Sudan Journal of Medical Sciences
Volume 15, Issue no. 2, DOI 10.18502/sjms.v15i2.5503
Production and Hosting by Knowledge E

Research Article

Comparison between Students' Perception toward an examination and item analysis, reliability and validity of the examination

Assad Ali Rezigalla[1], Ali Mohammed Elhassan Eleragi[1], Masoud Ishag Elkhalifa[2], and Ammar Mohammed Ali Mohammed[3]

[1]Department of Basic Medical Sciences, College of Medicine, University of Bisha, Bisha 61922, P.O. Box 551, Saudi Arabia
[2]Department of Medical Education, College of Medicine, University of Bisha, Bisha 61922, P.O. Box 551, Saudi Arabia
[3]Department of Anatomy, faculty of medicine, International university of Africa, Khartoum, Sudan.

Corresponding Author:
Assad Ali Rezigalla;
email: assadkafe@yahoo.com

Received 24 April 2020
Accepted 15 June 2020
Published 30 June 2020

Production and Hosting by Knowledge E

Abstract

Introduction: While student's perception of an exam is a reflection of their feelings toward the exmination items, item analysis is a statistical analysis of their responses to examination items. The study aims to compare students' perception toward the difficulty of an examination with the results of item analysis and examination reliability and validity.

Materials and methods: This is a cross-sectional study conducted in the College of Medicine between January and April 2019. The study uses a structured questionnaire and standardized item analysis of students' examination.

Results: Overall, 80 items were analyzed in this study. Kuder–Richardson Formula 20 of the examination was 0.906. The average difficulty index of the examination items was 69.4 (± 21.86). The response rate of the questionnaire was 88.9% (40/45). Students considered the examination as easy (70.4%). Students' perception toward the difficulty of the individual items shows a moderate positive correlation between easy perception and difficulty index ($r = 0.7033$, $p = 0.00001$), which means there is a tendency for high difficulty index to go with high easy perception (and vice versa). Moderate negative correlations were reported between moderate ($r = -0.2969$, $p = 0.008082$) and difficult ($r = -0.6094$, $p = 0.00001$) perception to individual items' difficulty and difficulty index. A significant moderate positive correlation ($r = 0.615$, $p = 0.00001$) was reported between the difficulty index and items covered within the specific learning outcomes.

Conclusion: Students' perception toward items difficulty is aligned with the standard difficulty index of items. Their perception can support the evidence of examination validity. The constructions of items from the covered outcomes result in an acceptable level of item and examination difficulties.

Keywords: Students' perception, item analysis, assessment, Difficulty Index, internal consistency

How to cite this article: Assad Ali Rezigalla, Ali Mohammed Elhassan Eleragi, Masoud Ishag Elkhalifa, and Ammar Mohammed Ali Mohammed (2020) "Comparison between Students' Perception toward an examination and item analysis, reliability and validity of the examination," *Sudan Journal of Medical Sciences*, vol. 15, issue no. 2, pages 114–123. DOI 10.18502/sjms.v15i2.5503

1. Introduction

Students' perception of an examination is defined as a reflection of their feelings toward the examination items. While item analysis refers to a statistical analysis of students' responses to examination items. The two represent different perspectives about examination items.

The assessment is considered to be valid if it measures what is intended to be measured and reflects the educational contents [1, 2]. Construct validity denotes "a unitary concept, requires multiple lines of evidence, to support the appropriateness and meaningfulness of the specific inferences made from test scores" [3]. The validity and reliability of exmination can be adversely impacted by the mismatch between the level of cognition in the assessment and the educational task [1, 4]. This mismatch can appear in the form of too many easy or difficult items.

Item analysis is used to evaluate the quality of items and consequently helps in the improvement of the assessment process. An assessment can be improved by refining the defective items or deleting the poorly constructed ones from the question bank [5–7]. The parameters of item analysis include the Difficulty Index (DIF) as well as the index of the internal consistency, that is, Kuder–Richardson formula 20 (KR-20). The DIF refers to the percentage of the examinees who answered the item correctly. It ranges from 0 to 100%, with a higher value indicating an easy item index (8). Meanwhile, internal consistency is commonly measured through Cronbach's α (Coefficient alpha) [5, 9, 10]. Coefficient alpha is known to be identical to the Kr-20 in the case of type A MCQs [5, 11, 12]. Different ranges and interpretations of item analysis parameters, as well as internal consistency, have been published in extant literature [9, 12–17].

