Item Analysis: A Veritable Tool for Effective Assessment in Teaching and Learning

Suleiman Sa'adu Matazu[1*]   Elizabeth Julius[2]

1.Department of Science and Vocational Education, Faculty of Education and Extension Services, Usmanu Danfodiyo University, PO box 2346, Sokoto Nigeria

2.Department of Science Education, Faculty of Education, Kebbi State University of Science and Technology Aliero, PO box 1144 Nigeria

Abstract
Effective assessment is central to education enterprise globally and its impact cannot be over emphasized. The paper focuses on item analysis as a veritable tool for ensuring effective assessment in education industry. The paper gives an overview of item analysis segments such as difficulty index, discrimination index and distractor efficiency. Efforts were made to explore these variables, their relevance and applicability in education, particularly teaching and learning processes. While the paper agued on the importance of item analysis in ensuring effective and efficient assessment in teaching and learning in line with global best practices, it was found out that the unconcerned of practitioners in determination of item analysis when constructing questions is hampering the relevance of item analysis, thus, the paper recommends among other things that practitioners should be engaged in addition to vetting question papers, in item analysis to enable them find empirically the suitability of a question or otherwise.
Keywords: Difficulty index, Discrimination index, Distractor efficiency, Effective assessment
DOI: 10.7176/JEP/12-21-04
Publication date: July 31st 2021

1. Introduction

Education has the objective of bringing about desirable changes in students in relation to the needs of the society. As desirous as this change is, there is the need for school system to find out the extent to which this change has taken effect in students. This over time has led to teachers assessing students' performance in the area of set objectives to reconcile taught with learnt. Students' assessment is central to educational system and is considered as one of the most important aspects of teaching and learning process. Appropriate implementation of accurate form of assessment encourages motivation for learning and provides educational feedback to teachers and other stakeholders. Furthermore, students' assessment is an integral part of the teaching and learning process (Popham, 2008; Trice, 2000). Haigh (2003) submitted that assessment is carried out continuously during or end of teaching and learning process. This is so, in order to determine the weak and strong position of the teaching methods or educational program. One of the primary purposes of assessment is to better students' learning and enhance teachers' teaching. Teachers at all levels of education prepare and administer many formal teacher-made tests and examination during the school year (Quaigrain & Arhin, 2017). Thus, this therefore, make test and examinations indispensable tools in the educational enterprise.

Strict adherence to the principles of test/examination construction, administration and analyses as well as reporting is very essential, especially when norm-referenced tests/examination are developed for instructional purposes (Quaigrain & Arhin, 2017). This is because, assessment gives information to students, teachers, parents and stakeholders that will assist in shaping educational decisions. However, it is important to note that the purpose of assessment taken during teaching and learning practice is multifold. It does not only assure students' capability to grasp the knowledge given but also to observe the extent to which teaching strategies are effective.

Among the global best practices in assessment and evaluation today is ensuring the validity and reliability of the instrument of such assessment. One of the sources by which this global practice can be achieved is through conducting an item analysis on the test or examination. Item analysis is a simple yet valuable procedure performs after the construction of the test or examination providing information regarding the reliability and validity of an item/test by calculating difficulty index, discrimination index, distractor efficiency, and their interrelationship (Rao, Kishan, Sajitha, Permi, & Shetty, 2016).

For instance, all Faculties of Education, Institute of Education and Colleges of Education are saddle with the responsibility of training and producing teachers on the basis of assessment. Therefore, a lecturer is placed in a sensitive and central role in the testing and evaluation process. This makes it imperative for the lecturers to be constantly and adequately reminded in the formalized testing and evaluation techniques to ensure sound facilitation of the enormous task that affects his or her profession. This paper, is aimed to improve the expertise of staff and students of Faculties of Education, Institute of Education and Colleges of Education to systematically use standardized and objective evaluation of students' achievement. The paper also tried to revealed the relevance of item analysis of essay and multiple choice' questions designed by members of staff on the reliability and validity

of the tests and examinations in line with global best practices.

