INTERNATIONAL JOURNAL OF DISTANCE EDUCATION (IJODE)

VOLUME 5

DETERMINING PSYCHOMETRIC PROPERTIES OF ACHIEVEMENT TESTS IN ODL BUSINESS MANAGEMENT

Esther O. Durowoju,
(Doctoral Candidate),

Adams O. U. Onuka, (Ph. D.)
&
Adesoye T. Onabamiro,
(Doctoral Candidate),
Institute of Education,
University of Ibadan, Ibadan, Nigeria
Durowjuesther@yahoo.com
or adamonuka@yahoo.com
or onabamiros@yahoo.co.uk.

Abstract

The study sought to expose ODL educators to the challenges of constructing achievement tests in Distance Learning School of Business Management. A pool of one hundred test items were drawn and administered on two hundred respondents, who were purposively selected among the University of Lagos Distance Learning Business Management students. The data collection instrument used was Business Management Achievement Test (BMAT). The discrimination and difficulty level of the test items were determined and the instrument was subjected to content and concurrent validity and Kuder-Richardson formula 20 was used to ascertain its reliability. The results showed that 48% of the test items were moderately difficult and discriminated well among the examinees, 12% of the items were too difficult, while 40% of the items were too simple for the testees. The study found out that the instrument was 0.94 reliable and valid (0.78 for content and 0.76 for concurrent). The implications of the findings were discussed with a view to exposing distance educators to the challenges involved in constructing a valid and reliable test.

INTRODUCTION

In the formal education system, knowledge is imparted to learners through personal contacts between the learners and the teacher. Basically, knowledge is passed on during the teachinglearning process, bringing about a desirable and steady change in learners' behaviour. For teaching to be effectively carried out, it is expedient for trained teachers to clearly outline achievable objectives that are specific, observable and measurable right from the outset. After teaching and learning had taken place, a good teacher would desire to know whether teaching has really taken place or whether learners have mastered the lesson taught. To do this, the teacher does not need to wait till the end of the course before assessing the learner, but has to do the assessment on a regular basis (this form of assessment is assessment for learning or formative evaluation). The process of assessing the students on a regular basis is called continuous assessment. Onuka (2008) states that in times past, distance learners were evaluated through the use of mainly tests, usually administered at the end of a year. It has since been discovered, however, that such a one-time terminal or summative test was bedeviled with lots of weaknesses. He went further to say that one major problem in the process of evaluation of open distance learning is that a one-time a year examination does not allow the learner to gauge his/her progress in the course of study, because it did not take into serious consideration the unique roles of the individual differences among learners to be so assessed. These weaknesses of the one-shot tests do not allow for the best picture of a distance learner's performance to be revealed in all its ramifications. The question one may be interested in is, how adequate is the one-time achievement test conducted in ODL? Pilot testing is another challenge posed to the ODL instructors, because it is difficult for these instructors to gather the ODL students together for pilot testing. Despite these challenges, the ODL instructor has no excuse of not making use of a good test.

Achievement tests are designed to measure present proficiency, mastery and understanding of general and specific areas of knowledge (Onuka, 2008). They are meant to determine the effectiveness of an instruction and learning, that is how well students have gained from the teachinglearning process. Achievement tests examines an individual on the level of skills acquired in a discipline or subject which the student has studied in some form. This type of test measures the actual level of previously acquired knowledge. Childs (1989) opines that achievement tests are well suited to provide educators with objective feedback on how much students are learning and understanding. In essence, it is designed to measure accomplishment.

Achievement tests can be classified in several ways, among which are: standardised and specially constructed tests. Specially constructed tests are ordinarily teacher made tests, which are used for promotion, counselling and remediation for students with learning difficulties. Specialised achievement tests can further be classified into general and special tests. General tests are typically batteries of tests that measure the most important areas of school achievements, such as language usage, vocabulary, reading, arithmetic, and social studies e.g. Numerical Aptitude Test, Perceptual Aptitude Test, while special achievement tests are tests in individual subjects such as History, English, Mathematics, Biology, Chemistry, Government etc.