The College of Medicine, University of Bisha, Saudi Arabia (UBCOM) has adopted innovative student-centered teaching, problem-based learning, an integrated curriculum, community-based teaching, electives with core, and the use of a systematic methods curriculum (SPICE). Problem-based learning is the principle educational strategy in addition to the team-based learning, seminars, case-based learning, and practical. The program offers Bachelor of Medicine, Bachelor of Surgery (MB, BS) following the successful completion of 12 semesters (six years) [18, 19].

Students' perception is widely used and recommended in the field of medical education. Data generated from students' perception can provide valuable information about faculty, the achievement of educational objectives, and instructional methods [20, 21] besides being considered as a reliable and valid indicator of effective teaching [22].

The study aims to compare students' perception toward the difficulty index of an examination with the results of the item analysis and examination reliability and validity.

2. Materials and Methods

2.1. Study area

The study was conducted at UBCOM between January and April 2019.

2.2. Study design

The study desgin is cross sectional study [23].

2.3. Study population

All students registered for the course of principles of human diseases (2016–2017) were included in the study (*n* = 40). The exclusion criteria of the study included students who refused to participate or those who did not fill the questionnaire. The particpating students filled the questionnaire immediately after completing the examination without identification.

2.4. Sample size

The sample size is the total coverage.

2.5. Materials

The study used a standardized item analysis of the final course examination and a questionnaire.

The examination used in this study was from the course of principles of human diseases. It is conducted in semester two of the second year (*n* = 45). The course is integrated and multidisciplinary. The course examination was developed by the course committee using course blueprint and then approved by the students' assessment committee (SAC) of UBCOM. It was comprised of type A MCQs. The number of examination items (*n* = 80) was adjusted according to the course blueprint and the tested domains [24]. Each item is composed of stem and four options, three distractors, and a single best answer. The correct answer is awarded one mark and no marks for blank or wrong selection.

The examination was marked automatically (DataLink 1200 – Apperson system) and double-checked by the examination officer and course coordinator. Standard item analysis was obtained and processed for the study.

The study used a questionnaire to evaluate students' perception of the examination items and standard item analysis of the examination.

The questionnaire was developed to gain a deep understanding of students' perception toward the examination in general and examination items in particular. It was developed by the authors in consultation with medical education experts and statisticians and consisted of two parts. Part consisted of a three-point Likert scale (easy, moderate, and difficult). The mode of covering specific learning outcomes (SLO) from which the items were constructed was assessed through a two-point Likert scale (covered or not covered). The second part encompassed the number of items, their mode of covering the course contents, and the ability of examination to assess students. This part was evaluated through a three-point Likert scale (yes, not sure, and no). The questionnaire was tested through a pilot study. The internal consistency of the questionnaire was 0.79. Data generated from the pilot study were not included in the study.

2.6. Data collection

The questionnaire was distributed to students in the last five min of the examination, and those who finished early were given the questionnaire after they left the examination hall. The data was collected by the authors. All students were informed that participation in the study had no impact on their academic performance in the long or short term.

2.7. Statistical analyses

The data obtained from the questionnaires and the standard item analysis were analyzed using SPSS, version 20 (Armonk, NY: IBM Corp, USA). Descriptive statistics and Pearson correlation coefficient were applied to measure the significance of difference and correlation among different variables. The level of significance was fixed at 95%, Confidence interval and *P*-value < 0.05 was considered as significant.

2.8. Ethical consideration

The study was approved by the research and ethics committee. All students who participated in the study gave a written consent.

3. Results

3.1. Item analysis

The total number of the analyzed items was 80. The average class score was 55.5 (69.38%). Class median was 56.0 (70.0%). KR-20 of the examination was 0.906. Students' passing rate was 32.5% (Passing marks = 60). The average DIF of the examination was 69.4 (±21.86). The exmination items were classified into difficult, moderate (acceptable), and easy (Table 1).

TABLE 1: Classification of examination items according to difficulty index.