2. Theoretical Position

In the standardized and objective evaluation of students' performances, the item analysis is a process in which both students' answers and test questions are examined in order to assess the quality and quantity of the items and the test as a whole. The current paper has attempted to promote valid interpretation of the characteristics, validity and reliability of quality assessment practices. Sireci and Parker, (2006) opined that the crucial issue in preparing test/examination items is to construct good questions. This requires first of all understanding of the assessment, having good skills in writing the items and an excellent knowledge of the content. There are some guidelines supported by experimental or quasi-experimental designs, but these are usually not adhered to. So the results are the preparation and administration of faulty tests (Walsh, 2008; Haladyna, 2004). The quality of a test depends upon each items of a test (Sharma, 2000). Item analysis allows us to observe the item characteristics, and to improve the quality of the test (Gronlund, 1993). According to Lange, Lehmann, & Mehrens, (1967) submitted that item analysis helps in identifying items too difficult or too easy, items not able to differentiate between students who have learned the content and those who have not, or questions that have un-plausible distractors. So, teachers can remove them from the pool of items or change the items or even modify instruction to correct a confusion, misunderstanding about the content or adjust the way they teach. Improving the skills in the test through item analysis can save time and energy on the part of teachers and test designers.

Two theoretical approaches are widely used for item analysis (Lange et al, 1967): -The first theory is the Classical Test Theory (CTT). According to Wahija, Saeed, Usman, Tahira, & Rubab (2018), the Classical Test Theory CTT was founded by Charles Spearman in 1904. The theory utilizes two main statistics: the item facility index (the percentage of students that correctly answered the item) and the Discrimination index (the point-biserial relationship between students' performance on individual item and total test score).

The second theory is the Item Response Theory (IRT) that describes both item statistics and students' ability with the aim of correlation between the score on a single item and overall test performance. The IRT assumes that there is a correlation between the score gained by a candidate for one (measurable) item/test and their overall ability on the latent trait which underlies test performance. The analysis of students' achievement and the item analysis are useful to study the validity and reliability of the tests/examination before and after their administration. Validity concerns the relationship between the indicators (the items) and the indicated variable. A test, a question or an item is reliable when it really measures what the researcher/evaluator (teacher) wants to notice.

3. Concept of Item Analysis

Item analysis is a procedure that is performed after an examination or test is constructed and administered that provides feedback information on the reliability and validity of the examination or test items (Considine, Botti, & Thomas, 2005; Khan, Ishrat, & Khan, 2015). It is basically meant for understanding students' level by evaluating the test questions. It tells how difficult or easy the questions given in an examination or test are (Ebel & Frisbie, 1991; Sim & Rasiah, 2006; Khan, Ishrat, & Khan, 2015). Furthermore, it is a process of determining the quality of a test or instrument by looking at each individual item or question and determine if they are sound. It helps in identifying individual items or questions that are not good enough and whether or not they should be discarded, kept or revised. It also allows the teacher to observe the item characteristics, and to improve the quality of the test (Gronlund, 1998), it allows revision of items too difficult or too easy and items that are not able to differentiate between students who have learned the content and those who have not, or questions that have un-plausible distractors. So, teachers can remove them from the pool of items or change the items or modify instruction to correct a confusing misunderstanding about the content or adjust the way they teach (Siri & Freddano, 2011). Item analysis is evaluated by calculating the difficulty index (DIF), the discrimination index (DI) and distractor efficiency (DE) of the test or examination items.

Wahija, Saeed, Usman, Tahira, & Rubab, (2018) pointed out that in order to enhance the quality of assessment, continuous analysis of students' assessment methodologies should be a key step. The process involves pre-validation and post-validation of the test/examination items, where the post validation refers to item analysis. Item analysis is useful in three aspects. First, is evaluation of difficulty index (DIF) which tells whether a test/examination items given to students is difficult or easy for the students to attempt. Secondly, the assessment of discrimination index (DI) shows the difference between high and low performing students. Thirdly, determining the distractor efficiency (DE) which helps the subject specialist to assess the credibility of incorrect options in the case of a multiple-choice questions (MCQ). This is equally highly needful in educational research where performance test is used as an instrument for data collection.

Anon (2006) submitted that item analysis is of great importance for effective assessment and further helps to achieve the following purposes;

    i.   To identify the concept that needs to be taught again upon discovering that students can't answer some particular questions

ii.   To identify and report the strength and weaknesses of curriculum parts
iii. To give feedback to students regarding their strength and weakness on items assessed.
iv.   To identify questions/items that are content biased

It is important to know that the analysis of items in test/examination for the evaluation of Difficulty Index (DIF) and Discrimination Index (DI) can be utilized in both objective and essay assessment while Distractor Efficiency (DF) is often done to objectives set of questions only.

