ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 2, No. 12, pp. 2561-2568, December 2012

doi:10.4304/tpls.2.12.2561-2568

Psychometrically Based Evaluation of Test Items Constructed by the Faculty and the Test-takers' Performances on Those Tests at Islamic Azad University, Dezful, Iran, in Winter Semester 2011-2012

Abdolreza Pazhakh
English Language Department, Dezful Branch, Islamic Azad University, Dezful, Iran
Email: pazhakh@iaud.ac.ir

Mohammad Hoseinpour
Department of Education, Science and Research Branch- Khuzestan, Islamic Azad University, Ahvaz, Iran

***Abstract*—This study aimed to see the extent the faculty members at Islamic Azad University of Dezful observe the psychometric indices of test items in designing, developing, and administering. Having got the formal permission from authorities in 2011-2012, 1000 final examination papers in different fields were selected through random cluster sampling in terms of students' population at various faculties. The items were analyzed both individually and collectively at different levels at: at individual instructors, relevant departments, colleges, and the entire university. The two ways to test the validity and internal consistency analysis were used and the estimated indices were 0.74 and 0.84, respectively. The findings showed that no uniform manner of evaluating students could be found among all those instructors under study. To the researcher, all these discrepancies might result from instructors' lack of familiarity with psychometrics and testing science across the colleges. The findings suggest that instructors' lack of familiarity with psychometric principles and the areas of cognitive expertise in designing and developing appropriate tests have brought unwanted measurement-related issues.**

***Index Terms*—coefficient of difficulty, difficulty level, psychometrics, reliability and validity**

I. Introduction

Regarding the fact that the educational quality development index is impossible unless university faculties familiarize themselves fully with the measurement principles to supervise, assess, and evaluate accurately and precisely. If they can actualize and achieve the educational quality development index in the light of their faith and commitment, relying on their scientific reserves, they can proceed towards nurturing talents not only in terms of scientific framework, course goals and syllabi but also in terms of quantitative and qualitative content.

The general purpose of this research project was to study the extent to which the faculty members, of Islamic Azad University, Dezful Branch, Iran, design and develop standard test items in their final exams in winter semester, 2011. The items were surveyed in terms of psychometric indices such as test facility, test difficulty, test discrimination, test reliability and validity of different types concerning the subjects' performances on those tests. This study also aimed to compare evaluative norms across the six faculties existing in this university. The hierarchy of these active faculties in terms the number of students is as follows: 1.Technical and Engineering, 2. Humanities, 3. Agriculture, 4. Nursing and Midwifery, 5. Architecture and 6. Graduate Studies. Of course, each faculty, in turn, includes different departments with their own variety of disciplines and their own particularities.

Research variables

The independent variables included in this research were those psychometric indices relevant to standard language testing such as test facility, test difficulty, test discrimination, test reliability and validity of different types concerning the subjects' performances on those tests.

The dependent variables included in this research were the extent to which the final exam test items- designed, developed, and administered by the faculty members of this university were based on the psychometric indices. The above-stated variables were studied not only at different department levels but also at various college levels and particularly at the entire university level.

Research questions

1. Do the tests items fall in the standard range of difficulty level?

2. Do the tests items fall in the standard range of discrimination level?
3. Is there any appropriate inter-item correlation among the items of each test?

Research hypotheses

1. The tests items do not fall within the standard range of difficulty level.
2. The tests items do not fall within the standard range of discrimination level.
3. There are no appropriate inter-item correlations among the items of each test.

II. Review of Literature

Testing and assessment are deeply rooted in history of education; they date back to 2500 BC. They have always been used as the instrument to prove the candidates capabilities in China. Even in the ancient Iranian Sassanid era, licensing of physicians to practice was provided on success in passing the test (Kyamanesh, 1992). From early days of religious scholastic schools, it has been customary for scholars with regard to their limited number to present the learned subjects orally before their (Isa Sediq, 1957).

