Biomedical & Pharmacology Journal Vol. **6**(1), 23-27 (2013)

Examining the Multiple Choice Educational Examinations of College Students

MOHAMMAD REZA SHARIF[1], MOHAMMAD HOSSEIN ASADI[2], ALI REZA SHARIF[3] and MANSOUR SAYYAH[4]

[1]Department of Pediatrics, Kashan University of Medical Sciences, Kashan, I.R. Iran.
[2]Faculty Member of Baghiyatollah (a.s.) University of Medical Sciences, Tehran, I.R. Iran.
[3]Department of Infectious Disease, Kashan University of Medical Sciences, Kashan, I.R. Iran.
[4]Faculty member of Trauma Research Center, Kashan University of Medical Sciences, Kashan, I.R. Iran.
*Corresponding author E-mail: mansorsayyah@yahoo.com

(Received: February 05, 2013; Accepted: March 13, 2013)

ABSTRACT

Evaluation of students' academic achievement is an unavoidable task for all faculty members of universities. The majority of academic institutions use multiple choice exams to evaluate their students. The question items that make up these exams need to possess certain psychometric properties to measure the student achievements. The purpose of this research was to examine the item properties of multiple choice exams used in the college of health of Kashan University of Medical Sciences. In this cross sectional descriptive study, a number of multiple choice exams used by the instructors were randomly selected. The difficulty index, discrimination index, and Cronbach alpha was calculated for every exam by using LERTAP 5.0 software prepared by the Assessment Systems Corporation of the United States to perform an item analysis. A total of 1343 multiple choice question items presented to the students at the college of health for evaluating 37 different subjects was analyzed. The results of analysis showed that the mean values of difficulty index and discriminating index were0.68 and 0.20, respectively. The average value of difficulty)0.68(and discrimination index)0.21(were relatively within the recommended values. However, some of the tests had value less than the recommended range and need careful reexamination. It was concluded that a feedback delivered to the instructors may improve these indexes. Further research is needed to examine this subject.

Key words: Item Analysis, Difficulty index, Discrimination index, Cronbach alpha, Multiple choice exams.

INTRODUCTION

Evaluation of students' academic achievement is an unavoidable task for all faculty members of universities. The majority of academic institutions use multiple choice exams to evaluate their students. The question items that make up these exams need to possess certain psychometric properties to measure the student achievements. Many of the institutions that apply tests for evaluation purpose do not carefully inspect the psychometric properties of the instruments such as the reliability of the total score obtained by the test or items constituting the scales or subscales intended within the test that are eventually used to judge their students. There are researches indicating that college-level faculty does not write exams well (Guthrie, 1992; Lederhouse & Lower, 1974; McDougall, 1997), and the side effects of such poor exams are reflected in student's performance by focusing on memorization only (Crooks, 1988; Shifflett, Phibbs, & Sage, 1997).

Different type of exams is used in educational institution including multiple choice and essay. Multiple choice tests are presently the most common and preferred types of tests that are in use in many educational settings. These types of tests are therefore subject to various types of evaluation by computer softwares in order to determine their psychometric properties. Item analysis is a procedure to check the psychometric properties of every item used in a multiple choice test. Item difficulty, item discrimination, and internal consistency are three important concepts in developing a good multiple choice examination. While difficulty index refers to the difficulty of an item for the respondents to correctly identify the correct alternative among the various choices, discrimination index indicates how well the item discriminate the strong students from the weak ones and the internal consistency demonstrate the consistency of response among the items measuring a concept (Nelson, 2001). There are rich sources of references in regard to the significance of these concepts as well as the acceptable values for these indices (Gronlund, 1985, Nelson, 2001, Ebel, & Frisbie, 1986 , Mehrens and Lehmann, 1991, Osterhof, 1990 , Linn & Gronlund, 1995 and Hopkins, Stanley, & Hopkins, 1990 ; Burch, *et, al,* 2008). For instance, Gronlund (1985) suggest item difficulty within the range 0.60 to 0.70 as an acceptable index for multiple choice exams while Nelson (2001) offers the range 0.30 to 0.70 as the desirable item difficulty.