Achievement tests can also be classified into essay and objective tests. Essay type of test is required, if the teacher is interested in testing the ability of students to organise materials, integrate ideas or interpret data, develop arguments, make comparisons and display other abilities involving original written verbal expression (Classroom Assessment, 2010). To achieve this, less structured response instruments that provide the students the chance to express, organise and produce ingenious answers to problem situations, are needed. Thus, an essay question is defined as a test item which requires a response composed by the examinee usually in the form of one or more sentences of a nature that no simple response or patterns of response can be listed as correct and the accuracy and quality of materials can be judged objectively only by skilled examiners. The essay test can be classified into restricted and extended essay type. Essay test is limited in use, as it is cumbersome and time consuming to mark, though easy to construct. It is also expensive and samples only a limited content area. Objective is a type of test that is made up of items, each of which has one or at least a few acceptable responses and in which the acceptable responses have been agreed upon in advance by the examiner. The objective test can be classified into multiple choice, fill-in-the-blank or short answer, true/false, matching, yes no. Objective test responses can be scored by machine or by a routine clerk from the answer key.

It is apparent that most of the tests taken by students in classrooms in most of our schools are "teacher-made". These tests are designed or selected at the teacher's initiative and tailored to unique classroom circumstances, students' ability, and daily instructional needs. These *instructional tests* provide the teacher with immediate feedback about a student's mastery of a subject area or specific skill. Most often, teacher-made tests differ considerably from teacher to teacher because of variations in classroom circumstances. In spite of the fact that tests are undeniably useful at the classroom level, results from instructional tests are unlikely to be comparable across classrooms or schools (Durowoju, 2010).

A good test that can be used to accurately measure the achievement of the general student population must be proven to be moderately difficult, able to discriminate among above average, average, and below average students. Validity and reliability are also part of the characteristics or qualities of a good test or any research instrument. Difficulty level refers to how appropriate the test items are with respect to the level of learners being tested. Item difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be (Wood, 1960). A rough "ruleof-thumb" is that if the item difficulty is more than .75, it is an easy item; if the difficulty is below 0.25, it is a difficult item.
(http://fcit.usf.edu/assessment/selected/responsec.html). It has to do with how many of the learners can answer the test items correctly. Factors considered in determining the difficulty level are the level and age of the learners, as well as the purpose of the test. It is expected that a test should not be too difficult or too easy for the students. It should be such that students would be able to score about 50% if they have been thoroughly and adequately taught.
It is pertinent to mention that a good test must possess the ability to discriminate among testees. If the test and a single item measure the same thing, one would expect people who do well on the test to answer that item correctly, and those who do poorly to answer the item incorrectly. A good item discriminates between those who do well on the test and those who do poorly. The discriminating power shows the degree to which the item measures the differences between the high scores and the low scores on each item. As a rule of thumb, in terms of discrimination index, .40 and greater are very good items, .30 to .39 are reasonably good but possibly subject to improvement, .20 to .29 are marginal items and need some revision, below .19 are considered poor items and need major revision or should be eliminated (Ebel & Frisbie, 1986).

Reliability refers to the consistency of assessment scores. On a reliable test, a student would be expected to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response. On an unreliable examination, a student's score may vary based on factors that are not related to the purpose of the assessment.

Many teachers are probably familiar with the reliability terms "test-retest reliability," "equivalent-forms reliability," and internal consistency reliability. Testretest reliability is also called stability reliability. It refers to the degree to which scores on the same test by the same individuals are consistent over time. It provides evidence that scores obtained on a test at one time (test) are the same or close to the same when the test is re-administered some other time (retest). Equivalent-forms reliability/Parallel form: Equivalent forms of an instrument are two instruments that are identical in every way except for the actual items included. The two forms measure the same concept; have the same number of items, the same structure, the same difficulty level, and the same directions for administration, scoring and interpretations. If the resulting coefficient of equivalence is high, the test has good equivalent forms of reliability.

Internal consistency is a commonly used form of reliability that deals with one test at one time. It is conceptualised through four different approaches, which are Split-half reliability (subdivided test), Kuder-Richardson method of rational equivalence, Cronbach Alpha (and Hoyt's analysis of variance procedure which is rarely used. Each approach provides information about the consistency among the items in a single test. These approaches help to eliminate sources of measurement errors such as differences in testing conditions. Each of these terms refers to statistical methods that are used to establish consistency of student performances within a given test or across more than one test. These types of reliability are of more concern on standardised or high stakes testing than they are in classroom assessment. In a classroom, students' knowledge is repeatedly assessed, and this allows the teacher to adjust as new insights are acquired.