And today with the development of science and large number of learners at different levels of education and conjoining massive growth in the behavioral sciences especially the many theories of psychology all resulted in innovation and emergence of new methods of evaluation. These days, tests are used for different educational purposes. In designing and developing such tests, it sometimes happens that a given test has low psychometric indices or even it may lack them. Such tests not only fail to identify type and level of learning, but they may follow certain obstacles and difficulties (Bachmann, 2005). In effect, the aim of observing the psychometric indices is to reduce the above-stated barriers and to allow us to achieve the desired results of learning how to evaluate properly. These aspects are as follows:

1- Lack of compliance of tests on desirable skills and activities so that after graduation they can apply their acquired subjects to the real situations.

2 - Lack of compliance of tests in terms of the main objectives of the lesson

3- Failure of the test to discriminate between different levels of learning

4- Encouraging the role of surface and mnemonic learning rather than meaningful and deep understanding of materials in terms of cognition

5- Test administration inducing tension, apprehension and anxiety that makes an obstacle for the test taker to have an actual performance on the test.

6- The researcher's main and inner motivation to conduct this experiment was to see if the instructors teaching at this university design and develop their tests in terms of psychometric principles, and if they observe such indices as coefficient of difficulty, discrimination coefficient, reliability and validity.

Sepasi (2005) found that the faculty members did not make use of the same test items in terms of psychometric indices across the type of subject they teach. Qadimi (2009) also showed that the faculty members of the educational sciences and psychology departments proved to have a much broader information and awareness of designing and developing test items. While the faculty members of the department of Basic Sciences showed the least amount of awareness of the testing principles in comparison to others. The experiment done by Pooladi (2001) showed that there is a significant difference among the extent to which Iranian high school teachers are aware of psychometrics and its application in their test designing and development. However, Salimi Zadeh (1998) and Sharifi (1995) argued that the test questions have to be constructed in terms of contentment and construct validities in such a way to represent perfectly the test takers' real performance to anticipate the possible success rate of learners' success and achievement.

On the significance of assessment in educational programs, Bransford, Brownand, and Kooking (2000, p 244) state that "the learning process necessarily involves the evaluation and feedback". Concerning the emphasis laid on the learning approach to assessment, Mc Keachi (2002, p 71) states that assessment does not merely mean the final exam to decide on test-takers final score, but it is also considered as a learning experience for learners. As Gagne (1985, P. 255) claims performance is associated with learning a new skill, it simply confirms the fact that learning has taken place. On the other hand, Zelif and Shawltz (1996, p.87) also criticize the reliable and objective assessment like MC, T/F, or even SCT tests, while confessing the importance of such tests. They claim that such tests are more applicable to assessing low levels of cognitive learning because the test taker selects the correct choice among the given alternatives. Concerning the disadvantages of the MC type of test, Zelif and Schultz continue to claim that objective tests fail to function as an accurate assessment. But the same authors in the field of evaluation strategies held that other varieties of assessment rather than objective ones can contribute greatly to education task, if they are properly designed and developed by instructors themselves.

Bachman (2005) maintained that a construct has to be both theoretically and operationally defined before it is tested or any decision be made on it. This makes the task difficult particularly the multiplicity of terms and tools, used to serve testing, makes assessment problematic and full of drawbacks. In all educational systems generally two types of evaluation are made: formative and summative. While the aim of the former is to supply the demands of universities to determine student eligibility for promotion, the goal of the latter is to enhance controlling restrictions in implementing the educational rules, and that's why it is prioritized (Gage, 1992). Page and Peterson, 2003). But Ebel and Frisby (1991) claim that most teachers lack such abilities at a high level. Mandrake (2000, p 41) acknowledges that valid tests must

have acceptable levels of validity and stability, although the design and construction of such tests with a high level of these two indices and their interpretation will not be an easy task.

Empirical validity or criterion-related validity

Predictive or criterion-related validity implies the precision with which we can predict an individual characteristic or some future behavior in terms of another independent criterion (Bachman, 2002).

