aksara

jurnal bahasa, seni, dan pengajarannya

aksara	Vol. XII	No. 1	Hal. 1 - 96	Bandar Lampung April 2011	ISSN 1411-2501

Terbitan Jurusan Pendidikan Bahasa dan Seni
FKIP Universitas Lampung

aksara

jurnal bahasa, seni, dan pengajarannya

aksara diterbitkan oleh Jurusan Pendidikan Bahasa dan Seni Fakultas Keguruan dan Ilmu Pendidikan Universitas Lampung. Terbit pertama kali pada April 2000. Terbit setiap April dan Oktober. Memuat artikel ilmiah tentang bahasa, seni, dan pengajarannya yang ditulis dalam bahasa Indonesia maupun bahasa Inggris. Tulisan berupa hasil penelitian dan ulasan hasil penelitian, teori, dan fenomena.

Alamat Penyunting dan Tata Usaha: Jurusan Pendidikan Bahasa dan Seni Fakultas Keguruan dan Ilmu Pendidikan Universitas Lampung Jalan Sumantri Brojonegoro 1, Bandar Lampung 35145 Telp 0721-701609, 08155501652; E-mail jurnal_aksara@yahoo.co.id.

aksara
Vol. XII No. 1 April 2011

DAFTAR ISI

Halaman

aksara Vol. XII No. 1 April 2011 ISSN 1411-2501

THE IMPLEMENTATION OF ITEMAN TO IMPROVE THE QUALITY OF ENGLISH TEST ITEMS AS A FOREIGN LANGAUGE: (AN ASSESSMENT ANALYSIS)

Ujang Suparman
FKIP Lampung University

Abstract : Assessment and evaluation are very important in education especially in English teaching as a foreign language (EFL). They play an important role to determine whether or not the objectives of teaching and learning have been successfully achieved. To make sure that the results of the assessment reflect what really happens, the instruments used to assess them should meet the requirements of good testing instruments. Consequently, the examiners, testers, lecturers, teachers or instructors should have assessed whether the instruments they use to measure the results of their teaching and learning are of good quality or not before they use them. At least, there are two major categories to assess such instruments – manual and non-manual, that is, using a software. To do the analysis of the test items professionally, fast and more comprehensively, this paper offers the procedure of the implementation of the second category, a software called ITEMAN.
Key words: *assessment, validity, reliability,* and *iteman*

INTRODUCTION

Evaluation plays an important role in education including in English teaching as a foreign language (EFL). To produce an ideal outcome of an evaluation, the instruments used should meet the criteria of a good test. One of the ways to measure the quality of test items is by using a software, called *Iteman*. This paper covers the procedure of how to analyze test items, for example, on *Advanced Reading* course consisting of 60 items of multiple choices with four options – A, B, C, and D. The objective of writing this paper is to share information about how to develop the quality of test items using the *Iteman* software. The analysis covered four major issues closely related to the assessment: validity, reliability, discriminating and level of difficulty, each of which is briefly defined in the following paragraph. To understand more on the implementation of the *Iteman* especially for EFL courses, readers are recomended to come to the first ITEC - International Teacher Education Conference 2013 to be organized by FKIP Lampung University. This topic will be presented completely based on empirical reserach on the conference.

Theoretical Foundation

Iteman

What is *Iteman*?

According to Assessment Systems Corporation (ASC) (1989-2006), ITEMAN can be defined as one of the analysis programs that comprise Assessment Systems Corporation's Item and Test Analysis Package. It is very important for lecturers and teachers of English who are responsible for administering tests (such as mid semester and final semester examinations) to know what ITEMAN is; why it is important; how it works, and what the example of an item analysis using ITEMAN. Basically, ITEMAN can be used to analyze test and survey item response data and provide conventional item analysis statistics (e.g., proportion/percentage endorsing and item-total correlations) for each item. Such function is very important for English lecturers in universities and teachers at school levels in order to assist them in determining the extent to which items are contributing to the reliability of a test and which response alternatives are functioning well for each item. Besides item-level statistics, more importantly the ITEMAN program also provides statistical indicators on the performance of the test as a whole (e.g., mean, standard deviation, reliability, median p-value).

How Does ITEMAN Work?