4. Difficulty Index (DIF) for Objective and Subjective Items/Questions

According to Mok (1995) difficulty index, also known as facility index is one of the ways to determine the difficulty level in examination and test questions by classifying it into three levels namely; easy, moderate and hard. It is important to note that difficulty index of an item in examination or test could be determined either in objective set of questions and/or essay set of questions. Cheang and Hasni (1998), defined difficulty index of objective set of test items as the ratio of the number of students that were able to answer the question correctly to the total number of students that took the examination or test. This implies difficulty index (DIF) represents a percentage of students who correctly answer the questions in a given examination. According to Cheand and Hasni (1998) the formula for calculating the difficulty index of a set of examination or test questions is given by:

$$DIF = \frac{HAG+LAG}{N}$$

Where:
HAG= High Ability Group,
LAG=Low Ability Group
N=Total number of considered students

In the case of scoring points for subjective or essay questions, flexibility is more than in objective questions, and all the marking points are dependent on the lecturer's marking scheme. The value of difficulty index of either objectives or subjective/essay is a number between 0.00 and 1.00 (Wahija et al., 2018). The bigger the value of a difficulty index DIF, the easier the question will be; on the other hand, if the value of the difficulty index is getting smaller, it implies the questions/items are getting harder. The relationship between difficulty index of questions and difficulty level is graphically represented in Figure 1.
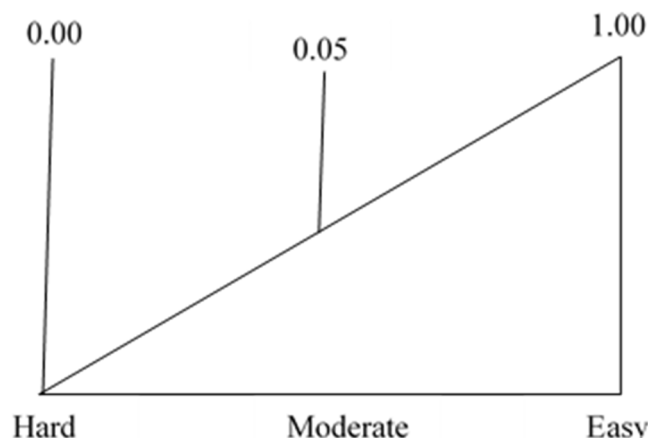
DIF Value

Figure 1. Graphical Relationship between Difficulty Index and Difficulty Level

5. Classification of Difficulty Index (DIF) Difficulty Level and Action to take

There is often the assumption that majority of students can answer set of questions (subjective or objective) presented to them correctly. However, not all students will be able to answer/ respond to hard questions correctly. In order to overcome this problem, Loon (2007) proposed classification of different level for different ranges of the difficulty index and action to be taken bearing in mind the value for any calculated difficulty index (DIF) ranges from 0.00 to 1.00

Table 1. Classification of Difficulty Index Level and Action

Difficulty Index (DIF)	Classification of Difficulty Level	Action
DIF < 0.3	Too hard	Modify Items
0.3 < DIF < 0.79	Moderate	Accept Items
DIF ≥ 0.8	Easy	Modify Items

Source: Loon (2007)

Table 1 shows that when the difficulty index (DIF) is found to be less than 0.3, the questions are considered to be too hard for the students to answer and therefore required modification. However, when the difficulty index (DIF) is found to be between 0. 3 to 0.79, the questions are said to be moderate thus are accepted. Consequently, when the difficulty index (DIF) is found to be 0.8 and above, the questions are believed to be too easy and therefore required modification as well. This assertion is further emphasized by Angel Group Support Center Florida (2014) that DIF value of between 0.3 to 0.79 is acceptable in examination/test questions. Outside the range implies items too difficult or too easy and required modifications.

6. Discrimination Index (DI)

Discrimination index is a way of determining how an examination or test question is able to discriminate between a higher and a lower ability student (Khan, Ishrat, & Khan, 2015; Sarina, Shahida, & Mohd, 2007). It can also be viewed as how the good students are doing versus the poor students on a particular test or examination questions. Furthermore, Gutar, Nazira, & Abdul, (2014) define discrimination index as the capacity of both subjective and objective tests to differentiate the students getting high scores from low performing ones. This simply implies, that the index indicates the ability of a question to discriminate between the higher and lower ability students. The value of discrimination index (DI) ranges between 0.00 to 1.00 (Wajiha, et al 2018).

