African Journal of Physics Vol. 1 (2), 4 pages, October, 2013. Available online at www.internationalscholarsjournals.org © International Scholars Journals

Full Length Research Paper

Evaluating reliability, validity and complex indices for teacher-built physics exam questions in first year high school

Rif Aziz

University of Tehran, Qom College, Iran. E-mail: rifaziz@yahoo.com.

Accepted 17 August, 2013

The purpose of the research is to determine high school teachers' skill rate in designing exam questions in physics subject. The statistical population was all of physics exam shits for two semesters in one school year from which a sample of 364 exam shits was drawn using multistage cluster sampling. Two experts assessed the shits and by using appropriate indices and z-test and chi-squared test, the analysis of the data was done. We found that the designed exams have suitable coefficients of validity and reliability. The level of difficulty of exams was high. No significant relationship was found between male and female teachers in terms of the coefficient of validity and reliability but a significant difference between the difficulty level in male and female teachers was found($P<0.001$). It means that female teachers had designed more difficult questions. We did not find any significant relationship between the teachers' gender and the coefficient of discrimination of the exams.

Key words: Teacher, built exam, content validity, face validity, reliability, coefficient of discrimination, coefficient of difficulty.

INTRODUCTION

Examination and testing is an important part of a teaching-learning process which allows teachers to evaluate their students during and at the end of an educational course. Many teachers dislike preparing and grading exams, and most students dread taking them. Yet tests are powerful educational tools that serve at least four functions. First, tests help you evaluate students and assess whether they are learning what you are expecting them to learn. Second, well-designed tests serve to motivate and help students structure their academic efforts. Crooks (1988), McKeachie (1986) and Wergin (1988) reported that students study in ways that reflect how they think they will be tested. In last 40 years the exams mostly used to evaluate the students have been designed by teachers. Some may have used tests which have been designed by outsider exam designers. These tests have not had enough efficiency (Seif, 2004). Regarding the importance of teacher-designed test in evaluation process of the students, many researches have been done in this area (Lotfabadi, 1997). Anderson and Rogan (2010) presented various tools that instructors could use, both to improve instrument design and validity before presentation to students and to evaluate the reliability and quality of the assessment after students have answered the questions. Kettler and Elliott (2009) conclude with a discussion of precautions, lessons learned and questions generated about the methods used to improve both access and test score validity for the students who are eligible for this new alternate assessment. Wuttiprom et al. (2009) conducted a survey administered to 312 students at the University of Sydney. Using the data from this sample, we performed five statistical tests (item difficulty index, item discrimination index, item point bi-serial coefficient, KR-21 reliability test, and Ferguson's delta) to evaluate the test's reliability and discriminatory power. The result indicates that our survey is a reliable test. This study also provided data from which preliminary findings were drawn on students' understandings of introductory quantum physics concepts. The main point is that questions, which require an understanding of the standard interpretations of quantum physics are more challenging for students than those grouped as non-interpretative. In theory, the best test for a subject is a test that includes all educational

Table 1. Exam characteristics by book chapters

Characteristic	Knowledge		Concept		Application		Total
chapter	Mark	Percent	Mark	Percent	Mark	Percent	
1	42.5	10.1	26.5	6.3	19.5	4.7	88.5
2	32.5	7.8	43.75	10.4	10.5	2.5	86.75
3	39.75	9.4	60.5	14.4	0	0	100.25
4	26	6.2	44.5	10.6	0	0	70.5
5	24.5	5.8	45.25	10.8	4.25	1	74
Total	165.25	39.3	220.5	52.5	34.25	8.2	420

objectives of the course. But if the test is too long, its preparation is impractical. Therefore, instead of including all content and objectives, one may choose some questions which are representative of the whole subject to achieve all objectives. Such a test is said to have content validity (Seif, 2004).

Content validity of a teacher-designed test can be assessed by a sample of the test questions. When a test does not have content validity, two possible outcomes may occur. First, the students can not present the skills that are not included in the test when the need arises. Secondly, instead some unrelated question may be included in the test that is answered wrongly. The important point here is that we should not mistake the face validity with content validity. Basically the face validity is a measure that determines whether a test is measuring logically and whether students think the test questions are appropriate (Lotfabadi, 1997).