The two forms of reliability that are typically considered in classroom assessment and in test development involve rater (or scorer) reliability. Rater reliability generally refers to the consistency of scores that are assigned by two independent raters and that are assigned by the same rater at different points in time. The former is referred to as "inter-rater reliability" while the latter is referred to as "intra-rater reliability". In Moskal (2000), a framework for developing marking schemes was presented and the issues of validity and reliability were given concise attention. Although, many teachers have been exposed to the statistical definition of the terms "validity" and "reliability" in teacher preparation courses, these courses often do not discuss how these concepts are related to classroom practices (Stiggins, 1999). A perfectly valid system is one that produces a result that is shown, through the passage of time, to have been correct. A valid blood test, therefore, is one that assess that the subject has Hepatitis B, and the subject indeed goes on to develop symptoms that confirm Hepatitis B assessment. A perfectly valid political poll would predict in advance the winner of the election.

Validity is the most important attribute to consider when preparing or selecting an instrument for use. It is worth mentioning that inferences cannot be made from data that has been collected with instruments not serving the purpose for which the instruments are intended. It refers to the degree to which the evidence from the instrument supports the correctness of the interpretation of the data from the instrument and that the manner in which the interpretations are used is appropriate (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999).

The ability of a test to produce findings that are in agreement with theoretical or conceptual values; to produce accurate results and to measure what is supposed to be measured is an indication that it is valid. Validity can also be described as a process of ascertaining substantial proofs that a test is appropriate when it measures what it is supposed to measure. A research instrument is said to be valid if it actually measures what it is supposed to measure. Rulers, thermometers, measures of weight and other instruments used to measure the physical world have demonstrable validity. Hence, validity means that it is true that the instrument measures what it is supposed to measure and that the data collected honestly and accurately represents the respondent's opinion. Since tests are designed for a variety of purposes and since validity can be evaluated only in terms of purpose, it is not surprising that there are several different types of validity. The different types of validity can be categorised into *logical, criterion-related and consequence* validity. Logical validity includes face, content and constructs validity, and is so named because validity could be determined through judgment (Durowoju, 2010).

Content validity is a process of ascertaining the extent to which a test adequately samples or measures behaviour that the test is designed to measure. An achievement test is said to be content valid when the proportion or amount of material covered by the test is equal or approximate to the proportion of material covered in the course. Some methods for determining content validity have been developed. One of such method was developed by Lawshe (1975), who proposed a simple formula for quantifying the degree of consensus by asking a panel of experts to determine the content validity of an employment test. This method can also be applied to other situations requiring a panel of experts to render some judgement, as in the examination of the content validity of mathematics achievement tests (Crocker et. al., 1988).

In validating a test, the content validity ratio is calculated for each item. Lawshe (1975) recommends that if the amount of agreement observed has more than a 5% chance of occurring by chance, the item should be eliminated. The minimal CVR values corresponding to this 5% level are presented in the table below:

Number of Panelists	Minimum Values
5	.99
6	.99
7	.99
8	.75
9	.78
10	.62
11	.59
12	.54
13	.51
14	.49
15	.42
20	.37
30	.33
35	.31
40	.29

Criterion-related or empirical validity includes concurrent and predictive validity and are so named because in each case, validity is determined by relating performance on an instrument to performance on another criterion. Concurrent validity is determined when test scores are obtained at the same time that the criterion measures are obtained, and the measure of the relationship between the tests scores and the criterion gives evidence of concurrent validity. On the other hand, predictive validity is determined when test scores are obtained at one time and the criterion measures are obtained at a future time after some intervening variables or events have taken place (e.g. training, experience, therapy, medication, sickness (Onabamiro & Durowoju, 2010). The measure of relationship between the test scores and a criterion measure obtained at a future time is referred to as predictive validity of the test. Another important type of validity is Consequential Validity, which refers to the process of examining the consequences or uses of the assessment results. A teacher may find out that the application of a test to evaluate the performances of male and female students on a given task consistently results in lower performances of the male students. The interpretation of this result may be that the male students are not as proficient within the area that is being investigated as the female students.