It is important to note that discrimination index (DI) value could extend from -1.00 to 1.00. The minus value here is called negative discrimination which means more students in the lower group are answering items correctly than students in the higher group (Loon, 2007). The formula for calculating the discrimination index is given by:

$$DI = \frac{2 \times (HAG - LAG)}{N}$$

Where:
HAG = High ability group
LAG = Low ability group
N=Total number of considered students

Table 2. Classification of Discrimination Index Level and Action

Discrimination Index	Action
> 0.36	Excellent
0.25-0.35	Acceptable
0.15 – 0.24	Acceptable with revision
< 0.15	Discard question

Table 2 indicated that a test with discrimination index of above 0.36 excellently discriminates between high ability students and low ability students, a value between 0.25 and 0.35 has an acceptable level of discrimination between high and low ability students. However, test with discrimination index between 0.15-0.24 need to be revised before being accepted as valid and reliable. Lastly value less than 0.15 is not discriminating and therefore that item/s in the test needs to be discarded.

For instance, assuming a lecturer is interested in calculating DIF and DI of 50 students that took essay test with 5 questions. The lecturer first and foremost needs to rank the individual scores of all the students that participated in the test/examination and then divide the scores into high ability group (HAG) and Low Ability Group (LAG). High ability group would consist of top ranked 25% students' scores and low ability group consists of low ranked 25% students' scores. According to Angel Group Support Centre Florida (2014), marks range for establishing DIF and DI in subjective test is for each question, students who achieve 5-3.5 will be considered as correct answer and will be categorized as A. For students that achieve 3.4 to 2.0 marks for each question will be considered as near to correct answer and will be classified as B. Students who have 1.9 to 0.5 marks for each question will be considered as near to incorrect answer and will be classified as C. While students who got 0.4 to 0.0 marks in each question are considered as incorrect answer and categorized as D. So, for each question, number of students obtaining marks in ABCD category should be counted as in the table below;

Table 3. Angel Group Showing Marks of a Test

Marks Range		5.0 – 3.5	3.4 -2.0	1.9 -0. 5	0.4-0.0	Total Number of Considered Students
Designated Sign		A	B	C	D	
HAG Students	Q1	15	9	1	0	25
	Q2	5	14	5	1	25
	Q3	16	9	0	0	25
	Q4	8	10	7	0	25
	Q5	22	3	0	0	25
LAG Students	Q1	1	22	2	0	25
	Q2	0	5	15	5	25
	Q3	2	13	8	2	25
	Q4	0	2	20	3	25
	Q5	4	18	3	0	25
Level of correctness		Correct answers	Near to correct answer	Near to incorrect answer	incorrect answers	

Source: Angel Group Support Center Florida (2014)

Table 5. Showing DIF & DI of each Question in the above Example with Action

Question	Difficulty Index (DIF)		Discrimination Index (DI)	
	Value	Range	Value	Range
Q1	0.32	Acceptable	0.56	Recommended
Q2	0.1	Not Acceptable	0.20	Acceptable
Q3	0.36	Acceptable	0.56	Recommended
Q4	0.16	Not Acceptable	0.32	Recommended
Q5	0.52	Recommended	0.72	Recommended

Source: Angel Group Support Center Florida (2014)

To establish DIF and DI for multiple choice questions same principle and procedures are employed. Thus

$$DIF = \frac{HAG + LAG}{N}$$

Where:

H= Number of students that gave correct option in high score group

L = Number of students that gave correct option in the low score group

Table 1 above is applicable in determination of difficulty index in relation to different level and action to be taken