Construct validity:

The aim of construct validity is to identify all factors influencing test performance and to determine the degree of influence of each. It concerns with the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs. For Carrol (1987), a construct of mental ability is defined in terms of a particular set of mental tasks that an individual is required to perform on a given test. However, for Cronbach and Meehl (1955) 'a construct is considered to be as a postulated attribute of people, assumed to be reflected in test performance' (p. 283).

Psychometrics and measurement

The aim of norm-referenced tests is to maximize the distinctions among individuals in a given group. Such tests are sometimes called psychometrics tests since most theoretical models of psychometrics, or psychological measurement, are based on the assumption of a normal distribution and maximizing the variations among individuals' scores. Besides, the term 'measurement' is a process of quantifying the characteristics of persons and human psychological phenomena in terms of explicit rules and procedures. Its main purpose is to come to "the individual recognition" of the mental and physical characteristics of the concerned individual. The more systematic and regular the information gathered, the more appropriately, and accurately the educators can tackle their problems by making (Nozari, 1995).

III. Methodology

Corpus

The corpus consisted of 1000 exam papers belonging to 100 instructors of Islamic Azad University of Dezful, Iran. They were randomly selected and were analyzed with respect to the percentage of students' distribution at various colleges. The distribution at each college was as follow: Technical College accounted for 55% (550 papers), Humanities College for 28% (280 papers), College of Agriculture for 4.5% (45 papers), Architecture College 4.5% (45 papers), Nursing and Midwifery College for 4% (40 papers), and College of Graduate Studies Faculty for 4% (40 papers). The collected data were analyzed and reported using such psychometric indices as difficulty level, discrimination level, reliability and validity co-efficiencies, students' distribution profiles and their performance distribution.

The data collection in this study was corpus-based. The corpus was the stratified random-based selection of 1000 exam papers in autumn semester 2011-2012. The collected data as shown in table 1 were analyzed and descriptive statistics such as frequency, percentage, cumulative percentage as well as psychometric indices such as facility, difficulty, discrimination co-efficiencies, reliability, and validity of the relevant tests under study were all estimated.

Instrumentations

The data collection in this study was corpus-based. The corpus was the stratified random-based selection of 1000 exam papers in autumn semester 2011-2012. The collected data as shown in table 1 were analyzed and descriptive statistics such as frequency, percentage, cumulative percentage as well as psychometric indices such as facility, difficulty, discrimination co-efficiencies, reliability, and validity of the relevant tests under study were all estimated.

Procedures

To carry out this research, first a formal permission from the authorities was received to have access to papers. Then 1000 papers of final examination in different fields in second semester of 2011 were selected by random cluster sampling. Regarding the statistical methods used in this research, special properties of each test type from psychometrical point of view were studied. These psychometric indices were difficulty level, discrimination level, reliability, and validity were estimated, respectively, at various levels- at individual, department, faculty, and university level. The two ways to test the validity and internal consistency analysis were used and the estimated indices were 0.66 and 0.74, respectively. The authorities were promised to be provided with the results as confidential documents to provide a solution to the problem and to fix it in the best way possible. All the papers were checked item by item twice by two teams of experts supervised by the researcher. Then the inter-raters' reliability was calculated to be 0.86.

IV. Data Analysis and Results

From this collected and analyzed data, it can be deduced that the classes, taught by the instructors whose papers were studied, ranged from 15 to 60 students. The male gender of the faculty members accounted for 74%, while the female gender for 26% of the sample under study. The highest mean was found to be in the theology department in humanities faculty, while the lowest mean was in the architecture faculty. In fact, in this faculty 52% of instructors used essay type and 48% of them used non-essay types of items. The exam tests were found in form of booklets with the required instructions on test rubrics, the points allotted to each item, time allotted to the test, the negative score had been administered to test-takers. Regarding the guidance, only 21% of papers showed such evidence. 77% of faculty members had followed the university policy regarding the provision of a uniform booklet with the required instructions

on such issues as test rubrics, the allotted points, the allotted time, and the negative score. Regarding the space needed for the questions to be answered, 48% out of those 77% had required students to elaborate on the essay types of questions in specialized space in the booklet, while 24% had required the test takers to put a tick in the appropriate answer sheet box. Although 16% had required students to answer the questions in a distinct paper, 11% had required test-takers to answer in front of the questions. Moreover, the findings showed that 94% of the faculty had predicted enough space for the required answers to the given questions though 4% had made the space a problem for the test takers to answer. Table 1 shows the descriptive statistics of the data as follows.