Parameters	UBCOM		Pande et al., 2013 [13]		Mitra et al., 2009 [16]	
	Interpretation	%	Interpretation	%	Interpretation	%
DI	Easy > 80	35	Easy (>70)	52.5	Easy (>80)	35
	Moderate (25–80)	62.5	Acceptable (30–70)	42.5	Acceptable (30–80)	60
	Difficult (0–25)	2.5	Difficult (< 30)	5	Difficult (< 30)	5

DI, Difficulty index

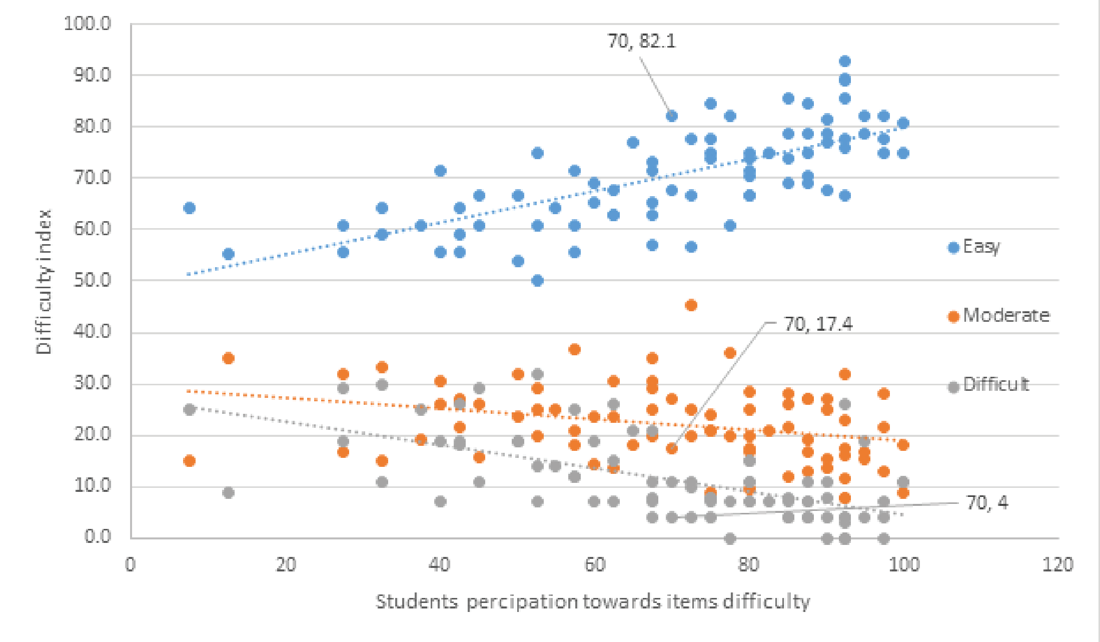

Figure 1: Correlation of student's perception toward item difficulty and the standard difficulty index of items.

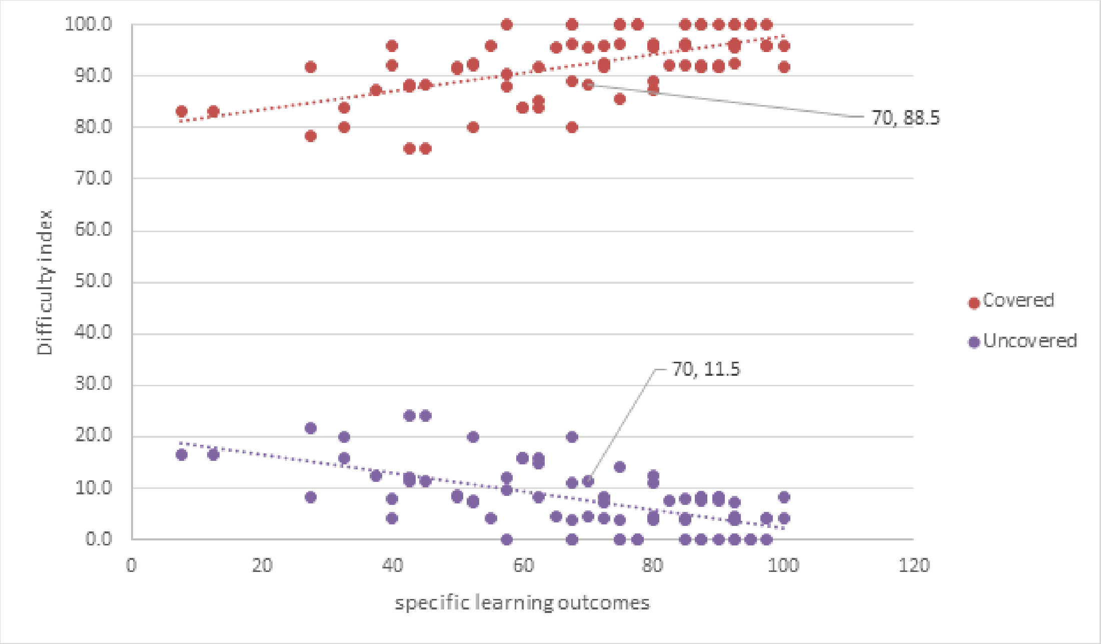

Figure 2: Correlation between the difficulty index of items and the specific learning outcome.