The input data in order to be analyzable by ITEMAN should be formatted in ASCII (text-only) files. This can be completed successfully through the use of the ITEMAN for Windows text editor, Notepad, a word-processing editor that produces true ASCII output, or a program written specifically to format your data. It is also very important to note that all the data to be included in the analysis must be contained in a single input file. A single analysis can cover up to 750 items, while the number of examinees is almost unlimited.

A data file in an ITEMAN can be put under five primary components:
1. A control line describing the data;
2. A line of keyed responses;
3. A line of the numbers of alternatives for the items;
4. A line specifying which items are to be included in the analysis; and
5. The examinee data. (ASC, 1989-2006: 2).

An example of a data file on an ITEMAN

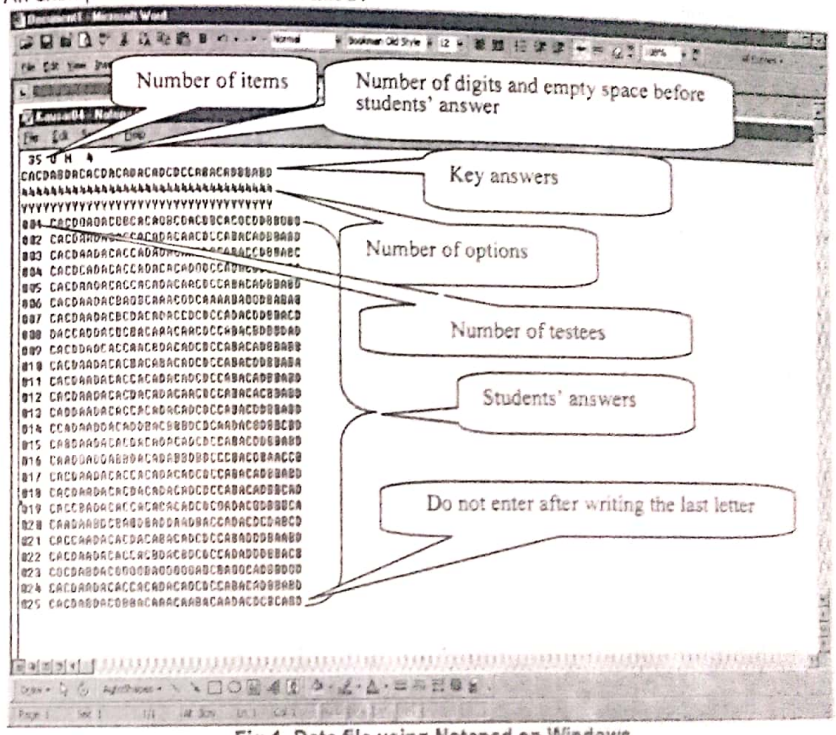

Fig 1. Data file using Notepad on Windows
(Thank Dr. Ngadimun)

Steps of Analysis with an ITEMAN

The ITEMAN program can work only with multiple choice items. It is relatively easy to analyze test items using the ITEMAN program. The most important thing to do is to be very careful in entering the data, because if you enter the data wrongly, it will result in unprecise results of data analysis. The following are the steps to enter the data using a new file.

1. Click *Star*
2. Select *Program*
3. Select *Accessories*
4. Choose and click *Notepad*
5. Save/ Click *File*
6. Select and click *Save as*, then name the data file. for example: Adhesd (make sure the file name must not exceed 8 letters/numbers)

7. Start data entry, it will be faster if you work with your friend – one of you reads students' answers and the other types them. If you work with your friend, please make sure to prounuce the letter clearly, e.g., a for apple; b for ball; c for charlie; d for doctor; and e for ent.
8. It's advisable for you to save it frequently by clicking *File* and then *Save* so that the typed data will not loss if the current suddenly cuts off.
9. The data will appear like shown on the Fig. 1 above.

How to Analyze the Data Using ITEMAN Program?

1. Open ITEMAN Program, by clicking *Star*,
2. Select *Program*/click *ITEMAN*.
3. Type the name of your data file (input) as you like on *Enter the name of the input file*. For example F:\Advread.txt then *Enter*.
4. Enter the name of the output file on *Enter the name of the output file*. For example, inthis case: F:\ Advread output then click *Enter*.
5. A question will appear, *Do you want the scores written to a file? (Y / N)*.
   Then type Y and click *Enter*.
6. Enter the name of your score file on *Enter the name of the score file*: For example, F:\ Advread scr
   Then click Enter. Finish. Have a good try!
   The data will appear like that in the following page.