TABLE 1:
DESCRIPTIVE STATISTICS OF THE PARTICIPANTS IN THE STUDY AND THE DISTRIBUTION OF EXAM PAPERS IN TERMS OF COLLEGES

College	Frequency	Percentage	Cumulative Percentage
Technical	550	0.55	1000
Humanities	280	0.28	450
Agriculture	45	0.4.5	170
Architecture	45	0.4.5	125
Nursing & Midwifery	40	0.4	80
Graduate Studies	40	0.4	40

As the data in table 1 show, the distribution of the papers analyzed in this research are represented in terms of the students' distribution across the colleges. The technical and engineering college and the college of graduate studies stood in extremes accounting for 0.55% and 4% of the students' population, respectively. Of course, the number of students in the college of nursing and midwifery was equal to that of graduate studies, 4%. Other colleges such as humanities, agriculture, and architecture accounted for 28%, 4.5%, and 4.5%, respectively. So from the whole corpus, 1000 exam papers, 550 from technical, 280 from humanities, 45 from agriculture, 45 from architecture, 40 from nursing and midwifery, and 40 from graduate school college were randomly selected and studied.

TABLE 2:
A SAMPLE OF THE RESULTS ANALYSIS OF TEN FACULTY MEMBERS RANDOMLY SELECTED BY SOFTWARE

Q	IF	Id	ID	∑pq	rtt	SE	MM	KR20
1	0.73	0.27	0.12	3.57	0.45	3.22	2.86	0.43
2	0.93	0.07	0.11	3.34	0.42	3.17	2.34	0.41
3	0.63	0.37	0.17	3.97	0.63	1.82	3.71	0.62
4	0.78	0.22	0.10	3.22	0.41	3.21	2.16	0.39
5	0.09	0.91	0.42	5.12	0.82	2.09	5.37	0.81
6	0.31	0.69	0.27	4.36	0.77	2.54	4.43	0.75
7	0.56	0.44	0.21	4.11	0.71	1.96	3.92	0.69
8	0.65	0.35	0.16	3.68	0.57	1.92	3.91	0.58
9	0.52	0.48	0.22	4.23	0.71	2.86	4.11	0.71
10	0.71	0.29	0.13	3.64	0.44	3.11	2.98	0.46

As you can see, table 2 shows the data pertinent to ten faculty members that were randomly selected by the software to show a sample sketch of the results analysis at the entire university level. Regarding the psychometric indices observed in the tests designed by the faculty members under study are shown in table 3. As this table shows from the viewpoint of psychometric level of difficulty, the faculty members at technical college showed the highest index 59%, and those at agriculture college showed the least 29%. Of course, humanities also had the second rand of difficulty 59%; other colleges such as graduate studies, nursing and midwifery and architecture showed 37%, 36%, 34%, respectively. Regarding discrimination level, the faculty members at humanities showed the highest index 34%, and those at Agriculture College showed 21% as the lowest index; other colleges such as nursing and midwifery, graduate studies, architecture, and technical college showed 0.28, 0.27, 0.25 and 0.23, respectively. Regarding the reliability index, humanities and technical colleges showed the extremes, 0.75, and 0.45, respectively. Of course, other colleges such as nursing and midwifery, graduate school, agriculture showed 0.71, 0.61, 0.59, respectively.