3.2. Students' perception

The response rate of the questionnaire was 72.5% (29/40) and 11 students refused to participate in the study. The averages of students' perception toward items' difficulty were easy (70.4%), moderately difficult (18.5%), and difficult (11.1%).

For 57% of the students, the examination coveredthe entire course. The distribution of examination items across the course content was equal for 64% of the students. While only 38% thought that the examination could assess them, 70% of the students reported that the number of questions was adequate to assess them. The majority of students (92%) believe that the examination items were covered during course instruction.

A moderate positive correlation was reported between easy perception and DIF (r = 0.7033, *p* = 0.00001), which means high DIF is associated with high easy perception. High DIF of an item indicates its easiness. Moderate negative correlation was reported between moderate (r = –0.2969, *p* = 0.008082) and difficult (r = –0.6094, p = 0.00001) students' perception and DIF (Figure 1).

A significant moderate positive correlation (r = 0.615, *p* = 0.00001) was reported between DIF and items from covered SLOs (Figure 2).

4. Discussion

The KR-20 of the examination (reliability coefficient) was 0.906, and the majority of examination items were within the acceptable range of difficulty (62.5%). These findings support the validity of the examination. According to some authors [1, 9, 14, 17, 25], KR-20 value of 0.8 or above is ideal and demonstrate excellent reliability of the examination. It has been reported that the presence of too many easy or difficult items can affect both examination validity and reliability [1, 4].

The average class score and class median were 55.5 and 56.0, respectively. These values suggest that the number of students who performed very well was the same as those with low performance.

Examination set-up, such as the construction of the examination by the expert staff who were involved in teaching and using the blueprint, supports the content form of validity [1, 2, 15, 24]. Also, the presence of an acceptable percentage of students who pass the examination supports the construct form of validity [26].

Students reported that the examination covered the course contents in a well-balanced manner, and the number of items was adequate to assess them. These findings of students' perception support the validity of the examination since an examination is considered valid if it measures what is intended to measure and reflects the educational contents. These findings are supported by the previous works of Carmines *et al.* and Brown *et al.* [15, 27] who reported that the validity is based on the extent to which a measurement reflects the specific intended content.

The average DIF of examination, according to the standard item analysis, was 69.4. The average student's perception of examination difficulty is easy (70.5%) and shows a significant positive correlation (r = 0.7033, p = 0.00001) with DIF. The average examination difficulty is considered good and acceptable according to the college assessment policy and literature [13, 14, 16]. In any examination or test, the average difficulty of items is adjusted according to the required competencies and student-level [1]. The current findings of students' perception toward examination difficulty suggested that they underestimated the examination difficulty. Students commonly underestimate their performance rather than the examination difficulty [1, 28, 29]. Van de Watering reported that students' perception toward examination difficulty differs according to their performance in the examination and students with higher scores underestimate their performance while students with lower scores have more accurate estimations [1]. According to the examination result, the upper students represent 72.5%. However, the class mean and average are relatively similar (55.5 and 56.0, respectively). The result suggested a good student's performance. These findings support the work of Van de Watering [1].

The limitation of the study includes the fewer number of students and the application of the study on one course. The strength of the study is that the test is considered valid and reliable through several pieces of evidence.

5. Conclusion

Students' perception toward items difficulty is aligned with the standard DIF of the examination items. Their perception support the evidence of examination validity. The constructions of items from the covered outcomes result in an acceptable level of item and examination difficulties.

Acknowledgments

The authors would like to thank the students who participated in the study. They are also thankful to Dr Elwathiq Khalid, Dr K. Salih, Dr E. Miskeen, Dr I. Jack, Dr M. Elhassan and all their colleagues. They are especially thankful to Mr. M. K. Abid for the statistical analysis and helpful comments and to the College Dean and administration for their help and allowing the use of facilities.

References

[1] van de Watering, G. (2006). Teachers' and students' perceptions of assessments: a review and a study into the ability and accuracy of estimating the difficulty levels od assessment items. *Educational Research Review*, vol. 2, no. 1, pp. 133–147.

