Proceedings of the Seventh National Conference on Private Higher Education Institutions (PHEIs) In Ethiopia

Major Theme: Charting the Roadmap to Private Higher Education in Ethiopia

Organized & Sponsored by St. Mary's University College

August 29, 2009
UN Conference Center
Addis Ababa, Ethiopia

Improving Test Construction Skills through Item Analysis: The Case of St. Mary's University College

By

Endale Ashagre

Abstract

It is widely believed that "assessment drives curriculum". Hence, it can be argued that if the quality of teaching, training, and leaning is to be upgraded, assessment is the obvious staring point. Instructors employ a variety of assessment tools in order to get an overview of students' performance. Among these, multiple choice quizzes and tests are the most reliable and commonly used assessment tools. A dependable multiple choice item is not just coming. It requires a thorough assessment and a continuous refinement. This implies the need to examine the quality, within the context the item is employed, through different mechanisms. One way to deal with this is through item analysis. It is a statistical procedure to analyze test items that combines methods used to evaluate the important characteristics of test items, such as difficulty, discrimination, and distractor analysis of the items in a test. Accordingly, the purpose of this study is to examine sample exam papers being administered by the different departments of St. Mary's University College and subsequently forward appropriate feedback on how to improve multiple choice items. The study employed both qualitative and quantitative analysis. Quantitatively; items were examined using basic item analysis statistics, which includes Item Difficulty and Discrimination index and Point-Biserial Correlation as well as Frequency Counts and Percentage. A total of seven hundred sixty one exam papers, which consisted 234 items, from nine courses, have been considered. To supplement the results obtained from this quantitative date, items were qualitatively reviewed in comparison with the basic guidelines of multiple choice item writing. Results of the study indicated that the majority (83%) of items examined have a moderate difficulty (a difficulty index: $.20<p<.80$) and more than half of the total number of items were found to be good and effective in discriminating (Discrimination index $\geq .20$ (72%) and Point-Biserial Correlation$\geq .20$ (52%)). On the other hand, those poor performing items, which are identified by quantitative analysis, were found violating the basic principles of multiple choice test item writing.

Introduction

Classroom assessment is an integral part of teaching and a central ingredient of every instructional process (Frey et al, 2004). As a result, educators in schools and universities routinely develop and administer classroom tests throughout the academic year to get a quick overview of students' mastery of relevant materials. Nonetheless, while dealing with test, quite repeatedly, a question arises as to what makes a test - a good test? The answer surely has direct implications for instruction and test construction. It becomes, therefore, imminent for teachers to assess the quality of assessment techniques they use and should evaluate them to determine if the scores they yield are valid.

White paper technical report (2006) indicated that instructors who construct their examinations can greatly improve effectiveness of their test items and the validity of test scores if they select and rewrite their items on the basis of item performance data. As Varma S. (2006) pointed out, item analysis is an important step in enhancing quality of tests and/ or exams. According to the Office of Educational Assessment, University of Washington (2005), item analysis is a process that examines student responses to individual test items (questions). This is done in an attempt to assess quality of test items and the test as a whole. It indicates which items are too easy and which ones are too difficult and fail to discriminate good and poor examinees. It suggests why an item has not functioned effectively and how it might be improved. Besides, it is assumed to be valuable in improving items which will be used again in later tests apart from serving as an input in eliminating ambiguous or misleading items in a single test administration. It is also valuable for increasing instructors' skills in test construction, and identifying specific areas of course content which need greater emphasis or clarity.

In light of the above advantages item analysis provides, an attempt has been made to review quality of sample examination papers administered by the various faculties and departments in St. Mary's University College. Results of the study are, thus, compared with standards set by various measurement and evaluation experts. Besides, possible qualitative explanations, as revealed by the item analysis, were forwarded in order to enhance instructors' skill in preparing better test items. Furthermore, the article offers some suggestions for the improvement of multiple-choice tests using "item analysis" statistics.

Study Objectives

The objectives of this study are to:

- evaluate sample exam papers being administered by the different faculties and departments of the University College.
- give appropriate feedback on how to improve test construction based on item statistics.