TABLE 3:
PSYCHOMETRIC INDICES OF TESTS CONSTRUCTED BY FACULTY MEMBER IN COLLEGES

College	Psychometric indices	Discrimination level	Difficulty level	Reliability Co-efficiencies
Technical & Engineering	Mean	0.23	0.66	0.45
	Mode	0.25	0.67	0.48
	Std.	0.14	0.13	0.27
Humanities	Mean	0.34	0.59	0.75
	Mode	0.37	0.61	0.64
	Std.	0.21	0.16	0.69
Agriculture	Mean	0.21	0.29	0.49
	Mode	0.23	0.31	0.44
	Std.	0.24	0.16	0.19
Architecture	Mean	0.25	0.33	0.61
	Mode	0.25	0.34	0.63
	Std.	0.17	0.13	0.39
Nursing & Midwifery	Mean	0.28	0.36	0.71
	Mode	0.30	0.37	0.60
	Std.	0.14	0.11	0.50
Graduate Studies	Mean	0.27	0.37	0.59
	Mode	0.29	0.38	0.62
	Std.	0.18	0.12	0.44

Concerning Bloom's fourth cognitive domain, analysis, the Architecture manifested the highest index (24%), while Nursing and Midwifery Colleges showed the lowest index (2%). The profiles for the Engineering College, Graduate School and Agriculture were (15%), (11%), (8%), respectively. With regard to Bloom's fifth cognitive domain, synthesis, both Humanities and Architecture colleges showed an index of (2%), the other colleges did not use synthesis level. Regarding Bloom's sixth cognitive domain, evaluation, the Humanities Faculty showed the highest record (10%). The Architecture and Nursing and Midwifery colleges were found to hold the second and third records (3%) and (1%), respectively. However, other colleges showed an index of zero which is worth rethinking (See table 3).According to Bloom (Bloom et.al, 1956), the deeper the levels of cognitive domains used by educators in education, the more successful they get in achieving their educational goals.

TABLE 4:
FINDINGS ON COGNITIVE DOMAINS AT COLLEGE LEVEL

Levels / College	Knowledge	Understanding	Application	Analysis	Synthesis	Evaluation
Technical & Engineering	42%	31%	12%	15%	--	--
Humanities	31%	47%	3%	7%	2%	10%
Agriculture	41%	27%	24%	8%	--	--
Architecture	14%	26%	31%	24%	2%	3%
Nursing & Midwifery	43%	26%	28%	2%	--	1%
Graduate Studies	37%	44%	8%	11%	--	--

Table 4 shows the percentage of colleges using Bloom's cognitive domains. Concerning Bloom's first cognitive domain, knowledge, the Nursing and Midwifery, Engineering, and Agriculture held the highest ranks (43%), (42%), and (41%), respectively. However, the Architecture faculty showed the lowest index (14%). Graduate School and Humanities manifested (31%) and (37%), respectively. Regarding Bloom's second cognitive domain, comprehension, Humanities Faculty and Graduate School held the first and second ranks (47%) and (44%), respectively, and the Nursing and Midwifery and Architecture showed the lowest index (26%). However, Technical and Engineering and Agriculture manifested (31%) and (27%), respectively. With regard to Bloom's third cognitive domain, application, the first record was held by the Architecture (30%), and the Humanities Faculty showed the lowest index (3%). However, other colleges were as follow: Nursing and Midwifery (28%), Agriculture (24%), Engineering (12%), and Graduate School (8%).

Concerning Bloom's fourth cognitive domain, analysis, the Architecture manifested the highest index (24%), while Nursing and Midwifery College showed the lowest index (2%). The profiles for the Engineering College, Graduate School and Agriculture were (15%), (11%), (8%), respectively. With regard to Bloom's fifth cognitive domain, synthesis, both Humanities and Architecture colleges showed an index of (2%), the other colleges did not use synthesis level. Regarding Bloom's sixth cognitive domain, evaluation, the Humanities Faculty showed the highest record (10%). The Architecture and Nursing and Midwifery colleges were found to hold the second and third records (3%) and (1%), respectively. However, other colleges showed an index of zero which is worth rethinking (See table 3). According to Bloom (Bloom et.al, 1956), the deeper the levels of cognitive domains used by educators in education, the more successful they get in achieving their educational goals. As table 4 shows, humanities, architecture colleges could reach the highest cognitive domain with indices of 10%, 3%, respectively. Of course, nursing and midwifery showed an index 1% which does not seem significant. But the descriptive statistics are presented in table 5.