For $$DI = \frac{2 \times (H\text{-}L)}{N}$$

7. Distractor Efficiency (DE) and its Classification

Distractor Efficiency (DE) is often utilized in an objective/ multiple choice questions (MCQ). Hingorjo and Jaleel, (2012) defined distractor efficiency (DE) as the ability of incorrect answer to distract the students. Objective or MCQ usually contains a question statement which presents a problem situation and usually four options i.e. one correct (key) and three incorrect (distractor) alternatives. According to Wajiha et al (2018) if < 5% of students that attempted the test choose the incorrect answer, then is called non-functioning distractors (NFD). However, distractors selected by > 5% of the students that took the test is called functional distractors (FD). The range of DE is 0-100%. If an item in a multiple choice question (MCQ) has 3 or more NFD, it's DE is 0%. If MCQ has 2 NFD, its DE is 33%. If MCQ has 1 NFD, its DE is 66% and if MCQ has no NFD, its DE is 100%. Thus, the higher the percentage of DE, the better the questions/items. Distractor efficiency (DE) is categorized on the basis of the number of the non-functioning distractors (NFD) present in a given multiple choice questions. The Table below shows the categorization and Distraction Efficiency (DE) and action to be taken.

Table 6. Distractor Efficiency Range

NFD	%	Action
≥ 3	0%	Discard Item
2	33%	Review Item
1	66%	Accept Item
0	100%	Excellent Item

Source Patel (2017)

8. Effective Assessment through Item Analysis

It is important to note that most of the times assessment in school is done with the help of an instrument (examination) either objective or subjective. So also instrument is indispensable for the collection of data when conducting educational research where performance test is often used. Effective assessment could be better achieved if instruments are both pre and post validated through the instrumentally of item analysis. Thus, the relevance of item analysis in ensuring effective assessment include but not limited to the following:

i.   It gives room to identify items/questions in examination that are not necessarily efficient and effective

ii.   It enables evaluator/lecturer to review, replace or even discard ineffective items identified. This will enhance the effectiveness of the test/examination items.

iii.   It enables the evaluator and the researcher to eliminate or reduce drastically the probability of weak students to give correct answer by guessing only.

iv.   In the area of research, item analysis could serve as a lacuna a particular research stand to bridge

v.   Item analysis could help identify area in the curriculum where individual teacher did not cover the subject matter or the students did not show enough interest to the study

9. Conclusion

The paper critically highlighted the concept of item analysis DIF, DI and DE as veritable tools for enhancing effective assessment in teaching and learning. The paper gives a clear insight that average DIF, high discrimination power and large number of functioning distractors is an effective way to improve validity and reliability of a given instrument. This is because, item analysis helps in detecting specific technical flaws in an instrument and furthermore provides information for improvement within and between the items. It also increases the skills of the examiner in item writing.

10. Suggestion/Need for Action

Going by the discussion above, the following suggestions were put forward in line with global best practices for effective assessment of students in the process of teaching and learning:

i.   In addition to vetting question papers, Faculties of Education, Institute of Education and Colleges of Education should be engaged in item analysis to enable them find empirically the suitability of a question or otherwise.

ii.   Faculties of Education, Institute of Education and Colleges of Education should check the oblivious attitude of their staff when it comes to determination of item analysis. The unconcerned of practitioners in determination of item analysis when constructing questions is hampering the relevance of item analysis, hence the paper strongly suggest that practitioners should be engaged in addition to vetting question papers, in item analysis to enable them find empirically the suitability of a question or otherwise.

iii.   Postgraduate students that are using performance test or achievement test as their instrument should be encouraged to run item analysis of their instrument to ensure fitness of individual item as this could be one of the gaps their research stands to bridge.

Reference

Ann, U. (2006) Module of item analysis. Universiti Teknologi Malaysia.

Anon, M. (2006). Module 5 Item Suitability and Testing. HBEF3203: Measurement and Evaluation in Education, p. 90-109. Master's Program Young Teaching of KPM-OUM: Open University of Malaysia.

Angel Group Support Centre Florida (2014). Angel 7.4 Instructor References Manual support centre of Eastern Florida State College.

Cheang, Q. B. & Hasni, Z. (1998) Item Analysis Questions. International Malaysia Education Technology Convention hlm 357-363.

Considine, J., Botti, M. & Thomas, S. (2005). Design, Format, Validity and Reliability of Multiple Choice Questions for use in Nursing Research and Education. *Collegian*, 12(1), pp. 19-24.

Ebel, R.L. & Frisbie, D. A. (1991) *Essentials of Educational Measurement.* In: Ebel RL, Frisbie DA, eds. A Book. 5th ed. Englewood Cliffs, NJ: Prentice-Hall.