Methodology

The study employed both quantitative and qualitative approaches to research. Sample exam papers on nine courses having multiple choice items ranging from 10-55 which are corrected and checked were selected from four departments. Accordingly, seven hundred sixty one exam papers were analyzed and the number of question items was 234. Students' response to each multiple choice item were coded and quantitatively analyzed to generate the item analysis statistics. Descriptive statistics like frequency counts; item difficulty and discrimination index formulas as well as percentage were employed. Furthermore, Point-Biserial correlation was employed as a supplement in order to show detailed analysis of item discrimination. Besides, items were reviewed qualitatively parallel with the basic guidelines of item writing in order to indicate common defects in item writing that adversely affect the item analysis statistics.

Literature Review

Measurement and evaluation experts define item analysis in almost the same manner. But the following definition which is forwarded by Varma (2006) seems comprehensive in addressing both the qualitative and quantitative aspects of the matter:

> *Item analysis is a method of reviewing items on a test, both qualitatively and statistically, to ensure that they all meet minimum quality-control criteria. The former uses the expertise of content experts and test review board. Such qualitative review is essentially useful during item development when no data are available for quantitative analysis. A statistical analysis, such as item analysis, is conducted after items have been administered and real-world data are made available for analysis.*

Item analysis assumes to provide some useful statistics which helps evaluate effectiveness of each item. According to Professional Testing Inc. (2006) the two most common statistics reported in item analysis are item difficulty, which is a measure of the proportion of examinees who responded to an item correctly, and item discrimination, which is a measure of how well the items discriminate between examinees who are knowledgeable in the content area and those who are not. An additional analysis that is often reported is the distracters analysis. It provides a measure of how well each of the incorrect options contribute to the quality of a multiple choice item.

Item Difficulty (P-value)

According to the Professional Testing Inc.'s report (2006), item difficulty index is one of the most useful and most frequently reported item analysis statistics. It is the proportion of respondents selecting the right answer to that question. It is a measure of how difficult the question was to answer. As Varma (2006) pointed out items should have indices of difficulty which is no less than .20 and no greater than .80. It is particularly desirable to have most items in the .30 to .50 range of difficulty. Hence, P-values above 0.80 are very easy items and should not be reused again for subsequent tests. If almost all of the students can get the item correct, it is a concept probably not worth testing. Whereas, P-values below 0.20 are very difficult items and should be reviewed for possible confusing language, removed from subsequent tests, and/or highlighted for an area for re-instruction. However, the office of Educational Assessment, University of Washington (2005), indicated that extremely difficult or easy items will have low ability to discriminate but such items are often needed to adequately cover sample course content and objectives. According to Patock J. (2002), the range of item difficulties on a good test depends on what one wishes to know. If the purpose of a test is to determine whether students have mastered a topic area or not, high difficulty values should be expected. If the purpose of a test is to discriminate between different levels of achievement, items with difficulty values between 0.3 and 0.7 are most effective. The optimal item difficulty depends on the question type and on the number of possible distracters. DIIA (2003) indicated the following values as ideal difficulty for various multiple-choice question formats.

FORMATS	**IDEAL DIFFICULTY**
Five-response multiple-choice	.60
Four-response multiple-choice	.62

Three-response multiple-choice	.66
True-false (two-response multiple-choice)	.75

Discrimination Index (DI)

Another equally important item statistics which aid in evaluating effectiveness of an item is item discrimination. Measurement and evaluation experts recommend two ways of calculating item discrimination. One of these is discrimination index. According to Michigan State University, Academic Technology Services (2004), DI is simply the difference between the percentage of high achieving students who got an item right and the percentage of low achieving students who got the item right. As Ballantyne C. (1998) asserted, the discrimination index is affected by the difficulty of an item, because by definition, if an item is very easy everyone tends to get it right and it does not discriminate. Likewise, if it is very difficult everyone tends to get it wrong. It is important to have such items in a test because they help define the range of difficulty of concepts assessed. According to the work of an anonymous author published in the *Journal of Chemical Education* (1980)**,** there are two factors that affect the ability of an exam to discriminate between levels of student ability: (1) the quality of individual test items, and (2) the number of test items.

DIIA (2003) indicated the following discrimination index in order to evaluate the ability of item to discriminate between the upper and lower group of students.