TABLE 5:
MEAN, MODE, STANDARD DEVIATION AND RELIABILITY TESTS CONSTRUCTED BY ALL FACULTY MEMBERS AT ENTIRE UNIVERSITY LEVEL

Statistical indices / Psychometric indices	Mean	Mode	Std.
Difficulty level	35.00%	37.66%	18.00%
Discrimination level	26.33%	28.16%	13.50%
Reliability level	51%	36.83%	41.33%

V. DISCUSSION

Concerning the first hypothesis, the collected data on the resultant of difficulty index of the total tests under study at entire university proved to be 0.47.66 which appears to be a relatively standard level. Of course, this index was not equally the same at various colleges and departments. For instance, the Engineering and Agriculture colleges with the mean of difficulty coefficients of 0.62 and 0.39 were on the extremes. Other colleges showed indices were as follows: Humanities 0.52, Graduate School 0.47, Nursing and Midwifery 0.45, and Architecture 0.41. Only 11% of the Engineering faculty members used items with difficulty index near the standard level, but the rest of the designed items displayed a higher difficulty index than what is termed standard. So the first hypothesis can be cautiously verified.

Regarding the second hypothesis concerning discrimination index, table 3 shows an index of 26.33% which is not a high index. In effect, various colleges displayed different discrimination level: Humanities 34%, Nursing and Midwifery 28%, Graduate Studies 27%, Architecture 25%, Engineering 23%, and Agriculture 21%. On the other hand, as none of the colleges' mean score was less than 20%, it implies that the discrimination coefficient was very low. Regarding the third hypothesis that tests have appropriate internal correlation, table 4 shows a KR20 index of 51% for all items surveyed at entire university level, which is a moderate index. This proves the third hypothesis. As far as face validity of tests is concerned, 45% of test designers had used the optimum space regarding the face validity principles. 34% of test designers had provided enough instruction. Moreover, 11% had clearly stated the time allotment and the point for each item.

Concerning the normal range of difficulty level which is between 0.30 and 0.70 (Farhady, Ja'farpour and Birdjandi, 2002), the technical engineering college and agricultural college accounted for the extreme indices, 0.66 and 0.29, respectively the rest could be terms as normally difficult.

As stated in Bachman (2002), indices beyond normal range are too easy or too difficult, non-standard. Regarding the discrimination level, Ganji and Sales (2004), hold that the higher the index, the better discrimination is made between test-takers. Therefore, a discrimination coefficient of 20% is suggested for teacher-made tests though a few items were not compatible with this criterion.

VI. CONCLUSIONS

We could come to conclusion that the faculty members in the fields of humanities seem to tend more towards objective tests and other varieties of test formats, but the faculty members in other faculties were mostly tended towards essay types of tests and rarely multiple choice types of tests. 77% of faculty members at humanities and only 23% of the faculty members in other colleges had followed the university policy regarding the uniformity of examination procedures like provision of question booklets with the required instructions on such issues as test rubrics, the allotted points, the allotted time, and the scoring procedures. The implication is the faculty members in Humanities College have been more conscious of the psychometric and testing principles, while those of the Agriculture College have had the lowest awareness of the psychometric and testing principles. Comparatively, the faculty members in Nursing and Midwifery College have been much more conscious of psychometric and testing principles than those of Graduate studies. The discrimination coefficient across different colleges was not the same. It ranged between 34% and 21%, Humanities, Agriculture, respectively. Although those colleges showed the least discrimination index, it can be concluded that their coefficients were not identical. The gross mean score was 26.33%.