MicroCat (tm) Testing System
Copyright ©1982,1984, 1986, 1988 by Assessment Systems Corporation
Beta-Test version – Univ. of Pittsburgh
Item and Test Analysis Program -- ITEMAN (tm) Version 3.00

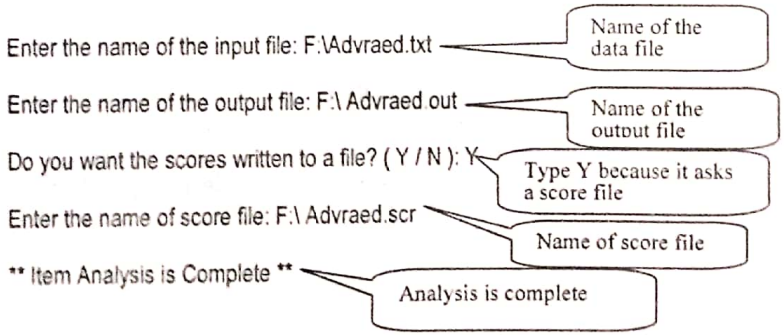

Fig. 2. Item analysis appearance using ITEMAN

The Results of Item Analysis with ITEMAN

The following is the steps that the researcher used to open the results of item analysis on MS Words program:

1. Click *Star*,
2. Select *Program*/click Microsoft Word
3. Click *File*/click *Open*, please look for the results on, for example, Drive F (depends on which one you choose).
4. The following is an example of the appearance of the results of test items analysis.

MicroCAT (tm) Testing System Page 1
Copyright (c) 1982, 1984, 1986, 1988, 1993 by Assessment Systems Corporation
Item and Test Analysis Program -- ITEMAN (tm) Version 3.50
Item analysis for data from file F:\Advraed.TXT
Date: 10/22/12 Time: 8:35 am

	Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	High	Point Biser.	Key
1	0-1	.32	.35	.36	A	.16	.30	.00	-.31	
					B	.32	.20	.55	.36	*
					C	.03	.00	.09	.12	
					D	.49	.50	.36	-.15	
					Other	.00	.00	.00		
2	0-2	.19	.17	.11	A	.05	.00	.09	.19	?
					B	.70	.80	.55	-.27	
CHECK THE KEY					C	.19	.10	.27	.11	*
C was specified, A works better					D	.05	.10	.09	.16	
					Other	.00	.00	.00		
4	0-4	.27	.45	.26	A	.22	.30	.27	.10	
					B	.03	.00	.00	-.02	
					C	.27	.00	.45	.26	*
					D	.49	.70	.27	-.31	
					Other	.00	.00	.00		

5	0-5	.78	.10	.19	A	.78	.90	1.00	.19	*
					B	.00	.00	.00		
					C	.16	.10	.00	-.17	
					D	.05	.00	.00	-.07	
					Other	.00	.00	.00		
6	0-6	.68	.41	.30	A	.08	.10	.09	.12	
					B	.05	.00	.00	-.01	
					C	.68	.50	.91	.30	*
					D	.19	.40	.00	-.44	
					Other	.00	.00	.00		

And so on.

In the following, the resume of the results of the item tests analysis is presented to the right of the score obtained by each of the participants.

Scale:	:	0
N of Items	:	60
N of Examinees	:	37
Mean	:	37.703
Variance	:	51.506
Std. Dev.	:	7.177
Skew	:	-0.677
Minimum	:	19
Maximum	:	51
Median	:	38
Alpha	:	0.798
SEM	:	3.224
Mean P	:	0.628
Mean Item-Tot.	:	0.288
Mean Biserial	:	0.405
Max Score (Low)	:	34
N (Low Group)	:	10
Min Score (High)	:	41
N (High Group)	:	11

No	Scores	Marks
1	37	62
2	24	40
3	45	75
Students' scores	40	67
	20	33
6	33	55
7	45	75
8	19	32
9	41	68
10	31	52

Table 1. The resume of the results of the item tests analysis

More importantly, ITEMAN can function as a powerful technique available to teachers or lecturers for improving the quality of instruction. To achieve this, the items to be analyzed have to meet the following requirements: first, they must be valid measures of instructional objectives; secondly, they must be diagnostic, in the sense that, knowledge of which incorrect

options that the students choose must be a clue to the nature of the misunderstanding, and, therefore, prescriptive of appropriate remediation; and finally, lecturers who construct their own examinations may greatly develop the effectiveness of test items and the validity of test scores if they select and rewrite their items on the basis of item performance data.