Discrimination value	**Item quality**
0.40 or higher	very good items
0.30 to 0.39	good items
0.20 to 0.29	fairly good items
0.19 or less	poor items

In addition to looking at how the candidates in the upper 27% group performed on a given item in comparison to the candidate in the lower 27% group , another way to assess the discriminability of the item is to look at Point-Biserial Correlation. According to Varma (2006), the point-Biserial correlation is the correlation between the right/wrong scores that students receive on a given item and the total scores that students receive when summing up their scores across the remaining items. Its value ranges from -1 to +1. A large positive point-Biserial correlation indicates that students with high scores on the overall test are getting the item right (which we would expect) and that students with low scores on the overall tests are getting the item wrong (which we would also expect). A low point-Biserial correlation implies that students who get the item correct tend to perform poorly on the overall test (which would indicate an anomaly) and that students who get the item wrong tend to do well on the test (also an anomaly). Debourgh G. (2001) described the following point. Biserial correlation results as a reference point so as to make judgments about the quality of items.

Point- Biserial Correlation (PBC)	**Item quality**
.30 And above	Very good
.20 to .29	Reasonably good
.10 to .19	Marginal, usually needs improvement
.00 to .09	Poor, to be rejected or revised

Varma (2006) also indicated that negative point-Biserial correlation coefficient for the keyed response is an indication that the item is problematic. The problem may simply be that the item has been miskeyed, or the item may be ambiguous, confusing, or malfunctioning for some other reason. The greater the number of candidates in the upper group who correctly answer the item, the higher the point-Biserial coefficient will be. The better items will be those which are answered correctly by all of the candidates in the upper group, and none of the candidates in the lower group**.**

According to the work of an anonymous author (2007), there is an interaction between item discrimination and item difficulty, and one should be aware of two principles:

1. Very easy or very difficult test items have little discrimination
2. Items of moderate difficulty (60% to 80% answering correctly) generally are more discriminating.

Distracters Analysis

Zurawisky (1998) indicated that item distracter analysis examines the percentage of examination which selects each incorrect alternative, to determine whether the distracters are functioning as intended. Distracters that are selected by a few or no students should be removed or replaced. Varma (2006) asserted that distracters should appeal to low scorers who have not mastered the material; whereas high scorers should infrequently select the distracters. Reviewing the options can reveal potential errors of judgment and inadequate performance of distracters. According to Debougrgh (2001), a perfect test item would have 2 characteristics:

1. Everyone who knows the item gets it right
2. People who do not know the item will have responses equally distributed across the wrong answers. It is not desirable to have one of the distracters chosen more often than the correct answer. This result indicates a potential problem with the question. The distracter may be too similar to the correct answer and/or there may be something in either the stem or the alternatives that is misleading.

A Caution in Interpreting Item Analysis Results

According to the Office of Educational Assessment, University of Washington (2005), each of the various item statistics provides information which can be used to improve individual test items and increase the quality of the test as a whole. Such statistics must always be interpreted in the context of the type of test given and the individuals being tested. Mehrens and Lehmann (1973), provide the following set of cautions in using item analysis results (Cited in the Office of Educational Assessment, University of Washington ,2005),

1. Item analysis data are not synonymous with item validity. An external criterion is required to accurately judge the validity of test items. By using the internal criterion of total test score, item analyses reflect internal consistency of items rather than validity.

2. The discrimination index is not always a measure of item quality. There is a variety of reasons an item may have low discriminating power: (a) Extremely difficult or easy items will have low ability to discriminate but such items are often needed to adequately sample course content and objectives; (b) An item may show low discrimination if the test measures many different content areas and cognitive skills. For example, if the majority of the test measures "knowledge of facts," then an item assessing "ability to apply principles" may have a low correlation with total test score, yet both types of items are needed to measure attainment of course objectives.

2. Item analysis data are tentative. Such data are influenced by the type and number of students being tested, instructional procedures employed, and chance errors. If repeated use of items is possible, statistics should be recorded for each administration of each item.

Results and Discussion

A total of 234 multiple choice items drawn from nine different courses were analyzed to generate the item statistical data. Besides, as it is indicated earlier a total of 761 sample exam papers were included in the analysis. Accordingly, the results obtained are presented as follows.

Item Difficulty

In order to asses the difficulty level of each multiple choice item, an index of difficulty was used. The following table summarizes the results obtained from the analysis. The table shows the percentage of items which fall into three groups; namely, difficult, moderately difficult and easy.