Though there are essential qualities that a good test must possess, it is apparent that most ODL course facilitators are having the challenge of how to construct valid achievement tests in the courses they facilitate or write materials on, with all the qualities of a good test. Therefore, this study sought to expose educators to the techniques of constructing valid achievement tests in ODL.

This study sought answers to three research questions as follows:

1. *What are the ranges of difficulty level and discriminating index respectively of the achievement tests in Business Management?*
2. *What is the degree of the reliability of the achievement tests in Business Management?*
3. *What is the degree of:*

a. The content validity of the achievement tests in Business Management?

b. The concurrent validity of the achievement tests in Business Management?

METHODOLOGY

This is a survey research adopting ex-post facto procedure to collect data since the researchers have no direct control over independent variables as their manifestations have already occurred (Kerlinger & Lee, 2000). Population, Sampling and sample. The target population for this study comprised of all 300 Level ODL Business Management students of Lagos

State University, Nigeria. 200 respondents were randomly selected from 356 three hundred level students of the university of Lagos Dis tance Learning Business Management course.

Instruments/ Instrumentation

Instrument

Business Management Achievement Test (BMAT)

The Business Management Achievement Test consisted of 100 multiple choice items. These items were distributed by content of Analysis for Business Decision course into cognisance of three out of the six domains of cognitive learning by Benjamin Bloom. The behavioural objectives used were knowledge, comprehension and application in line with the course contents. The item distribution is illustrated in the table of specification stated below. The most recent scores of the students in the course were also collected.

TABLE 1: TEST BLUEPRINT TABLE/ TABLE OF SPECIFICATION

CONTENTS	ANALYSIS	SYNTHESIS	EVALUATION	TOTAL
The nature of organisations and organisation theory	3	7	6	16
The classical -mechanistic theory of organisation and management	4	8	5	17
The behaviour -humanistic theory of organisation and management	3	8	4	15
Modern theories of organisation and management	4	9	4	17
Authority and power in organisations	3	9	7	19
Conflict and change in organisations	3	6	7	16
Total	18	47	35	100

instrument was administered on 200 students and scored using the multiple choice test marking guides. The test scores were then used to determine the discrimination index and difficulty level. The reliability of the instrument was also ascertained using KuderRichardson formula 20, which provides an estimate of what is called internal consistency and concurrent validity. The instrument was also determined using the scores of the students from their class on the same course as the criterion-reference.

Data Collection Procedure

The pool of items of (BMAT) were administered on the students in order to determine the good items whose reliability was determined using Kuder-Richardson 20 formula after which it was correlated with the most recent scores of the students in the same course. The content validity was determined using the opinion of ten experts, which was later subjected to content validity ratio.

Data Analysis

Data were scored and the resulting data were then collated and analysed using the following formulae:

Research Question 1:

The above formula was used to calculate the ***range of difficulty level of the test where:***

H = (No of items with high scores)

L = (No of items with low scores)

N = (No of students involved in the test analysis)

Research Question 1b The formula for calculating discrimination is

D = H - L, where

H = No of high scorers

L = No of low scorers.

Research Question 2: Kuder-Richardson 20 was used to determine reliability.

$$R=\frac{N}{N-1}\left[\frac{\ddot{a}x^2\ \acute{O}pq}{\ddot{a}x^2}\right]$$

Where $\ddot{a}x^2$ = variance of testees' scores

P= proportion of testees that answered each item correctly.

Q= proportion of testees that answered each item wrongly.

Research Question 3a:

I. Content Validity BMAT was determined by using the formula developed by Lawshe (1975).The formula is used in quantifying the degree of consensus by asking a panel of experts to determine the content validity of the instrument. Each panel member responds to the following questions for each of the test items: "Is the skill or knowledge measure by this item -

- Essential
- Useful but not essential
- Not necessary.

Hence, the formula for quantifying the content validity is referred to as content validity ratio

CVR =

CVR = Content Validity Ratio

N_c = No of panels indicating "essential"

N = Total number of panels

b. Concurrent Validity To determine the concurrent validity of BMAT, the measure of relationship (comparison) between the current test scores of the respondents (test B) and the BMAT scores (test A) was ascertained.

Results and Discussion

Research Question 1

What are the ranges of difficulty level and discrimination indices respectively of achievement tests in Business Management?