Furthermore, the internal correlation coefficient was variant across colleges, Humanities (75%), Engineering (45%), Nursing and Midwifery (71%), Architecture (61%), Graduate Studies (59%), and Agriculture (49%). Concerning the levels of cognitive domains, humanities and architecture showed a commonality but using different test formats. For example, the faculty members at engineering college used more essay types of tests which are more appropriate for measuring conceptually more subjective issues. But lack of objectivity, consistency and precision in scoring, as well as lack of sampling from all the course contents are among the drawbacks of such tests of which the faculty have to be aware. Moreover, analyses showed that 32% of the given answers to a certain item in essay types of questions were the same. This means that test takers have depended more on their memories rather than creativity, which is not of the major goals of essay types of tests. So teachers are required to know the limitations of essay tests.

An interesting point was that no test-taker had been faced with any problem where tests had enough and clear instructions on administering and scoring procedures, while considerable problems were observed in cases where tests lacked clear instructions. For example, some test-takers had performed differently on the tests with lack of clear instruction, particularly in essay types of questions. Even 4% of test-takers had used the paper margins with arrows to

show their answers and even some had left some questions unanswered due to space limitation. Moreover, it was found that 65% of the questions had been typed, although 35% were written by hand on one or more sheets of paper. Another important point was that in 2% of the cases misspelled words and ineligibility were found that had confused test-takers. These all imply lack of sufficient awareness of the majority of faculty members of testing and psychometric principles. On the one hand, after applying appropriate teaching methodology and getting feedback from the learners, one of the main concerns on the part of the faculty members is to be concerned about designing and developing standard tests. On the other hand, of students' concerns is to sit for exams and to perform on tests. As Barlo and Canning (2000) put it, the more teachers get familiar with psychometrics and testing principles, the better they can design, develop, administer, analyze and interpret the data collected. In effect, such teachers not only will consider all psychometric indices of test items as essential, but they will also try to tap those higher cognitive domains which are of qualitatively higher and more value than sticking to lower Bloom's cognitive domains such as memory-based knowledge domain (Mehrens and Lehman, 1984).

REFERENCES

[1] Bachman, L. F. (2002). Fundamental considerations in language testing. Oxford: Oxford University Press. PP. 47-115.
[2] Bachman, L. F. (2005). Statistical analyses for language assessment. Oxford: Oxford University Press. PP. 75-117.
[3] Barlow, L. & Canning, C. (2000). Alternative assessment acquisition in the U nited Arab Emirates. (*ERIC Document Reproduction Service No. ED 448599*).
[4] Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., &Krathwohl, D. R. (1956). Taxonomy of educational objectives: the classification of educational goals; Handbook I: *Cognitive Domain New York, Longmans, Green, 1956*.
[5] Bransford, J. D., A. L. Brown, and R. R. Kooking, eds. (2000). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press. http://www.nap.edu/catalog.php?record_id=6160#toc.
[6] Cronbach, L. J., &Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
[7] Ebel, L. E. & Frisbie, D. A. (1991). Essentials of educational measurement. Cliff , NJ: Engelwood, Printic Hall. PP. 86-112.
[8] Farhady, F. Ja'farpour, J. and Birdjandi, P.(2002). Foreign Language Testing. SAMT Publication.
[9] Gagn é, R. M. (1985). The conditions of learning and theory of instruction, (4th ed.). New York: CBS College.
[10] Ganji, H. and Sales, M. (2004). Pschometrics (Theoretical Fundamentals of Psychometric Tests). Savlan Publication. PP. 16-64.
[11] Mehrens, W. A. & Lehman, I. J. (1984). Measurement and evaluation. (3rd ed.). New York: Holt, Inehard and Winston.
[12] Kiyamanesh, A. (1992). Multiple choice items and their washback effects on the performance of test-takers, teaching content and methodology. *Education Quarterly*, 31, 31-34.
[13] McKeachie, W. J. (2002). Teaching tips, (11th ed.). Boston: Houghton Mifflin. PP. 29-32.
[14] Mundrake, G. A. (2000). The evolution of assessment, testing, and evaluation. In: Rucker, J. (Org.). *Assessment in business education, 38*, NBEA Yearbook. Reston: NBEA.
[15] Nozari, A. (1995). Introduction to educational measurement. Tehran Avay-e-nour Publication. PP. 44-52.
[16] Pooladi, H. (2001). A study on elementary teachers' specialized knowledge of psychometrics and testing principles in Qouchan, Islamic Azad University, Tehran Central Branch, Tehran, Iran, MA thesis. PP. 56-57.
[17] Qadimi, M. M. (2009). A study on faculty members' specialized knowledge of psychometrics and testing principles in Islamic Azad University, Region 9. MA thesis.
[18] Salimizadeh, M. K. (1998). Familiarity with Test Item Analysis in National Assessment of Education Organization (NAEO). NAEO Publication. P. 65-80.
[19] Sediq, I. (1957). History of Iranian Culture. Tehran University Publication, 424.
[20] Sepasi, H. (2005). Item analysis on tests designed by faculty members at Chamran University, Iran, in terms of psychometric indices. *Journal of Education and Psychology*, (3), 4, 1-20.
[21] Sharifi, H. (1998). Open-ended questions and their objective scoring procedures. Tehran. National Assessment of Education Organization (NAEO) Publication. PP. 97-107.
[22] Zeliff, N. D., & Schultz, K. A. (1996).Authentic assessment. In: Perreault, H. R. (Org.). *Classroom strategies: the methodology of business education, 34*, NBEA Yearbook, Reston: NBEA. PP. 117-123.