Based on what is said, an ideal test in addition to measuring what is supposed to measure must be consistently constant in different times. This characteristic is called reliability. Other measures of an ideal test are difficulty level and discriminant index. The total percent of the individuals who answer the question correctly is known as difficulty coefficient denoted by P (Seif, 2004). The discriminant index is a measure of discrimination between strong and weak groups. In this study, we intend to evaluate the extent of ideal quality measures (validity, reliability) in teacher-designed test for first year high school.

MATERIALS AND METHODS

The statistical population in this study consisted of all physics exam papers for final physics exams in first and second semester for first year of high school in Qom province of Iran of which a sample of 364 was taken. A multistage cluster sampling was used to draw samples. In first stage one of four education districts was chosen and in second stage three schools was randomly selected. In third stage a number of exam papers from each school were selected according to the number of students in each school.

In this study the content validity of the exam questions was assessed in two ways. In the first method we used a two dimensional table. One dimension was educational goals and the second dimension was the content of the course materials (Seif, 2004). The second method applied for assessing content validity was a questionnaire with Likert scale in which two physics education expert evaluated the extent of compatibility of exam questions with course contents. For assessment of face validity of teacher-built exams we used a 12-item questionnaire answered by two physics experts.

Reliability

To assess the reliability of the tests, we needed to use a number of experts to mark the exam papers in order that the marking does not affect the marker's opinion (seif, 2004). In this study, we asked two teachers to mark the exam papers separately and used Kendal agreement coefficient to check the agreement of the two markings.

Difficulty coefficient and discriminant coefficient

Because all of physics exam questions were open questions, we used the following formula for calculating the difficulty coefficient (DifCo).

$$DifCoef_{question(i)} = \frac{M_{S(i)} + M_{W(i)}}{N_B * m_i}$$

where $M_{S(i)}$ = sum of marks for Strong group in question I, $M_{W(i)}$ = sum of marks for Weak group in question I, N_B=number of students in both groups and Mi = total mark of question 1

And the Discriminant Coefficient (DisCo) was calculated based on the following formula (Kiamanesh 2002).

$$DisCoef_{question(i)} = \frac{M_{S(i)} - M_{W(i)}}{n_g * m_i}$$

where $M_{S(i)}$ = sum of marks for Strong group in question I, $M_{W(i)}$= sum of marks for Weak group in question I, n_g=number of students in one group and m_i=total mark of question 1

RESULTS

The percentages of papers were almost equal in terms of students' sex (49% males and 51% females). The characteristics of the exam questions are summarized in Table 1.

Table 1. Chi- square test for comparison of difficulty coefficients between female and male teachers.

Difficulty level	# of questions from female teachers	# of questions from female teachers	Chi-squared value	Degrees of freedom	p-value
0 - 0.2	9	18			
0.21- 0.4	18	30			
0.41 - 0.6	32	81			
0.61 - 0.8	119	25	96.079	4	0.000
0.81 - 1	36	12			

Table 1. Chi-square test for comparison of discriminant coefficients between female and male teachers

Discriminant level	# of questions from female teachers	# of questions from female teachers	Chi-squared value	Degrees of freedom	p-value
0-0.2	32	28			
0.21 - 0.4	74	46			
0.41 - 0.6	50	37			
0.61 - 0.8	31	31	2.902	4	0.574
0.81 - 1	27	24			

Table 2. Chi-square test for comparison of discriminant coefficients between female and male teachers

Discriminant level	# of questions from female teachers	# of questions from female teachers	Chi-squared value	Degrees of freedom	p-value
0 - 0.2	32	28			
0.21 - 0.4	74	46			
0.41 - 0.6	50	37			
0.61 - 0.8	31	31	2.902	4	0.574
0.81 - 1	27	24			

Table 1 shows that almost half of the physics questions were on concept (52.5%) and smaller percentages on knowledge (39.3%) and application (8.2%).There were no questions on analysis, combination and evaluation in the exams.