Item analysis:

The answer scripts were marked using the marking scheme. The scores of each testee on each item were added together. The totals were arranged in descending order and 27% upper scorers and 27% lower scorers were selected. The total number of the testees who got each item correctly was also taken into consideration. Likewise, the difference between upper scorer and lower scorer on each item was recorded. Based on these, the discrimination index and difficulty index were calculated.

For discrimination index, the formula; was used.

Table 1 below shows the number of the items and their properties

TABLE 2.1: SAMPLE OF THE ITEM ANALYSIS OF DISCRIMINATION AND DIFFICULTY INDICES

ITEM NO	UPPER SCORERS = 54	LOWER SCORERS = 54	TO-TAL	DIFFE-RENCE	DISCRIMI-NATION INDEX	DIFFI-CULTY INDEX	REMARKS
1	54	32	86	22	.41	.80	Too simple
2	54	32	86	22	.41	.80	Too simple
3	54	23	77	31	.57	.71	Moderately difficult
4	27	5	32	22	.41	.30	Too difficult
5	54	50	104	4	.07*	.96	Too simple

Items with low discriminating power

Item selection:

The table below shows the number of items selected with their discrimination and difficulty indices.

TABLE 2.2: SAMPLE OF THE FINAL ITEMS SELECTED AFTER THE ANALYSIS

ITEM NO	UPPER SCORERS = 54	LOWER SCORERS = 54	TOTAL	DIFFERENCE	DISCRIMINATION INDEX	DIFFICULTY INDEX
1.	54	23	77	31	.57	.71
2.	53	20	73	33	.61	.68
3.	45	20	65	25	.46	.60
4.	54	20	74	24	.44	.69
5.	30	13	43	17	.31	.40

The above table shows the discrimination and difficulty indices of the items finally selected. From the table, the discrimination indices varied from 0.40 and above are good items and the difficulty indices varied from 0.25 to 0.75.

From the result, 26 items were too simple, and 12 items were too difficult. 40% of the items did not discriminate well, while 60% of the items have a high discriminating power. Hence, only 48 items out of the 100 items were selected. See appendix II for details. This agrees with the recommendation of Sidhu (2005) that items with 0.40 and above discrimination index is a good item.

The result shows that only 48 items out of the 100 items, were moderately difficult and discriminate well among the testees and these items were suitable for the category of testees for which it was designed.

The above table shows the discrimination and difficulty indices of the items finally selected. From the table, the discrimination indices varied from 0.30 and above. The difficulty index varied from 0.25 to 0.75. The item number only represented the items that were selected. See appendix I for detail.

The result revealed that 48% of the test items discriminate well and were moderately difficult while 12% of the items were either too difficult or did not discriminate well and 26% items were too simple or did not discriminate well among the testees. Hence, the 12% items that were too difficult and 26% items that were too simple were discarded from the items. This is in line with the assertion of Wood (1960), that the larger the percentage getting an item right, the easier the item and the lower the percentage getting an item right, the more difficult the item. It also corroborates the rule of thumb quoted by who? Classroom Assessment (2010), in
That if item difficulty is more than 0.75, it is an easy item and if the difficulty is below 0.25, it is a difficult item.

Research hypothesis 2

Reliability of BMAT:
Kuder-Richardson 20 (KR20) was used to establish the reliability coefficient.

$$R=\frac{N}{N-1}\ \left[\frac{\text{äx}^2\ \text{Ópq}}{\text{äx}^2}\right]$$

$$\text{Therefore, } R=\frac{100}{99}\ \text{x}\left[\frac{258.52\ \ 19.04}{258.52}\right]$$

$$=1.01 \text{ x } 0.926 = .93526. \qquad \text{Therefore } R^2=0.87$$

The reliability coefficient of 0.9 obtained from the tryout exercise shows that the instrument possesses a high internal consistency and, therefore, highly reliable as a measure of Distance Learners' achievement in Business Management. The R^2 of 0.875 means that 87.5% variation in DLC students' Business Management achievement is measured by BMAT and 12.5% is traceable to other factors. This result corroborates the finding of Onabamiro (2007) who found the reliability coefficient of Mathematics achievement test to be 0.923 with the R^2 value 0.85. The agreement in the reliability coefficient of Business Management and that of Mathematics must mean that thorough work was done in the construction and validation of the instrument.