Table 1: Item Difficulty

No.	Difficulty index	Frequency	Percentage
1	≤ .20	19	8.11
2	.20 < p < .80	193	82.48
3	≥ .80	22	9.40
	Total	**234**	**100**

Among the total number of items being analyzed, the majority (82.48%) fall within what is assumed appropriate (moderate) difficulty level. However, some of the items were found to be easy (9.40%) and difficult (8.11%). A few number of easy items, particularly at the beginning of the exam are recommended as office of Educational Assessment, University of Washington (2005) indicated. But at the same time, those difficult items, which students couldn't respond to correctly, need further review for possible problems**.** Such items as Varma (2006) indicated may have problems with regard to possible confusing language and/or highlighted for an area of re-instruction.

Item Discrimination

Items were analyzed in order to determine their effectiveness in discriminating between the upper and lower group of students. Accordingly, the following table presents the results obtained.

Table 2: Item Discrimination

No.	Discrimination index		Frequency	Percentage
1	.40 and above		89	38.03
2	.30 to .39		39	16.67
3	.20 to .29		41	17.52
4	0.19 or less		51	21.79
5	Negative		14	5.98
		Total	**234**	**100**

The majority of the items (38.03%) were excellent items as DIIA (2003) indicated in discriminating between students of upper and lower groups. 16.67% of the items were, however, within the range of what is assumed to be very good in discriminating. Still 17.52% were fairly good in discriminating. The remaining 21.79% and 5.98% were poor items and items with negative discrimination, respectively. This result of the study implies that the majority of the items were effective in discriminating between the upper and lower group of students. In the mean time, there are some items, particularly those of poor and with negative discrimination, which need due attention for further improvement and revision.

Point- Biserial Correlation (PB)

Unlike the previous one, the Point-Biserial Correlation reviews the discriminating power of each multiple choice rather than dwelling on group performance. It indicates a one to one correlation between each item and students' overall performance on that particular exam. Hence, the following table indicates PB Correlation results of the whole items.

Table 3: Point - Biserial Correlation

No.	PB index		Frequency	Percentage
1	.30 and above		66	28.2
2	.20 to .29		54	23.07
3	.09 to .19		53	22.65
4	.00 to .09		31	13.25
5	Negative		30	12.8
		Total	**234**	**100**

Results of the analysis indicate that quite significant number of items were assumed to be very good (28.2%) and good (23.07%) in discrimination. Whereas, 22.65% and 13.25% were marginal and poor items, respectively. The remaining 12.8% of the items have a negative correlation which implies that these items were problematic. This indicates the fact that instructors should look into poor performing items in order to make some improvement and enhance their test item writing skill.

Alternative (Distracter) Analysis

This analysis provides opportunity to study responses students make to each alternative of an item. The efficiency of alternatives can be judged by looking at the data presented on the following three tables (Table 4, Table 5 and Table 6). Results show the number and proportion of students in the lower, middle and upper group who selected the correct answer as well as the number of students choosing each alternative. However, all the tables indicate situations where distracters could not function properly when the item is easy, difficult or confusing.

Table 4: Alternative Analysis for Item 3 (Principle of Marketing)

Alternative Analysis for Item 3 DI = 0.23						
Answer key: A						
Observations						
Alt	Lower	Middle	Upper	All	Difficulty	Point Biserial
A	**46**	**78**	**70**	**194**	**.88**	**.17**
B	9	3	0	12		
C	4	0	0	4		
D	4	2	3	9		
E	0	1	0	1		
Total	**63**	**84**	**73**	**220**		

Item 3 (Table 4) is the easiest item in the test as 88 percent of the students have correctly answered it. All the distracters do not appear to be serving any function as only few students have selected them. This item could be the first on the test.

Table 5: Alternative Analysis for Item 15 (Production and Operation Mgt.)

Alternative Analysis for Item 15 DI=.08						
Answer key: D						
Observations						
Alt	Lower	Middle	Upper	All	Difficulty	Point Biserial
A	27	31	17	75		
B	5	8	4	17		
C	4	10	2	16		
D	**6**	**13**	**10**	**29**	**.16**	**.011**
E	6	20	15	41		
Total	**48**	**82**	**48**	**178**		

Item 15 is a difficult item. Alternative 'A' is a possible alternative answer and should be examined to ensure that it is not, in fact, a correct answer, especially as a number of students in the upper group have selected it. Besides, the number of students who selected alternative 'E' significantly exceeds those who selected the correct answer. This also indicates another problem with the item in general and distracters in particular. Alternatives ‘B’ and ‘C’ are weak and need to be revised. Very few students are 'distracted' by these alternatives.

