Journal of Experimental and Clinical Medicine
https://dergipark.org.tr/omujecm

Experimental Research

J. Exp. Clin. Med., 2017; 34(4): 257-262
doi: 10.5835/jecm.omu.34.04.006

Assessment of Medicine Faculty Biostatistics Exam with different models

Leman Tomak

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

ARTICLE INFO

Article History
Received 05 / 09 / 2017
Accepted 08 / 12 / 2017
Online Published Date 10 / 10 / 2019

*** Correspondance to:**
Leman Tomak
Department of Biostatistics and Medical Informatics,
Faculty of Medicine,
Ondokuz Mayis University,
Samsun, Turkey
e-mail: lemant@omu.edu.tr

Keywords:
Item difficulty
Item discrimination
KR-20
Rasch analysis

ABSTRACT

The most important goal of a test is having reliable and valid measurement tools. The purpose of this study is to conduct item analysis to questions for an exam which measures the biostatistics knowledge of students and to assess the Biostatistics Program. The study group consists of a total of 261 students in their second year of Ondokuz Mayıs University Faculty of Medicine. 132 (50.6%) of the students are female, while 129 (49.4%) are male. Item analysis was assessed by using classical method and Rasch analysis. The average value of the test which consisted of a total of 60 multiple choice questions was 47.47± 6.99; the lowest score was 15, while the highest score was 57. KR 20 value was found as 0.86. When all the questions were analyzed with Rasch analysis, item difficulty of 75% of the items was between -1.60 and 1.60. As a conclusion, the exam was found to be reliable and it was shown to be a moderately difficult exam which assessed the knowledge of the students. Future studies are planned to assess Biostatistics teaching in different levels of class.

1. Introduction

Miller's pyramid is the most important component in tracking, measuring and assessing the development of education process and it attracts attention especially in the assessment of health education. While the base of the pyramid forms knowledge and comprehension skills, the top of the pyramid forms performance (Vegada et al., 2016).

During the process of measuring knowledge, multiple choice questions are frequently preferred among different types of measurement tools used for error-free and objective measurement results (Mukherjee and Lahiri, 2015). Multiple choice questions are easy to conduct and classify and they enable an easy assessment of the students taking the exam. Suitably created multiple choice questions measure and analyze knowledge, comprehension and application. They can measure knowledge from a narrow field to a wide field and thus they are the type of questions used most frequently in medicine (Epstein, 2007; Brookhart, 2015).

Having reliable and valid measurement tools is the most important goal of a test. Thus, item analysis is conducted by using the application results of the test (Atılgan et al., 2011; Mukherjee and Lahiri, 2015). Item analysis of multiple-choice questions first of all assesses the quality of the items, while at the same time the test measures the skills in the area. Different

theoretical methods assess whether the items have desired characteristics (Abdalla 2011; Tomak and Bek, 2015). The most frequently used method, Classical Test Theory (CTT) assesses the test as a whole. This method based on correlation assumes that all the questions are equal indicators of an individual's knowledge in that subject (Allen, 2012; Petrillo et al., 2015). An alternative method to this is Rasch analysis, which is one of the Item Response Theory (IRT) models. This method, which was initially developed for educational purposes, can also be used especially to develop psychological assessment tools. When compared with the classical method, Rasch analysis is more stable in different populations (Trakman et al., 2017). In this theory, responses are modeled on individual items. Individual's response on test item shows the characteristics of the item and the individual. It is assumed that the individual's response (test taker's performance) depends on one or more factors, which are called latent characteristics or skills. Each item in the test or questionnaire is assumed to measure an underlying latent characteristic (De Grutijter and Van der Kamp, 2008; Pallant, 2016).

In this study, the purpose was to conduct item analysis and distractor analysis for multiple choice questions which measured the biostatistics knowledge of Ondokuz Mayis University (OMU) Faculty of Medicine students and to assess the Biostatistics Program.

2. Materials and methods

Participants

The study group consists of a total of 261 students in their second year of Ondokuz Mayıs University Faculty of Medicine. 132 (50.6%) of the students are female, while 129 (49.4%) are male.