Gutar, N., Nazira, I. & Abdul, Q. (2014) Using Item Analysis on Essay Questions given in Summative Examination: Facility Value and Discrimination Index. *International Journal of Research in Medical Sciences,* 3(1), pp. 178-182.

Gronlund, N. E. (1998). Assessment of student achievement. 6th edition. Boston: Allyn and Bacon.

Hingorjo, M. R. & Jaleel, F. (2012) Analysis of MCQ: Difficulty Index and Distractor Efficiency. *Journal of the Pakistan Medical Association*, 62(2), pp. 142-149.

Haigh, J. H. (2003). Classroom Assessment: Principles and Practice for Effective Instruction. 3rd Edition. Boston; Pearson Education

Haladyna, T.M. (2004). MC formats. In: Haladyna TM (ed). Developing and validating multiple-choice test item. 3rd edn. Mahwah, New Jersey: Lawrence Erlbaum Associates; 67-96.

Juridah, J. (2010) Difficulty index of examination and their relation to the achievement of programme outcome. *Procedia Social and Behaviuoral Science* 18, 71-80.

Khan, G. N., Ishrat, N. & Khan, A. Q. (2015) Using item analysis on essay types questions given in summative examination of medical college students facility value, discrimination index. *International Journal of Research in Medical Sciences,* 3, (1), pp. 178-182. DOI: 10.5455/2320-6012.ijrms20150131.

Loon, L. H. (2007) Validity, Reliability and Difficulty indices for instructor-built examination questions. *Journal of applied quantitative method,* 3(2), pp. 151-155.

Lange, A., Lehmann, I.J. & Mehrens, W.A. (1967). Using item analysis to improve tests. Journal of Educational Measurement, 4(2), 65-68; http://www.jstor.org/stable/1434299.

Mok, S.S. (1995). Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions. *Journal of Biomirrow 4(6)*.

Olurinda, D. O. (2017) University and students' preferences for assessment format. *Journal of Science Teachers Association of Nigeria* (STAN) Vol, 52(1) pp. 73-86.

Wajiha, M., Saeed, A., Usman, H., Tahira, Z. & Rubab, B. (2018) Difficulty Index, Discrimination Index and Distractor Efficiency in multiple choice questions. ResarchGate. Anals of PIMS.

Patel, R. M. (2017) Use of Item Analysis to Improve Quality of Multiple Choice Questions. *Journal of Educational Technology*, 4(1), pp. 22-29.

Popham, W. J. (2008). Classroom Assessment: What Teachers Need to Know (5th ed.). Boston: Allyn and Bacon.

Quaigrain, K. & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. Cogent Education, 4(1), pp. 2-11. http://dx.doi.org/10.1080/2331186X.2017.1301013

Rao, C., Kishan, P. H. L., Sajitha, K., Permi, H. & Shetty, J. (2016). Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *Int J Educ Psychol Res* 2:201-204.

Remmers H.H., Gage, N.L. & Rummel, J.I. (1967). A Practical introduction to measurement and evaluation (2nd ed.). Delhi: Universal Book Stall.

Sim, S.M. & Rasiah, R. I. (2006) Relationship between Item Difficulty and Discrimination Indices in True/False - Type Multiple Choice Questions of a Para-clinical Multidisciplinary Paper. *Ann Acad Med Singapore*. 35, pp. 67-71.

Siri, A. & Freddano, M. (2011). The use of item analysis for the improvement of objective examinations. *Procedia Social and Behavioral Sciences* 29, 188 – 197. doi:10.1016/j.sbspro.2011.11.224.

Sarina, S., Shahida, S. & Mohd, N. A. M. L. (2007) Pengkelasan Tahap Kesukaran Soalan Menggunakan Rangkaian Neural. 1st International Malaysian Educational Technology Convection, hlm. pp. 357-363.

Sireci, S. & Parker, P. (2006). Validity on Trial: Psychometric and Legal Conceptualizations of Validity. Article first published online: 15 NOV 2006; DOI: 10.1111/j.1745-3992.2006.00065.

Sharma, S.R. (2000). Modern teaching strategies. New Delhi: Omsons Publications.

Trice, A. D. (2000). A handbook of classroom assessment. New York: Longman.

Walsh, K. (2008). Answering multiple choice questions. BMJ Careers 2005. Available at http://careers.bmj.com/careers/advice/view-article.html?id=891 (accessed on 23 May 2008).