Research Question 3

A. Content validity of BMAT

To determine the content validity, the researchers made use of ten panellists. The formula for quantifying the content validity is referred to as content validity ratio

CVR =

For Item 1

CVR= = 3/5 = 0.6

For Item 2

CVR = = ´4/5 = 0.8

The CVR was done for all the items and average was found for the instrument which was 0.78. This agrees with the assertion of Nunnaly (1978) that 0.7 is an acceptable validity coefficient but that lower thresholds are sometimes used in the literature. The value, which is a bit higher than the recommended 0.70 put the BMAT at advantage.

b. Concurrent Validity of BMAT:

To determine the concurrent validity, the result of the students in their first semester 300 level in Business Management was correlated with their rèsults in BMAT. The correlation was 0.76. This result corroborates the assertion of Nunnaly (1978) that 0.7 is an acceptable validity coefficient but that lower thresholds are sometimes used in the literature. The value, which is a bit higher than the recommended 0.70 put the BMAT at advantage.

CONCLUSION

To a layman, construction of test seems to be easy and simple, but in the real sense, it requires expertise, skill or knowledge, and mental application. It is also time-consuming most especially, the construction of objective tests.

To construct Business Management Achievement multiple choice objective items, involved critical thinking of what should be the distracters of the right options, endurance, patience and dedication.

However, it is not an easy task to prepare a set of good and acceptable test items. This is shown in the analysis, where only 48 out of the 100 items developed were good items.

Suggestions

Based on the administration and analysis of the test and findings, the researchers therefore suggest that:

a. To get a reasonable pool of good items such as sixty and above, the researcher should try to construct about 200 items,

b. Remove the simple and difficult items and add moderately difficult items.

C. The larger the test, the higher the reliability and the longer the number of the good items that will be derived and the more the test will be able to sample the content.

Recommendations

Government and the school authorities should organise interactive workshops for teachers where they will have the opportunity to share and solve areas of difficulty in their courses and also learn how to construct reliable and valid tests. Such workshops will also afford the teachers the forum to suggest ways of reducing tension in students.

REFERENCES

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1990). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Childs, Ruth Axman (1989). ERIC Clearinghouse on Tests Measurement and Evaluation. Washington DC, American Institutes for Research Washington DC.

Classroom Assessment (2010). Item Analysis (retrieved from http://fcit.usf.edu/assessment/selected/responsec.html on 4 - 09 - 10)

Crocker, L., Llabre, M., & Miller, M. D. (1988). The Generalizability of

Content Validity Ratings. *Journal of Educational Measurement, 25, 287 299.*

Durowoju, E. O. (2010). Standardisation of Test Items in Business Studies. A term paper presented at the Institute of Education, University of Ibadan, Nigeria.

Ebel, R. L., & Frisbie, D. A. (1986). Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall.

Lawshe, C. H. (1975). A Quantitative Approach to Content Validity. *Personnel Psychology,* 28, 563 575.

Moskal, B. M. (2000). Scoring Rubrics: What When and How? Practical Assessment, Research and Evaluation, 7(3)

Nunnaly, J. (1978). Psychometric Theory. New York: McGraw-Hill.

Olubodun, O. J. (2007). Test Construction Techniques and Principles.

Onabamiro, A. T. (2007). Standardisation of Test Items in Mathematics. A Term paper presented at the Institute of Education, University of Ibadan, Nigeria

Onabamiro, A. T. & Durowoju, E. O. (2010) Practical Steps in Test Construction; in Onuka, A. O. U. (ed.) Some Fundamentals of

Evaluation in Distance Learning. Ibadan: Distance Learning Centre, University of Ibadan, Nigeria

Onuka, A. O. U. (2008). Teacher-initiated Student Peer-assessment: A Means of Improving Learning-assessment in Large Classes. A

paper presented at the West African Examinations Council Monthly Seminar, Lagos: Research Division April 30, 2008.

Sidhu K. S. (2005). New Approaches to Measurement and Evaluation.: *Sterling Publishers Private Limited). Stinggins, R. J.*

(1999). Evaluating Classroom Assessment Training in Teacher Education Programs. Educational Measurement: Issues and Practice, 18 (1), 2327. Wood, D. A. (1960). Test Construction: Development and

Interpretation of Achievement Tests. Columbus, OH: Charles E. Merrill Books, Inc.