Measures

The data set used in the study consists of multiple-choice questions for measuring biostatistics knowledge of students during the 2015-2016 academic year. Students' level of knowledge was assessed with a total of 60 questions including basic biostatistics subjects. The topic titles included in the exam were listed as summary of the data, sample, descriptive statistics, statistical significance, error types and hypothesis tests. The questions were prepared by taking the lesson hours into consideration and with content validity and applied on the students.

Analysis

Rasch method was used with CTT for the analysis of multiple-choice questions. In the analysis of the data, NCSS model was used for CTT model, while RUMM programs were used for Rasch model (Hintze, 2007; Andrich et al., 2012).

For multiple choice questions, the stage of creating and analyzing the distractors is very important for a healthy measurement. Distractors are choices given to mislead those who do not know or those who know little (Tarrant et al., 2009; Deepak et al., 2015). A good distractor should consist of expressions which are correct but which do not meet the requirements of the problem and which are incorrect although they seem to be correct. Distractors are reasonable, but clearly incorrect expressions (Collins, 2006; Kilgour and Tayyaba, 2016).

CTT is the most widely and most frequently used item analysis method and it connects some existing structures to hypothetical basis. In this theory, the degree of having a characteristic is found by adding up the answers given to items related to that subject in the scale. The score obtained is the observed score and it consists of the combination of actual score and error score (Crocker and Algina, 2008; Cappelleri et al., 2014). In a reliable measurement, error score will decrease and the observed score will get closer to the real score, in other words, reliability is decreasing the random errors in the measurement on which reliability is conducted. In order to find out the reliability of a measurement tool, internal consistency analyses, item difficulty and item discrimination can be used (Biswas et al., 2015; Abozaid et al., 2017).

Kuder- Richardson 20 (KR-20) coefficient was used to assess the internal consistency of the measurement tool analyzed in this study. KR-20 is used in the assessment of reliability of tests which consist of multiple choice questions the correct answers of which are coded as 1 and the incorrect answers of which are coded as 0 (Brennan 2011).

In the assessment of reliability in terms of items, average values of the items, item difficulty and item discrimination were analyzed (Zaman et al., 2010; Hingorjo and Jaleel, 2012). Item difficulty is the rate of individuals who answer the item correctly (Deepak et al., 2015). Biserial point correlation (rpbis), which is used to find out item discrimination, is a measurement which shows the association between test score and the score of answering the item correctly (Biswas et al., 2015; Abozaid et al., 2017).

Another method used in the study is Rasch analysis. It is the only model in Item Response Theory models which includes item difficulty. The probability of answering an item correctly can be found with the function of the rate of that individual's ability level to item difficulty (Demars, 2010; Petrillo et al., 2015). It has two assumptions as unidimensionality and local independence. The responses given to questions are defined by a unidimensional latent variable and the responses given to questions are independent of each other (Li et al., 2011; Pallant, 2016). With Item Characteristics Curve (ICC) in Rasch analysis, item

response possibility of individuals with different ability levels and difficulty values of the items can be found with graphs (De Grutijter and Van der Kamp, 2008; Trakman et al., 2017).

3. Results

The data obtained from the answers given to 60 multiple choice questions by a total of 261 students were analyzed. The average value of the test was 47.47±6.99; while the lowest score was 15 and the highest score was 57. KR 20 value of the test was 0.86. The questions, which assessed basic biostatistics knowledge, were assessed with CTT and Rasch analysis. Difficulty and discrimination of the questions were found, at the same time how the questions worked was analyzed with distractor analysis. In terms of analysis of all the questions with Rasch analysis, there were 45 (%75) items with an item difficulty between -1.60 and1.60.

