VALIDITY, RELIABILITY AND DIFFICULTY INDICES FOR INSTRUCTOR-BUILT EXAM QUESTIONS[1]

Gholamreza JANDAGHI[2]
PhD, Associate Professor Faculty of Management, University of Tehran, Qom Campus, Iran

E-mail: jandaghi@ut.ac.ir

Fatemeh SHATERIAN[3]
MSc, Academic member of Islamic Azad University, Saveh Branch, Iran

E-mail: shaterian@yahoo.com

Abstract: *The purpose of the research is to determine college Instructor's skill rate in designing exam questions in chemistry subject. The statistical population was all of chemistry exam scripts for two semesters in one academic year from which a sample of 364 exam scripts was drawn using multistage cluster sampling. Two experts assessed the scripts 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 instructors in terms of the coefficient of validity and reliability but a significant difference between the difficulty level in male and female instructors was found(P<.001). It means that female instructors had designed more difficult questions. We did not find any significant relationship between the instructors' gender and the coefficient of discrimination of the exams.*

Key words: *instructor-built exam; content validity; face validity; reliability; coefficient of discrimination; coefficient of difficulty*

1. Introduction

Examination and testing is an important part of a teaching-learning process which allows instructors to evaluate their students during and at the end of an educational course. Many instructors 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) report that students study in ways that reflect how they think they will be tested. In last 40 years the most exams used to evaluate the students have been designed by instructors. 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 instructor-designed test in evaluation process of the students, many researches have been done in this area (Lotfabadi 1997). In theory, the best test for a subject is a test that includes all educational 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 instructor-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 they need. Second, instead some unrelated question may be included in the test that are 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 instructor-designed test for first year college.

Materials and methods

The statistical population in this study consisted of all chemistry exam papers for final chemistry exams in first and second semester for first year of college in Qom province of Iran of which a sample of 364 was taken. A twostage cluster sampling was used to draw samples. In first stage three colleges was randomly selected. In second stage a number of exam papers from each college was selected according to the number of students in each college.

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 chemistry education expert evaluated the extent of compatibility of exam questions with course contents. For assessment of face validity of instructor-built exams we used a 12-item questionnaire answered by two chemistry 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 instructors 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 chemistry 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
m_i=total mark of question i

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
m_i=total mark of question i

Results

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

Table 1. Exam characteristics by book chapters

Characteristic chapter	knowledge		concept		application		total
	mark	percent	mark	percent	mark	percent	
1	41.5	11.5	78	21.7	8.25	2.3	127.75
2	36	10	85.5	23.7	2	0.6	123.5
3	30.75	8.5	20.5	5.7	2.25	0.6	53.5
4	32.5	9.1	22.75	6.3	0	0	55.25
total	140.75	39.1	206.75	57.4	12.5	3.5	360

Table1 shows that most chemistry questions were on concept(57.4%) and percentages on knowledge(39.1%) and small percentage on application(3.5%).There were no questions on analysis, combination and evaluation in the exams.

As stated before, the agreement of instructors evaluations was calculated using Kendal's agreement coefficient. The value of the coefficient was 0.49 which was significant at p-value of 0.05. The Kendal's agreement coefficient for face validity of the questions based on the evaluation of expert instructors was 0.42 and significant at p-value<0.05). The reliability coefficient based on markers' evaluations was) 0.971 and significant(p<0.0001). The minimum and maximum difficulty coefficients estimated were DifCoef(min)=0.14 and DifCoef(max)=1 with standard error of 0.16 which indicates that the questions have moderate difficulty level. The minimum and maximum discriminant coefficients were DisCoef(min)=0.07 and DisCoef(max)=0.98 with standard error of 0.20 indicating that the questions have good discriminant coefficient.

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

Table 2. Chi- sqaure test for comparison of difficulty coefficients between female and male instructors

Difficulty level	# of questions from female instructors	# of questions from female instructors	Chi-squared value	Degrees of freedom	p-value
0-0.2	4	7	28.230	4	0.000
0.21-0.4	19	22			
0.41-0.6	21	41			
0.61-0.8	45	28			
0.81-1	60	20			

Table2 shows that there is a significant relationship between diffculty level of the questions and the sex of instructors. Female instructors tend to design more difficult chemistry questions than males.

Table 3. Chi-square test for comparison of discriminant coefficients between female and male instructors

discriminant level	# of questions from female instructors	# of questions from female instructors	Chi-squared value	Degrees of freedom	p-value
0-0.2	23	26	5.212	4	0.266
0.21-0.4	54	37			
0.41-0.6	36	22			
0.61-0.8	17	21			
0.81-1	19	12			

Table 3 shows no relationship between the instructor'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 instructors is also of great importance. One of the factors in the performance of instructors is good examination and good marking. Exam questions plays 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 instructors in terms of instructor's sex. Female instructors tend to design more difficult questions than males. This may be because of the performance of the female students (Jandaghi 2008). We also found that a high percentage of exam questions concentrate on concept(57.4%) and knowledge(39.1%) whereas the small percentages on other characteristics such as applications. This may be because of the nature of chemistry. These percentages may of course change when the topic of the course changes. In summary, instructors need to be assessed and evaluated during their teaching process to ensure the quality of their performance.

References

1. Crooks, T. J. **The Impact of Classroom Evaluation Practices on Students,** Review of Educational Research, 58 (4), 1988, pp. 438-481
2. Jandaghi, Gh. **The Relationship between Undergraduate Education System and Postgraduate Achievement in Statistics,** International Journal of Human Sciences, 5 (1), 2008
3. Kiamanesh, A. R. **Assessment and Evaluation in Physics,** Ministry od Education Publications, Tehran, Iran, 2008
4. Lotfabadi, H. **Assessment and Evaluation in Psychological Sciences,** Samt Publication, Tehran, Iran, 1997
5. McKeachie, W. J. **Teaching Tips** (8th ed.), Lexington, Mass., Heath, 1986
6. Seif, A. A. **Assessment and Evaluation in Education,** Doran Publication, Tehran, Iran, 2004
7. Wergin, J. F. **Basic Issues and Principles in Classroom Assessment,** in McMillan, J. H. (ed.), „Assessing Students' Learning New Directions for Teaching and Learning", no. 34, San Francisco, Jossey-Bass, 1988

[1] **Acknowledgement:**
This research was funded by Education and Teaching Organization of Qom Provice of Iran.

[2] Gholamzreza Jandaghi Graduated from University of Toronto in Canada in 1994.
He has published as an author and co-author over 50 articles in journals and more than 30 articles on national and international conferences. He is also author of 3 books(Descriptive Statistics for Human Sciences, Probability and Statistics for Engineers and Applied Probability and Statistics for Management and Ecomics).
He teaches Mathematcs and Statistics courses and supervises master and Ph.D. students at Qom Campus of University of Tehran.
His current research areas are EFQM , Statistics Education and Simulations.

[3] Fatemeh Shaterian was born in 1978 in Qom Province of Iran. She is holding B.Sc. in Clinical Psychology, M.Sc. in Assessment and Measurment and is doing her Ph.D. in Behavioral Psychology at Islamic Azad University of Iran. Shaterian is now an academic member at Saveh branch of Islamic Azad University. Her research interests include "job satisfaction", "meta analysis" and "validity and reliability".