APPENDIX I
ITEM ANALYSIS OF DISCRIMINATION AND DIFFICULTY INDICES

ITEM NO	UPPER SCORERS = 54	LOWER SCORERS = 54	TOTAL	DIFFERENCE	DISCRIMINATION INDEX	DIFFICULTY INDEX	REMARKS
1	54	32	86	22	.41	.80	Too simple
2	54	32	86	22	.41	.80	Too simple
3	54	23	77	31	.57	.71	Moderately difficult
4	53	20	73	33	.61	.68	Moderately difficult
5	45	20	65	25	.46	.60	Moderately difficult
6	54	20	74	24	.44	.69	Moderately difficult
7	30	13	43	17	.31	.40	Moderately difficult
8	27	5	32	22	.41	.30	Too difficult
9	48	27	75	21	.39	.69	Moderately difficult
10	54	40	94	14	.30	.87	Too simple
11	53	41	94	12	.22	.87	Too simple
12	13	7	20	6	.11	.19	Too difficult
13	53	39	92	14	.30	.85	Too simple
14	52	45	97	7	.13	.90	Too simple
15	45	34	79	11	.20	.73	Moderately difficult
16	36	35	71	1	.02	.66	Moderately difficult
17	52	27	97	25	.46	.90	Too simple
18	23	8	31	15	.28	.29	Too difficult
19	54	22	76	32	.59	.70	Moderately difficult
20	50	38	88	12	.22	.81	Too simple
21	52	41	93	11	.20	.86	Too simple

22	49	41	90	8	.15	.83	Too simple
23	45	34	79	11	.20	.73	Moderately difficult
24	54	50	104	4	.07*	.96	Too simple
25	30	24	54	6	.11	.50	Moderately difficult
26	43	29	72	14	.26	.67	Moderately difficult
27	31	26,	57	5	.09	.53	Moderately difficult
28	50	30	80	20	.37	.74	Moderately difficult
29	53	40	93	13	.24	.86	Too simple
30	50	20	70	30	.56	.65	Moderately difficult
31	43	26	69	17	.31	.64	Moderately difficult
32	56	31	86	25	.46	.80	Too simple
33	35	12	47	23	.43	.44	Moderately difficult
34	50	46	96	4	.07	.89	Too simple
35	44	18	62	26	.48	.57	Moderately difficult
36	26	14	40	12	.22	.37	Moderately difficult
37	49	28	77	21	.39	.71	Moderately difficult
38	54	42	96	12	.22	.89	Too simple
39	54	41	95	13	.24	.88	Too simple
40	10	1	11	9	.17	.10	Too difficult
41	40	30	70	10	.19	.65	Moderately difficult
42	51	30	81	21	.39	.75	Moderately difficult
43	51	39	90	12	.22	.83	Too simple
44	50	27	77	23	.43	.71	Moderately difficult
45	28	8	36	20	.37	.33	Too difficult
46	36	16	52	20	.37	.48	Moderately difficult
47	48	30	78	18	.33	.72	Moderately difficult
48	47	28	75	19	.35	.69	Moderately difficult
49	48	26	74	22	.40	.69	Moderately difficult
50	33	16	49	17	.31	.45	Moderately difficult

51	51	35	86	16	.30	.80	Too simple
52	54	34	88	20	.37	.81	Too simple
53	48	23	81	25	.46	.75	Moderately difficult
54	44	17	61	27	.50	.56	Moderately difficult
55	55	30	85	25	.46	.79	Too simple
56	55	38	93	17	.31	.86	Too simple
57	34	21	55	13	.24	.51	Moderately difficult
58	49	26	75	23	.43	.69	Moderately difficult
59	53	31	84	22	.40	.78	Too simple
60	35	28	63	7	.13	.58	Moderately difficult
61	54	36	90	18	.33	.83	Too simple
62	47	15	62	32	.59	.57	Moderately difficult
63	33	10	43	23	.43	.39	Moderately difficult
64	37	14	51	23	.43	.47	Moderately difficult
65	44	36	80	8	.15	.74	Moderately difficult
66	11	4	15	7	.13	.14	Too difficult
67	53	16	69	37	.69	.64	Moderately difficult
68	49	30	79	19	.35	.73	Moderately difficult
69	40	19	59	21	.39	.55	Moderately difficult
70	51	36	87	15	.28	.81	Too simple
71	25	3	28	22	.41	.30	Too difficult
72	52	26	78	26	.48	.72	Moderately difficult
73	52	30	82	22	.41	.80	Too simple
74	54	35	98	19	.35	.91	Too simple
75	48	28	76	20	.37	.70	Moderately difficult
76	23	14	37	9	.17	.34	Too difficult
77	53	26	79	27	.50	.73	Moderately difficult
78	48	16	64	32	.59	.59	Moderately difficult
79	49	17	66	32	.59	.61	Moderately difficult
80	40	16	56	24	.44	.52	Moderately difficult