The questions which showed different intervals were analyzed in more detail by taking the difficulty and discrimination of the questions into consideration. In variance analysis table, for Item 38 which asked the equivalent of variance value, the correct answer A was given by 141 (54%) students, while the other answers were given as B by 15 (6%) students; C by 49 (19%) students; D by 16 (6%) students and E by 40 (15%) students, respectively. In the assessment conducted with classical method for this question, item difficulty was found as 0.54, while discrimination was found as 0.40. Item difficulty was found as 1.87 with Rasch analysis. Fit residual value was found as -2.02. ICC for Item 38 is given in Fig. 1.

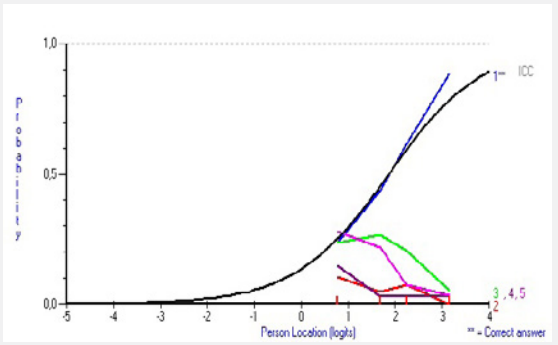

Fig. 1. ICC for Item 38.

The correct answer of Item 39 which assessed hypothesis test was E and it was given by 92 (35%) students, while the other answers were given as A by 12 (5%) students; B by 28 (11%) students; C by 8 (3%) students and D by 121 (46%) students, respectively. With CTT, item difficulty was found as 0.3, while discrimination was found as 0.31. With Rasch analysis, item difficulty was found as 2.70. Fit residual value was -0.42 (Fig. 2).

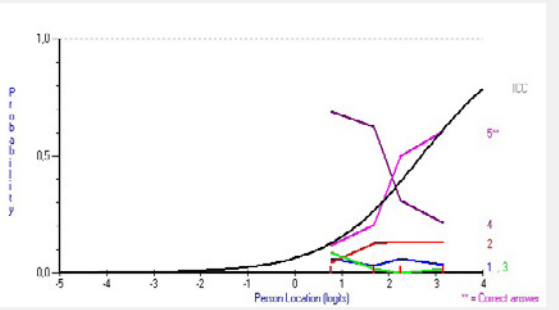

Fig. 2. ICC for Item 39.

The correct answer for Item 56 which assessed the graphic summary of data was C and it was answered correctly by 76 (29%) students. The other answers were given as A by 3 (1%) students, B by 81 (31%) students, D by 15 (6%) students and E by 84 (32%) students. Item difficulty was found as 0.29, while discrimination was found as -0.22. Item difficulty was found as 2.92 with Rasch analysis. Fit residual value was found as 0.64. Rasch analysis of the question is given with Fig. 3.

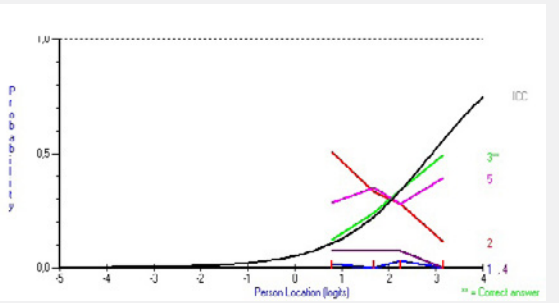

Fig. 3. ICC for Item 56.

The correct answer for Item 37 which assessed variation coefficient was C and it was given by 190 (73%) students, while the answer A was given by 42 (16%) students, B by 5 (2%) students; D by 4 (2%) students and E by 20 (8%) students. CTT assessment showed that item difficulty was 0.73 and discrimination was 0.37. Item difficulty was found as 0.87 with Rasch analysis. Fit residual value was -0.98. ICC for Item 37 is given with Fig. 4.

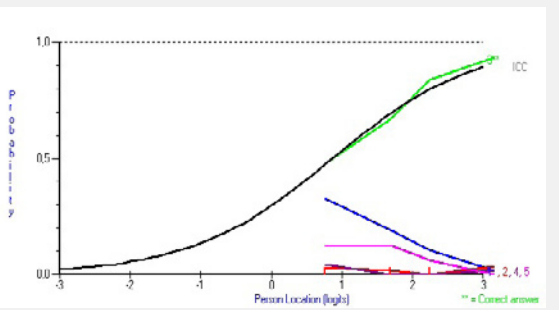

Fig. 4. ICC for Item 37.