Table 6: Alternative Analysis for Item 4 (Production and Operational Mgt.)

Alternative Analysis for Item 4 DI = 0.06						
Answer key: D						
Observations						
Alt	Lower	Middle	Upper	All	Difficulty	Point Biserial
A	6	14	6	26		
B	0	2	0	2		
C	29	39	10	78		
D	**3**	**3**	**5**	**11**	**.06**	**-.051**
E	10	24	27	61		
Total	**48**	**82**	**48**	**178**		

Item 4 is one of the most difficult items in the test and it discriminates negatively (Point-Biserial). Few students have selected the correct answer, ‘D’. The majority of students have chosen alternative ‘A’, ‘C’ and ‘E’. These distracters need to be examined to ensure they are not ambiguous, or, in fact, that all are not the correct answer. Item 4 needs to be revised or discarded.

Implication of Item Analysis Statistics

Based on the item analysis statistics, qualitative review was made to analyze some of the items that were found to be poor performing as revealed by the item statistics. As indicated in the literature, one of the important factors which affect item discrimination and difficulty is quality of individual test items. Such quality can be reviewed along with the basic guidelines of writing test items. Accordingly, it was observed that there were defects in writing test items which account for the results obtained through item analysis results. As Millman & Greene (1993) outlined, a quality teacher-made test should follow valid item-writing rules (Cited in Cheng &

Bucat, 2002). To increase the validity of teacher-made tests, many item-writing rules-of-thumb are made available in the literature.

One of the common defects observed was the use of brief and meaningless stem. According to Cheng & Bucat (2002), a common fault in MC (multiple choice) item writing is to have a brief, meaningless stem with problem definition revealed in the options. In such cases, it can be difficult to see the intent of the item after reading the stem. To write a focused item, it should include the central idea in the stem instead of the options.

The use of "None of the above" and/or "All of the above" consistently as an option is the most frequent problem identified. Cheng & Bucat (2002) indicated that the use of *none of the above* and/or *all of the above* as options in multiple choice items is tempting to many teachers because they appear to fit easily into many items. Furthermore, Rodrigueze (1997) indicated, the problem with *all of the above* as an option is that it makes the item too easy. If students can recognize at least one incorrect option, they can eliminate it as a viable option. On the other hand, if they can recognize at least two correct options, then they know that it is the correct answer. The same thing applies for "None of the above"

The use of blank space at the beginning of a multiple choice item is another problems observed with regard to item writing. As Cheng & Bucat (2002) stated, measurement specialists have advised not to use the completion format because a student has to retain the stem in short term memory while completing the stem with each option. Test anxiety is even higher if the student is not a native English speaker.

Another problem which was observed and significantly contributes for an item to become easy is the use of clue while writing items. Debougrgh G. (2001) asserted that items written with clues or which indicate verbal association with the correct answer are likely to be easy items. This is because students can easily eliminate the remaining options.

The use of relatively lengthened options is still another problem. As Rodrigueze (1997) mentioned teachers are mostly unaware of this item writing principles. Besides Chase (1964) indicated that the longer options tend to result in higher response rate and students can easily pick the option.

The use of negatively stated item is one among those problems observed. As Cheng & Bucat (2002) indicated, most students have difficulty in understanding the meaning of negatively phrased items. They often read through the negative terms such as not, no, and least and forget to reverse the logic of the relation being tested

Conclusions

Item analysis is a completely futile activity unless the results help instructors improve their assessment practices and item writers improve their test construction. Careful consideration of the results of item analysis can lead to significant improvements in the quality of exams written by an instructor.

- An item must be of appropriate difficulty for students to whom it is administered. Items that virtually everyone gets right are useless for discriminating among students and should be replaced by more difficult items. The same is true of those difficult items.
- An item should discriminate between upper and lower groups. As a general rule, DI or PBC greater than 0.20 is desirable. A negative discrimination

both in discrimination index and Point-Biserial Correlation is undesirable as they indicate a potential problem in the items.