81	29	10	39	19	.35	.36	Moderately difficult
82	51	29	80	22	.41	.74	Moderately difficult
83	31	17	48	14	.30	.44	Moderately difficult
84	27	7	34	20	.37	.31	Too difficult
85	49	33	82	16	.30	.80	Too simple
86	51	33	84	18	.33	.77	Too simple
87	37	22	59	15	.28	.55	Moderately difficult
88	39	23	62	16	.30	.57	Moderately difficult
89	38	27	65	11	.20	.60	Moderately difficult
90	34	21	55	13	.24	.51	Moderately difficult
91	45	25	70	20	.37	.65	Moderately difficult
92	24	14	38	10	.19	.35	Moderately difficult
93	20	0	20	20	.37	.19	Too difficult
94	46	12	48	34	.63	.44	Moderately difficult
95	11	1	12	10	.19	.11	Too difficult
96	46	14	60	32	.59	.56	Moderately difficult
97	30	9	39	21	.39	.36	Moderately difficult
98	10	1	11	9	.17	.10	Too difficult
99	32	14	46	18	.33	.92	Too simple
100	50	23	73	27	.50	.68	Moderately difficult

APPENDIX II

FINAL ITEMS SELECTED AFTER THE ANALYSIS

ITEM NO	UPPER SCORERS = 54	LOWER SCORERS = 54	TOTAL	DIFFERENCE	DISCRIMINATION INDEX	DIFFICULTY INDEX
1.	54	23	77	31	.57	.71
2.	53	20	73	33	.61	.68
3.	45	20	65	25	.46	.60
4.	54	20	74	24	.44	.69
5.	30	13	43	17	.31	.40
6.	27	5	32	22	.41	.30
7.	48	27	75	21	.39	.69
8.	54	22	76	32	.59	.70
9.	50	30	80	20	.37	.74
10.	50	20	70	30	.56	.65
11.	43	26	69	17	.31	.64
12.	35	12	47	23	.43	.44
13.	44	18	62	26	.48	.57
14.	49	28	77	21	.39	.71
15.	51	30	81	21	.39	.75
16.	50	27	77	23	.43	.71
17.	28	8	36	20	.37	.33
18.	36	16	52	20	.37	.48
19.	48	30	78	18	.33	.72
20.	47	28	75	19	.35	.69
21.	48	26	74	22	.40	.69

22.	33	16	49	17	.31	.45
23.	48	23	81	25	.46	.75
24.	44	17	61	27	.50	.56
25.	49	26	75	23	.43	.69
26.	47	15	62	32	.59	.57
27.	33	10	43	23	.43	.39
28.	37	14	51	23	.43	.47
29.	53	16	69	37	.69	.64
30.	49	30	79	19	.35	.73
31.	40	19	59	21	.39	.55
32.	25	3	28	22	.41	.30
33.	52	26	78	26	.48	.72
34.	48	28	76	20	.37	.70
35.	53	26	79	27	.50	.73
36.	48	16	64	32	.59	.59
37.	49	17	66	32	.59	.61
38.	40	16	56	24	.44	.52
39.	29	10	39	19	.35	.36
40.	51	29	80	22	.41	.74
41.	31	17	48	14	.30	.44
42.	27	7	34	20	.37	.31
43.	39	23	62	16	.30	.57
44.	45	25	70	20	.37	.65
45.	46	12	48	34	.63	.44
46.	46	14	60	32	.59	.56
47.	30	9	39	21	.39	.36
48.	50	23	73	27	.50	.68