The correct answer B for Item 36 which assessed hypothesis knowledge was given by 237 (91%) students, while A was given by 5 (2%) students, C was given by 3 (1%) students, D was given by 14 (5%) students and E was given by 2 (1%) students. In the assessment of this question made with classical method, item difficulty was found as 0.91 and discrimination was found as 0.38. Item difficulty was found as -0.63 with Rasch analysis. Fit residual value was 0.12. ICC for Item 36 is given with Fig. 5.

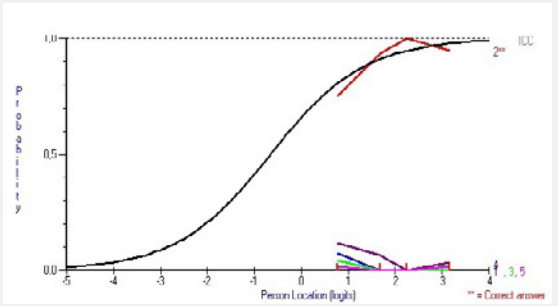

Fig. 5. ICC for Item 36.

4. Discussion

In the second year of their 6-year-long education, students of OMU Medicine Faculty get theoretical and practical classes about biostatistics in the block which includes basic information about biostatistics. The exam at the end of the block which includes multiple choice questions assesses students' level of knowledge and what they have learned.

Multiple choice questions are preferred especially when there are too many people taking the exam, when there is a need for selection and placement, when there are too many questions, when there is a need for quick results and when high reliability and validity is required. These questions which are developed for a specific content provide a consistently tested content (Collins, 2006; Mukherjee and Lahiri, 2015). The disadvantages of these questions are the fact that they are answered by remembering rather than recognizing the questions, there is the factor of luck in answering and developing them are difficult and time consuming. In this question type, since the information is mostly remembered within the item, it is insufficient to measure top level behaviors. This in turn causes negative effect on construct validity (Atılgan et al., 2011).

A test's being valid and reliable will help it to reach its goal. With this study, the test's analyses were conducted in general and on the basis of items (questions) and the characteristics of items which worked well and which did not work well were assessed over a few items.

In the general assessment of the reliability of the exam, KR-20 value was found as 0.86 and it is the equivalent of Cronbach Alpha in the assessment of the reliability of dichotomous questions (DeVellis 2006). A KR- 20 value of over 0.80 is required for a reliable measurement tool and it was found that the exam met this requirement. For KR-20, it is assumed that the questions are not equally difficult, while for KR-21, which is another type of KR, it is assumed that the questions are equally difficult. Thus, KR-20 is more reliable than KR-21 (Brennan 2011).

With assessment based on questions, item difficulty, discrimination and whether the distractors were well built were assessed. In the classical method, as the item difficulty level increases, it is understood that the item gets easier (Abozaid et al., 2017). Although the most ideal values for item difficulty are between 0.5 and 0.6 in an exam, 0.3-0.7 and 0.2-0.8 are recommended intervals (Sim and Rasiah, 2006; Hingorjo and Jaleel, 2012; Biswas et al., 2015). For item discrimination index, values over 0.25 are indicators of discrimination. Values over 0.40 are accepted as perfect, those between 0.30 and 0.40 are accepted as good and those between 0.25 and 0.30 are accepted as acceptable levels of discrimination (Deepak et al., 2015; Abozaid et al., 2017). In assessment with Rasch analysis, as the item difficulty becomes negative, the items become easier, while they become more difficult as it increases positively (De Grutijter and Van der Kamp, 2008; Trakman et al., 2017). Within this context, some questions –which had different difficulty and discrimination and distractor characteristics- were analyzed.