Validity

Validity is a characteristic of a good test which requires it to be usable to measure what it is intended to measure and nothing else (Power, 2012). According to Hatch and Farhady (1982), there are two types of variability – internal and external validity. Internal validity refers to the extent to which the outcomes of a test serve the uses for which they were intended. Validity concerns the results of the test not to the test itself. Validity can be stated in high validity, moderate validity, and low validity rather than absolute validity. However, it should be kept in mind that a test can be highly valid for one purpose but not for another. There are three basic types of validity – content validity, construct validity, and criterion related validity.

Content Validity

It refers to the extent to a test measures a representative sample of the subject matter content which is usually described comprehensively in a curriculum or syllabus. The content validity is focused on the adequacy of the sample and not only on the appearance of a test. Content validity must be carefully defined. If a test is dealing with course content, the test items should match to the materials covered in the course.

Construct Validity

Sometimes the examiners expect to establish the validity of certain general psychological constructs, such as *self-esteem, extrovert, acculturated,* and *motivated.* If the examiners wish to interpret test performance in terms of psychological aspects, they are concerned with construct validity. Any examiner who has considered at test of such constructs as *self-esteem, field dependent/independent, integrative/instrumental motivation,* and *right /left hemisphere problem solving* may ask himself/herself questions of how the tests actually test these constructs. However, many people have an objection concerning with such construct validity because they doubt whether the test items really comprise the construct. Therefore, construct validity is not as important as content validity in EFL discussion.

Criterion Related Validity

An examiner concerns with criterion related validity whenever test scores will be used to predict future performance or to estimate current performance on some valued measure other than the test itself. For example, we have constructed a new language aptitude test and we consider that it is a good one. Then the aptitude test is administered to a group of beginning English Education Study Program learners at the FKIP Unila. Then to show it is a valid test, we compare the results with an established test, for example Test of English as a Foreign Language (TOEFL) administered before the students start learning in a university, which is the criterion expected to be able to predict. We predict from our aptitude test scores to performance

on the TOEFL. If a test, for example, English placement test, is used to predict some future performance, say the success in the English Education Study Program, then we are concerned with predictive validity.

However, if we two measures (e.g. TOEFL and IELTS) are administered at the same time and compared, we are checking concurrent validity. Each of the measure is the criterion for the other. Then we are claiming that one measure is a valid test as compared with the other.

In sort, it can be stated that content and construct validity are concerned with some specific practical use of test outcomes. The outcomes help us determine how well test scores represent certain learning objectives (content validity) or how well they predict or estimate a particular performance (criterion-related validity). By contrast, when we want to interpret test performance in terms of psychological aspects, we are dealing with construct validity.

Reliability

In this study the definition of *reliability* is straightforward: a measurement is reliable if it reflects mostly true score, relative to the error. For example, an item such as "Red foreign cars are particularly ugly" would likely provide an unreliable measurement of prejudices against foreign- made cars. This is because there probably are ample individual differences concerning the likes and dislikes of colors. Thus, this item would "capture" not only a person's prejudice but also his or her color preference. Therefore, the proportion of true score (for prejudice) in subjects' response to that item would be relatively small.

There are three major important functions of *Reliability & Item Analysis*. First, they may be used to construct reliable measurement scales, secondly, to improve existing scales, and finally to evaluate the reliability of scales already in use. Specifically, *Reliability & Item Analysis* will aid in the design and evaluation of *sum scales*, that is, scales that are made up of multiple individual measurements (e.g., different items, repeated measurements, different measurement devices, etc.). Numerous statistics can be computed to allow you to build and evaluate scales following the so-called *classical testing theory* model.