[2] McMillan, J. H. (2012). *SAGE Handbook of Research on Classroom Assessment*. SAGE.

[3] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.

[4] Linn, R. L. (1993). Educational measurement. *American Council on Education Series on Higher Education*. ORYX PR.

[5] Considine, J., Botti, M., and Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*,

vol. 12, no. 1, pp. 19–24.

[6] Abdulghani, H. M., Ahmad, F., Ponnamperuma, G. G., et al. (2014). The relationship between non-functioning distractors and item difficulty of multiple choice questions: a descriptive analysis. *Journal of Health Specialties*, vol. 2, no. 4, p. 148.

[7] Lai, H., Gierl, M. J., Touchie, C., et al. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and Learning in Medicine*, vol. 28, no. 2, pp. 166–173.

[8] Shete, A. N., Kausar, A., Lakhkar, K., et al. (2015). Item analysis: an evaluation of multiple choice questions in physiology examination. *Journal of Contemporary Medical Education*, vol. 3, no. 3, pp. 106–109.

[9] Al-Osail, A. M., Al-Sheikh, M. H., Al-Osail, E. M., et al. (2015). Is Cronbach's alpha sufficient for assessing the reliability of the OSCE for an internal medicine course? *BMC Research Notes*, vol. 8, no. 1, p. 582.

[10] Peterson, R. A. and Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology*, vol. 98, no. 1, p. 194.

[11] Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, vol. 4, no. 10, pp. 20–24.

[12] Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, vol. 78, no. 1, p. 98.

[13] Pande, S. S., Pande, S. R., Parate, V. R., et al. (2013). Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology. *South-East Asian Journal of Medical Education*, vol. 7, no. 1, pp. 45–50.

[14] Bland, J. M. and Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, vol. 314, no. 7080, p. 572.

[15] Carmines, E. G. and Zeller, R. A. (1979). *Reliability and Validity Assessment*. SAGE Publications.

[16] Mitra, N., Nagaraja, H., Ponnudurai, G., et al. (2009). The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1, multidisciplinary summative tests. *IeJSME*, vol. 3, no. 1, pp. 2–7.

[17] Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, vol. 80, no. 1, pp. 99–103.

[18] Ibrahim, M. E. and Al-Shahrani, A. M. (2018). Implementing of a problem-based learning strategy in a Saudi medical school: requisites and challenges. *International Journal of Medical Education*, vol. 9, p. 83.

[19] Ibrahim, M. E., Al-Shahrani, A. M., Abdalla, M. E., et al. (2018). The effectiveness of problem-based learning in Acquisition of Knowledge, soft skills during basic and preclinical sciences: medical Students' points of view. *Acta Informatica Medica*, vol. 26, no. 2, p. 119.

[20] Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, vol. 13, no. 2, pp. 153–166.

[21] Zhao, J. and Gallant, D. J. (2012). Student evaluation of instruction in higher education: exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, vol. 37, no. 2, pp. 227–235.

[22] Marsh, H. W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential baises, and utility. *Journal of Educational Psychology*, vol. 76, no. 5, p. 707.

[23] Rezigalla, A. A. (2020). Observational study designs: synopsis for selecting an appropriate study design. *Cureus*, vol. 12, no. 1.

[24] Abdellatif, H. and Al-Shahrani, A. M. (2019). Effect of blueprinting methods on test difficulty, discrimination, and reliability indices: cross-sectional study in an integrated learning program. *Advances in Medical Education and Practice*, vol. 10, p. 23.

[25] Sullivan, G. M. (2011). *A Primer on the Validity of Assessment Instruments*. Chicago, IL: The Accreditation Council for Graduate Medical Education.

[26] van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *BMJ*, vol. 321, no. 7270, pp. 1217–1219.

[27] Brown, S. and Knight, P. (2012). *Assessing Learners in Higher Education*. Abingdon, United Kingdom: Routledge.

[28] Rezigalla, A. A. (2015). Angoff's method: The impact of raters' selection. *Saudi Journal of Medicine and Medical Sciences*, vol. 3, no. 3, p. 220.

[29] Dochy, F. J. R. C. (1992). Assessment of Prior Knowledge As a Determinant For Future Learning: The Use of Prior Knowledge State Tests and Knowledge Profiles. Centre for Educational Technology and Innovation, Open University.