For item 38, item difficulty with CTT was 0.54, which can be accepted as the optimal value and a discrimination value of 0.40 can be assessed as good (Deepak et al., 2015; Abozaid et al., 2017). Since Fit residual values were between -/+ 2.5 for all items, the items were made fit for the model. With Rasch, item difficulty was 1.87, which was over the average (De Grutijter and Van der Kamp, 2008; Pallant, 2016). In terms of distractors, as ability increases, the rate of answering the question correctly also increases and the rate of incorrect answers decreases. According to these, it can be said that Item 38 is a good functioning, ideal item.

With Item 39, degree of freedom was questioned. In the question, degree of freedom (df) was asked for Student t test, the rate of highest distractor was higher than the correct answer and here df value was given for paired t test. With CTT, the difficulty of the question can be said to be acceptable and it can be said to have good discrimination (Biswas et al., 2015). With Rasch analysis, it can be described as a difficult question (Trakman et al., 2017). The important point here is the result that more talented students had higher rates of answering correctly and the highest distractor was marked by students who had lower levels of talent. The question was very distractive for students who had insufficient levels of information.

For Item 56 which was about graphical summary (box graph) in categorical data, the correct answer and the two distractors were answered in similar rates. According to the results obtained with both methods, the question is difficult and it does not have sufficient discrimination (Deepak et al., 2015; Pallant, 2016). It was found that the frequency of the correct answer increased as the level of talent increased, the frequency of the distractor in choice B (histogram answer) decreased as the level of talent increased and the distractor in choice E (line chart answer) was found to be high in high talent level. It was found that the students had superficial knowledge about this subject.

Item 37 which questioned the expression of variability with percentage had values close to the ones an achievement exam should have. Its discrimination can also be considered as good (Hingorjo and Jaleel, 2012). When analyzed with Rasch, it was found that it was an easy question and with ICC it was found that the correct answer was marked easily by the students (Pallant 2016). Since this question about the subject of descriptive analysis-variability measures is among the basic subjects students should know about, the percentage of correct answers can be accepted as normal.

Item 36, which was for the calculation of the expected value with Chi-square test, was found to be an easy question with both methods (De Grutijter and Van der Kamp, 2008; Trakman et al., 2017). In addition, discrimination had a value higher than 0.35 (Deepak et al., 2015; Abozaid et al., 2017). This result shows that this subject which was repeated in the applied lesson as well as the theoretical lesson was comprehended well by the students.

This study analyzed the exam which was conducted for measuring the knowledge of biostatistics lesson which requires the students to have specific basic knowledge during and after their medicine education and assessed the reliability of the exam in general and showed how the questions worked on the basis of item. These assessments made with different theoretical methods were also presented to readers visually with graphs. Since it was not possible to show all of the questions, five items which exemplified different difficulty and discrimination intervals were analyzed more closely.

As a conclusion, the reliability of the exam was revealed and it was shown that the exam was a moderately difficult exam which assessed the knowledge of the students. At the same time, multiple choice questions were assessed with analyses through different methods and their discriminations were analyzed according to different levels of talent.

Biostatistics has an important duty especially in terms of showing the power of evidence in medical practices based on evidence. Medical students need to have sufficient basic knowledge not only to be able to conduct their own practices but also to be able to follow scientific publications. Students should be given sufficient and necessary education during their medical studies and this education should be assessed. Thus, this study will play an important role in organizing the future programs and assessment of Biostatistics Education in different levels of classes is intended in future studies.

REFERENCES

Abdalla, M.E., 2011. What does item analysis tell us? Factors affecting the reliability of multiple choice questions. Gezira J. Health Sci.7,17-25.

Abozaid, H., Park, Y.S., Tekian, A., 2017. Peer review improves psychometric characteristics of multiple choice questions. Med. Teach. 39, 50-54.

Allen, D.D., 2012. Validity and reliabilityof the movement ability measue: A self-report instrument proposed for assessing movement across diagnoses and ability levels. Physical Thearapy. 87, 899-916.

Andrich, D., Sheridan, B., Luo, G., 2012. RUMM 2030 Version 5.4 for windows. RUMM Laboratory Pty Ltd.