As stated before, the agreement of teacher's evaluations was calculated using Kendal's agreement coefficient. The value of the coefficient was 0.54 which was significant at p-value of 0.002. The Kendal's agreement coefficient for face validity of the questions based on the evaluation of expert teachers was 0.49 and significant at (p-value<0.006). The reliability coefficient based on markers' evaluations was 0.975 and significant (p<0.003). The minimum and maximum difficulty coefficients estimated were DifCoef (min) = 0.01 and DifCoef (max) = 1 with standard error of 0.20 which indicates that the questions have moderate difficulty level. The minimum and maximum discriminant coefficients were DisCoef (min) = 0 and DisCoef (max) = 1 with standard error of 0.21 indicating that the questions have good discriminant coefficient.

We also found no significant difference for content validity and reliability between female and male teachers. Then we compared the difficulty coefficient and discriminant coefficient between two sexes of teachers. The test results are shown in Tables 2 and 3.

Table 2 shows that there is a significant relationship between difficulty level of the questions and the sex of teachers. Female teachers tend to design more difficult physics questions than males.

Table 3 shows no relationship between the teacher's sex and the discriminant level of the questions.

DISCUSSION AND CONCLUSION

One of the important issues in any teaching and learning system is the quality of the students. There should be some standards for exam questions so that we have the same and high level of quality among all educational

organizations' output. Although the achievement of students in their course of study is important, the performance of teachers is also of great importance. One of the factors in the performance of teachers is good examination and good marking. Exam questions play a vital role in students' achievement. The level of difficulty, discrimination, validity and reliability of exam questions must be ensured in order to have good outputs. In this study, we concluded that some of these factors can differ among different teachers in terms teacher's sex. Female teachers tend to design more difficult questions than males. This may be because of the performance of the female students (Jandaghi, 2007). We also found that a high percentage of exam questions were concentrated on concept (52.4%) and knowledge (39.3%), whereas a small percentage was concentrated on applications. This may be due to the nature of quantitative sciences like physics. These percentages may of course change when the topic of the course changes. In summary, teachers need to be assessed and evaluated during their teaching process to ensure the quality of their performance.

ACKNOWLEDGEMENT

This research was funded by Education and teaching organization of Qom Province of Iran.

REFERENCES

Anderson TR, Rogan JM (2010). Bridging the Educational Research-Teaching Practice Gap: Tools for Evaluating the Quality of Assessment Instruments. Biochem. Mole. Biol. Educ., 38(1): 51-57.

Crooks TJ (1988). The Impact of Classroom Evaluation Practices on Students. Rev. Educ. Res., 58(4): 438-481.

Jandaghi GH (2007). The Relationship between Undergraduate Education System and Postgraduate Achievement in Statistics, Submitted.

Kettler RJ, Elliott S (2009). Modifying Achievement Test items: A Theory-Guided and Data-Based Approach for Better Measurement of What Students with Disabilities Know, Peabody J. Educ., 84(4): 529-551.

Kiamanesh AR (2002). Assessment and Evaluation in Physics, Ministry of Education Publications, Tehran, Iran.

Lotfabadi H (1997). Assessment and Evaluation in Psychological Sciences, Samt Publication, Tehran, Iran.

McKeachie WJ (1986). Teaching Tips (8th ed.), Lexington, Mass: Heath.

Seif AA (2004). Assessment and Evaluation in Education, Doran Publication, Tehran, Iran.

Wergin JF (1988). Basic Issues and Principles in Classroom Assessment, In J. H. McMillan (ed.), Assessing Students' Learning. New Directions for Teaching and Learning, no. 34. San Francisco: Jossey-Bass.

Wuttiprom S, Sharma MD, Johnston ID, Chitaree R, Soankwan C (2009). Development and Use of a Conceptual Survey in Introductory Quantum Physics. Int. J. Sci. Educ., 1(5): 631-654.