. Hotiu A. The Relationship between Item Difficulty and Discrimination Indices in Multiple-Choice Tests in a Physical Science Course. Florida: Florida Atlantic University; 2006.
24. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singapore 2006;35:67-71.
25. Escudero EB, Reyna NL, Morales MR. The level of difficulty and discrimation power of the basic knowledge and skills examination. Rev Electron Invest Educaitva 2000;2:1.
26. Osterlind S. Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats. 2nd ed. New York: Kluwer Academics; 2002.
27. Demars C. Item Response Theory. New York: Oxford Universtiy; 2010.
28. Li J, Liu H, Liu H, Feng T, Cai Y. Psychometric assessment of HIV/STI sexual risk scale among MSM: A Rasch model approach. BMC Public Health 2011;11:763.
29. Thorpe GL, Favia A. Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications. Psychology Faculty Scholarship; 2012.
30. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. 1st ed. USA: Sage; 1991.
31. Goetz C, Ecosse E, Rat AC, Pouchot J, Coste J, Guillemin F. Measurement properties of the osteoarthritis of knee and hip quality of life OAKHQOL questionnaire: An item response theory analysis. Rheumatology (Oxford) 2011;50:500-5.
32. Andrich D. The application of an unfolding model of PIRT type to the measurement of attitude. Appl Psychol Meas 1988;12:33-51.
33. Wilson M. Constructing Measures: An Item Response Modeling Approach. USA: Lawrence Erlbaum; 2005.
34. Zheng X, Rabe-Hesketh S. Estimating parameters of dichotomous and ordinal item response models with gllamm. Stata J 2007;7:313-33.
35. Baylari A, Montazer G. Design a personalised e-learning system based on item response theory and artificial neural network approach. Expert Syst Appl 2009;36:8013-21.
36. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and Evaluation in Psychology and Education. 5th ed. New York: MacMillan; 1991.
37. Karaca E, Yurdabakan I, Cetin B, Nartgun Z, Bicak B, Gomleksiz M. Measurements and Evaluation in Education. 2nd ed. Ankara: Nobel; 2010.
38. Hambleton RK, Jones RW. Comparison of classical test theory and item response. Educ Meas Issues Pract 1993;12:253-62.
39. Fan X. Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educ Psychol Meas 1998;58:357-81.
40. Cohen AS, Bottge BA, Wells CS. Using item response theory to assess effects of mathematics instruction in special populations. Except Child 2001;68:23-44.
41. Dapueto JJ, Francolino C, Servente L, Chang CH, Gotta I, Levin R, *et al.* Evaluation of the Functional Assessment of Cancer Therapy-General (FACT-G) Spanish Version 4 in South America: Classic psychometric and item response theory analyses. Health Qual Life Outcomes 2003;1:32.
42. Stage C. Classical Test Theory or Item Response Theory: The Swedish

Experience. Sweden University, Umea. Publications in Applied Educational Science; 2003.
43. Lin CJ. Comparisons between classical test theory and item response theory in automated assemly of parallel test forms. J Technol Learn Assess 2008;6:1-41.
44. Hambleton RK, Robin F, Xing D. Item response models for the analysis of educational and psychological test data. In: Tinsley H, Brown S, editors. Handbook of Applied Multivariate Statistics and Mathematical Modeling. San Diego, CA: Academic Press; 2000.
45. Prieto L, Alonso J, Lamarca R. Classical Test Theory versus Rasch analysis for quality of life questionnaire reduction. Health Qual Life Outcomes 2003;1:27.

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.