Atılgan, H., Kan, A., Doğan, N., 2011, Eğitimde Ölçme ve Değerlendirme (5.Baskı). Anı Yayıncılık, Ankara.

Biswas, S.S., Jain, V., Agrawal, V., Bindra, M., 2015. Small group learning: Effect on item analysis and accuracy of self-sssessment of medical students. Educ. Health. 28, 16-21.

Brennan, R.L., 2011. Generalizability theory and classical test theory. Appl. Meas. Educ. 24, 1-21.

Brookhart, S., 2015. Making the most of multiple choice. Educ. Leadership. 73, 36-39.

Cappelleri, J.C., Lundy, J.J., Hays, R.D., 2014. Overview of classical test theory and item response theory for quantitative assessment of items in developing patientreported outcome measures. Clin. Ther. 5,648-662.

Collins, J., 2006. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. Radiographics. 26, 543-551.

Crocker, L., Algina, J., 2008. Introduction to classical and modern test theory. Cengage Learning, USA.

Deepak, K.K., Al-Umran, K.U., AI-Sheikh, M.H., Dkoli, B.V., Al-Rubaish, A., 2015. Psychometrics of multiple choice questions with nonfunctioning distracters: Implications to medical education. Indian J. Physiol. Pharmacol. 59, 428-35.

De Grutijter, D.N., Van der Kamp, L.J., 2008, Statistical test theory for the behavioral sciences. Chapman&Hall, London.

Demars, C., 2010, Item Response Theory. Oxford Universtiy, New York.

DeVellis, R.F., 2006. Classical test theory. Med. Care. 44, 50-59.

Epstein, R.M., 2007. Assessment in medical education. N. Engl. J. Med. 356, 387-96.
Hingorjo, M.R., Jaleel, F., 2012. Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. J. Pak. Med. Assoc. 62,142-147.
Hintze, J., 2007. NCSS and GESS. Kaysville, Utah.
Kilgour, J.M., Tayyaba, S., 2016. An investigation into the optimal number of distractors in single-best answer exams. Adv. Health Sci. Educ. Theory Pract. 21, 571-585.
Li, J., Liu, H., Liu, H., Feng, T., Cai, Y., 2011. Psychometric assessment of HIV/STI sexual risk scale among MSM: a Rasch model approach. BMC Public Health. 11, 1-8.
Mukherjee, P., Lahiri, S.K., 2015. Analysis of multiple choice questions (MCQs): Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. IOSR J. Dent. Med. Sci. 1, 47-52.
Pallant, J., 2016, Scale development, Rasch analysis and item response theory. Melbourne Australian consortium for Social & Political Research inc (ACSPRI), Melbourne: ACCPRI.
Petrillo, J., Cano, S.J., McLeod, L.D., Coon, C.D., 2015. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patientreported outcome measures: a comparison of worked examples. Value Health. 1, 25-34.
Sim, S.M., Rasiah, R.I., 2006. Relation between item difficulty and discrimination indices in true/false type multiple choice questions of a para-clinical multidisciplinary paper. Ann. Acad.Med. 35, 67-71.
Tarrant, M., Ware, J., Mohammed, A.M., 2009. An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. BMC Med. Educ. 9, 1-8.
Tomak, L., Bek, Y., 2015. Item analysis and evaluation in the examinations in the faculty of medicine at Ondokuz Mayis University. Niger. J. Clin. Pract. 18, 387-394.
Trakman, G.L., Forsyth, A., Hoye, R., Belski, R., 2017. The nutrition for sport knowledge questionnaire (NSKQ): Development and validation using classical test theory and Rasch analysis. J. Int. Soc. Sports Nutr. 3,14-26.
Vegada, B., Shukla, A., Khilnani, A., Charan, J., Desai, C., 2016. Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. Indian J. Pharmacol. 48, 571-575.
Zaman, A., Niwaz, A., Faize, F.A., Dahar, M.A., 2010. Analysis of multiple choice items and the effect ofitems' sequencing on difficulty level in the test of mathematics. Eur. J. Soc. Sci. 17, 61-67.