- All of the incorrect options, or distracters, should actually be distracting. Preferably, each distracter should be selected by a greater proportion of the lower group than of the upper group. Distracters that are not chosen by any examinees should be replaced or eliminated. They are not contributing to the test's ability to discriminate the top, average and low achievers.

Recommendation

Based on the findings obtained and the conclusion reached, the following suggestions are forwarded:

- Taking the multiple advantages and importance of item analysis, the Academic Development and Resource Center at SMUC has to promote and conduct an extensive study on the area so as to enhance instructors' skill in preparing better test items and improve the assessment practices.
- *Item* banking is an essential tool for the *development* of valid and reliable *exams*. It is a collection of test *items* that can be readily accessed for future use. Hence, the writer strongly suggests that the two things (Item analysis and Item bank development) should go hand in hand so that items of better quality can be maintained and reused again in the future.
- Study on item analysis is contextual. It requires extensive study at different level on continuous basis. Hence, the researcher recommends the need to expand and replicate such studies across different academic institutions in order to secure and establish quality assessment strategy.

References

Academic Technology Service, Michigan State University (2002). *Interpreting the index of Discrimination*. Retrieved on January, 2009 from http://www.goodthinking in nursing.com/pdf/test_scoring_and_analysis.pdf.

Ballantyne, C. (2004). *Multiple Choice Tests: Test scoring and analysis.* Retrieved on January, 2009 from http://www.tlc.murdoch.edu.au/eddev/evaluation/mcq/score.html.

Chase, C. I. (1964). Relative length of option and response set in multiple choice items *Educational and Psychological Measurement*, 24(4), 861-866. Retrieved on January, 2009 from http://epm.sagepub.com/cgi/pdf_extract/24/4/861

Cheung, D., & Bucat, R. (2002). *How can we construct good multiple-choice items?* Retrieved on October, 2008 fromhttp://www3.fed.cuhk.edu.hk/chemistry/files/constructMC.pdf

DeBourgh, G. (2001). *Test scoring and Analysis*. Retrieved on August, 2008 from http://www.goodthinkinginnursing.com/pdf/test_scoring_and_analysis.pdf.

DIIA (2003). *Test item analysis & decision making*. Retrieved 0n August, 2008 from http://www.utexas.edu/academic/mec/research/pdf/itemanalysishandout.pdf

Frey, B. et.al (2004). *Item-Writing Rules: Collective Wisdom.* Retrieved 0n February, 2009 from http://people.ku.edu/~bfrey/itemwritingrules.pdf .

Office of Educational Assessment, University of Washington (2005). *Item analysis.* Retrieved on August, 2008 from http://www.washington.edu/oea/pdfs/resources/item_analysis.pdf

Patock, J. (2004). *A Guide to Interpreting the Item Analysis Report*. University Testing Services, Arizona State University. Retrieved 0n September, 2008 from http://www.asu.edu/uts/pdf/InterpIAS.pdf.

Professional Testing Inc. (2006). *Conduct the Item Analysis*. Retrieved 0n September, 2008 from http://www.proftesting.com/test_topics/steps_9. http://www.goodthinkinginnursing.com/pdf/test_scoring_and_analysis.pdf

Rodriguez, M. C. (1997). *The art & science of item writing: A meta-analysis of multiple- choice item format effects*. Retrieved 0n September, 2008 from http://www.edmeasurement.net/aera/papers/artandscience.pdf

Varma, S. (2006). *Preliminary item statistics using point-biserial correlation and P-values.* Retrieved on September, 2008 from www.eddata.com

White Paper Technical Report (2006).*Improve your written tests using item analysis*. retrieved on September, 2008 from ftp://ftp.spss.com/pub/web/wp/IAWP-*02001.pdf*

Zurawski, R. (1998). *Making the most of exams: Procedures for Item analysis.* The National Teaching and Learning Forum. 7(6), Retrieved on January, 2009 from http://www.ntlf.com/html/pi/9811/v7n6smpl.pdf

------------------- (1980). Statistical analysis of multiple choice exams**.** *Journal of Chemical Education. 57*, 188-190. Retrieved on March, 2009 from http://chemed.chem.purdue.edu/chemed/stats.html

------------------ (2007). *Item Analysis*. Retrieved on March, 2009 from http://ese.escambia.k12.fl.us/eval/EOC_Exams/Item%20Analysis%20Hints.pdf