High quality multiple choice items are difficult to construct but easily and reliably scored. Ebel (1986) states that the item difficulty less than 0.20 for an multiple choice exam indicate the item is a poor item and believes that this level should not be less than 0.40 whereas Mehrens (1991) and Osterhof (1990) set a less restricted criterion and suggest the 0.20 to 0.40 as a sufficient level for an item to be included in a multiple choice exams. The internal consistency criterion knows as the Cronbach alpha is another index that is used to judge a multiple choice test. In this regard, different level for different test purposes has been offered. Linn (1995) states that the value for the internal consistency should be between 0.60 to 0.85 while Hopkins (1990) suggests that this value should be 0.90 or higher. Burch (2008) claims that it is necessary to determine reliability of a test for issuing certificate of competency for medical practice. In addition to the criterion described, when designing multiple choice test items, the distracters offered to the test takers are also important (Nelson, 2001).

Considering the importance of such criterion in designing multiple choice examination, this descriptive research was designed to determine the item difficulty, item discrimination, internal consistency used in final examinations of college of health of Kashan University of Medical Sciences.

MATERIAL AND METHODS

This descriptive research was conducted in collaboration with education development center of education undersecretary of the university. All the 37 multiple choice exams given by the instructor at the college of health in education year were randomly selected and used as the data for item analysis by Laboratory of Educational Research1 Test Analysis Package (LERTAP version 5.0). Every exam was item analyzed separately by LERTAP and then the results of analysis of these 37 exams including item difficulty, item discrimination, Cronbach alpha were calculated by SPSS:pc version 14. The results of all analysis were reported in tables prepared by MS-Word.

RESULTS

Overall, 1343 multiple choice exams for different subjects given by 37 instructors were analyzed. Descriptive statistics including mean, standard deviation and other indices are presented in table 1.

Table 1: Descriptive statistic of Item difficulty and Item discrimination of 37 exams

Index	Mean	Standard Deviation	Max.	Min.	Number
Item difficulty	0.68	0.06	1	0	1343
Item discrimination	0.20	0.07	0.96	-0.69	1343

The frequency of item difficulties and discrimination indexes are presented in table 2. The frequency of item difficulties were categorized according to what Nelson (2001) and other literature recommend. The difficulties index categories were set less than 0.30, 0.30 to 0.70 and above 0.70. Table 2 shows that 8 percent of exams had item difficulty less than 0.30, 39.6 percent had difficulty index within the recommended range, that is, 030 to 0.70 and 52.3 percent of the exams had items difficulty over 0.70.

Similar procedure was used to present the discrimination index of the exam. The index was classified into five categories. The base for categorization was negative to zero, more than zero to 0.20, 0.21 to 0.40, and 0.41 to 0.80 and over 0.81, respectively. The result of this analysis is presented in table 3.

Table 2: Frequency distribution of classified difficulty and discrimination index

Difficulty	Range	Frequency	Percent	Cumulative
	> 0.30	108	8	8
	0.30 – 0.70	532	39.7	47.7
	0.71- 1	703	52.3	100
	total	1343	100	-
discrimination				
	Negative to zero	337	25.1	25.1
	0.0 to 0.20	370	27.6	52.6
	0.40 –0.21	377	28.1	80.7
	0.41-0.80	233	17.3	98.1
	1 – 0.81	26	1.9	100
	total	1343	100	-

A visual inspection of the table 2 reveals that the discrimination index for the items with negative or zero were 25.1% , between 0 to 0.20 were 27.6% , between 0.21 to 0.40 were 28.1%, between 0.41 to 0,80 were 17.1% and above 0,81 to 1 were 1.9%, respectively.

The third index calculated for the 37 tests was Cronbach alpha. The average of this index was 0.60. The frequency of this index for the entire test was classified into 5 categories as 0 to 0.20, 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80 and 0.81 and higher. The result of this classification is presented in table 3.

Table 4: Frequency distribution table of classified *cronbach* values

Range	Frequency	Percent	Cunulative
Less than 0.20	4	10.8	10.8
0.20 – 0.40	3	8.1	18.9
0.41 – 0.60	12	32.4	51.4
0.61 – 0.80	12	32.4	83.8
0.81 - 1	6	16.2	100
Total	37	100	-

In table 4, it can be seen that 10.8% of the exams have internal consistency less than 0.10 and 16.2% have consistency index over 0.81 or more. The percent of cronbach alpha ranges 0.20 – 0.40, 0.41–0.60 and 0.61–0.80 are 8.1%, 32.4% and 32.4%, respectively.