Abdolreza Pazhakh of the Islamic Azad University of Dezful was born in Dezful, Iran, in 1959. He was awarded both his B.A. and MA in TEFL from the state-run University of Shiraz, Iran, in 1989 and 1992, respectively. Later, he went to Esfahan to pursue his Ph.D. in TEFL. He was awarded with his Ph.D. degree in TEFL from the Islamic Azad University, Science and Research Branch-Khorasgan, Esfahan, Iran, in 2002.

As the initiator of the English language departments in the Islamic Azad University, Dezful and MIS branches, he had been the head of both departments for 13 years. He became the president of the Islamic Azad University of Behbahan (IAUB) between 2005 and 2007. After rejoining the IAUD, he was again assigned as the dean of the humanities faculty since 2007. He has published two books "A Semantic Approach to Natural Languages" and "ESP for Midwifery Students" in addition to numerous articles in international journals. He has both supervised more than twenty MA theses and presented numerous articles at international conferences. His teaching domain has included such English major courses as English Teaching Methodology, Language Testing, Applied Linguistics, CA, EA, and ESP.

Dr Pazhakh has been a member of Provincial promotion committee of Khuzestanese faculty members for two years, and he is now both the director of Khuzestan Provincial Recruitment Scientific Teamwork of the faculty members, and one of the tripartite

members of Khuzestan provincial recruitment committee in Iran. He is a reviewer of a few International journals as well as the scientific member of certain local and global conferences.

Mohammad Hoseinpour of the Islamic Azad University, Science and Research Branch-Khuzestan, was born in Masjed-e-Soleyman, Iran, in 1966. He was awarded his BA in educational sciences from Chamran University, Iran, in 1990. He got two MA degrees: one in educational management from the Islamic Azad University, Science and Research Branch-Khorasgan, Esfahan, Iran, and the other in educational research from Chamran University in Ahvaz, Iran, in 2000. He was awarded his Ph.D. in curriculum development from the Islamic Azad University, Science and Research Branch-Tehran, Iran, in 2006.

He has published different articles on the issues of the curricular development in Iranian local journals. As well as, he has presented some articles in Iranian local conferences. Moreover, he has supervised tens of MA theses in his field.

Dr Hoseinpour has been the president of the Islamic Azad University, Behbahan Branch, Behbahan, Iran, between the years 2001 and 2003. He has been has been a member of Provincial promotion committee of Khuzestanese faculty members for two years. He has also been the students' assistant of the Islamic Azad University, Science and Research Branch-Khuzestan, Ahvaz, Iran, in 2008 through 2010. Now, he is the head of management department in graduate school of the Islamic Azad University Science and Research Branch-Khuzestan, Ahvaz, Iran.