DISCUSSION

The results of analysis of data by performing item analysis and other statistical procedures showed that he average value of difficulty and discrimination index were relatively within the recommended values by the experts. The rresults of this research showed that the average of item difficulty for the test conducted at the college of health was 0.68. This value is approximately close to what Gronlund (1985) recommends and is with the range 0.3 to 0.70 that Nelson (2001) suggests. However, 8 percent of tests items showed item difficulties over the 0.70 criterion. This condition indicates that some of the test items need to be re evaluated. When an item difficulty approaches high value such as those found in this research, it implies that either the instructor did not cover the subject matter thoroughly or the student did not show enough interest to study it well. The other index evaluated was the discrimination index. In this research, the average of discrimination index was 0.21. this value is with the range Nelson (2001) has suggested. However, 25 percent of items used in the exams used at the college of health had negative discrimination values or values close to zero. Such item is not making any true contribution to the evaluation that the instructor has in mind. These items need complete revisions since they cannot discriminate the test takers and score as such should not be used as the criterion for making important decisions. An item with negative discrimination index indicates that the strong students were not able to answer the question correctly, while the week students answered the question correctly. A reevaluation of these items may reveal serious flaws in the questions such as typing error or some other critical structure in the test stem. Considering the range 0.60 to 0.85 purposed by Linn (1995) , the average observed in this research was low. However, it should be added that some of the tests had values within this range or even higher. In fact, 32.4% of the tests had internal consistency within the 0.60 to 0.80 and even 16.2% had values over 0.81. The value of internal consistency may change by eliminating test items with low coefficient[11].

However, some of the tests used in this research had item difficulty, discrimination, or reliability index value less than the recommended range. These tests need careful reexamination and may fit for further use. It was concluded that a feedback delivered to the instructors may improve these indexes. Further research is needed to examine other tests that are used regularly at different colleges.

ACKNOWLEDGEMENTS

The researcher would like to express his appreciation to the Educational Development Center for its generous contribution to the conduct of this research.

REFERENCES

1. Angoff, W. H., Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education (1971).
2. Ebel, R.L. & Frisbie, D.A., *Essentials of Educational Measurement (*4th ed.). Sydney: Prentice-Hall of Australia (1986).
3. Gronlund, N.E., *Measurement and evaluation in teaching* (5th ed.). New York: Collier Macmillan Publishers (1985).
4. Nelsonæ L. R. *Item Analysis for Tests and Surveys,* Curtin University of Technology (2001).
5. Kaplan, R.M. & Sacuzzo, D.P., *Psychological testing: principles, applications, and issues.* Pacific Grove, California: Brooks/Cole (1993).
6. Mehrens, W.A. & Lehmann, I.J., *Measurement and evaluation in education and psychology* (4th ed.). London: Holt, Rinehart (1991).
7. Oosterhof, A.C., *Classroom applications of educational measurement*. Columbus, Ohio: Merrill (1990).
8. Linn, R.L. & Gronlund, N.E., *Measurement and assessment in teaching* (7th ed.).

Englewood Cliffs, NJ: Prentice-Hall (1995).

9. Hopkins, K.D., Stanley, J.C., & Hopkins, B.R., Educational and *psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall (1990).
10. Burch, V. C.; Norman, G. R.; Schmidt, H. G.; van der Vleuten, C. P. M., Are Specialist Certification Examinations a Reliable Measure of Physician Competence? *Advances in Health Sciences Education,* **13**(4): 521-533 (2008).
11. Basic Item Analysis for Multiple-Choice Tests. 1995-10-00 Basic Item *Analysis for Multiple-Choice Tests*. ERIC/AE Digest.
12. Guthrie, D. S., "Faculty goals and methods of instruction: Approaches to classroom assessment." In *Assessment and Curriculum Reform*. New Directions for Higher Education No. 80, 69-80. San Francisco: Jossey-Bass (1992).
13. Lederhouse, J. E., & Lower, J. M., Testing College professor's tests. *College Student Journal*, **8**(1): 68-70 (1974).
14. McDougall, D., College faculty's use of objective tests: State-of-the-practice versus state-of-the-art. *Journal of Research and Development in Education*, **30**(3): 183-193 (1997).
15. Crooks, T. J., The impact of classroom evaluation practices on students. *Review of Educational Research,* **58**(4): 438-481 (1988).
16. Shifflett, B., Phibbs, K., & Sage, M., Attitudes toward collegiate level classroom testing. *Educational Research Quarterly,* **21**(1): 15-26 (1997).