Measures of reliability. Based on the discussion above, one can easily infer a measure or statistic to describe the reliability of an item or scale. Specifically, an *index of reliability* may be defined in terms of the proportion of true score variability that is captured across subjects or test takers, relative to the total observed variability. In equation form, it can be stated:

Reliability = $\sigma^2_{(true\ score)} / \sigma^2_{(total\ observed)}$

Number of items and reliability. This concept describes a basic principle of test design. That is, the more items there are in a scale designed to measure a particular concept, the more reliable will the measurement (sum scale) be. Let us examine the following example to clarify the concept. Suppose you want to measure the height of 10 persons, using only a crude stick as the measurement device. Note that we are not interested in this example in the absolute correctness of measurement (i.e., in inches or centimeters), but rather in the ability to distinguish reliably between the 10 individuals in terms of their height. If each person is measured only once in terms of multiples of lengths of your crude measurement stick, the

resultant measurement may not be very reliable. However, if each person is measured 100 times, and then take the average of those 100 measurements as the summary of the respective person's height, then you will be able to make very precise and reliable distinctions between people (based solely on the crude measurement stick).

Discriminating Power

There are two indicators of the item's discrimination effectiveness: point biserial correlation and biserial correlation coefficient (Matlock-Hetze, 1997). The choice of correlation depends upon what kind of question we want to answer. The advantage of using discrimination coefficients over the discrimination index (D) is that every person taking the test is used to compute the discrimination coefficients and only 54% (27% upper + 27% lower) are used to compute the discrimination index, D.

The point biserial (rpbis) correlation is used to find out if the right people are getting the items right, and how much predictive power the item has and how it would contribute to predictions. There are three ways that can be used to measure discriminating power (D) of test items: discriminating index; correlation index; and harmonious index. A discri-minating power is usually symbolized with a capital D. The steps to determine the level of discriminating power are: First, rank order the answer sheet top-down from the highest to the lowest scores based on the total number of test takers; then multiply N with 27%, the results is n score, after that, calculate n from the Upper Group (the answer sheets with high scores are counted from the top) while n from the Lower Group (the answer sheets with low scores are counted from the bottom). And finally, determine the proportion of the test items answered correctly by each group. That is, the correct answers from each of the Upper Group (pU) and Lower Group (pL) are divided by n. the discriminating power is in fact the differences of the proportion of the correct answers between the UG and the LG. So, it can be stated that D = pU – pL.

To determine whether a test item is accepted, revised or rejected, the following parametric criteria is used:

Parameter of D Coefficient	Decision
D = > 0.30	accepted
D = 0.10 - 0.29	revised
D = <0.10	rejected

Level of Difficulty

Level of item difficulty can be defined as the percentage of students taking the test who answered the item correctly. In short, it can be stated that 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). Matlock-Hetzel (1997) states that to compute the item difficulty,

the examiner can divide the number of people answering the item correctly by the total number of people answering item. The proportion for the item is usually denoted as p and is called item difficulty (Crocker & Algina, 1986). An item answered correctly by 85% of the examinees would have an item difficulty, or p value, of .85, whereas an item answered correctly by 50% of the examinees would have a lower item difficulty, or p value, of .50.

The easiest way to measure the level of difficulty of an item is by using proportional scale or proportion correct (p), that is, the number of test takers answering correctly on the items under analysis is compared with the total number of test takers. The equation is as follows:

$$p = \frac{\sum B}{N}$$

where p = the proportion of test takers who answer correctly a certain item under analysis
$\sum B$ = the number of test takers who answer correctly
N = the total number of test takers.

The level of difficulty ranges from 0 through 1. It can be categorized into three classifications as follows:

Proportion Correct (p)		Category
$p \geq 0.70$	:	easy
$0.30 < p < 0.70$	:	Average
$p < 0.30$	:	difficult

METHOD

The design of the research is descriptive assessment, that is, a study describing the results of an analysis of the topic under discussion, which was adjusted with standardized criteria. The research anlyzed the test items used to assess the students' Intermediate Reading Comprehension abilityt on the third semester. The tests were adapted from Test of English as a Foreign Language (TOEFL). The research took place in the English Study Program, the Department of Language and Arts Education, the Faculty of Teacher Training Education, Lampung University. It was carried out during the First Semester of the 2011/2012 academic year. The research participants consist of 35 English students taking an Intermediate Reading course on the First Semester.

The data were collected by means of documentation, that is, using the students' answer sheets on the Intermediate Reading Semester Final Examination on the academic year mentioned above. The data, that is, the students' scores on Intermediate Reading were analyzed using *Iteman* software. The analysis covered four major issues relating to the assessment: validity, reliability, discriminating power and level of difficulty.

HOW TO INTERPRET THE RESULTS OF ITEM ANALYSIS

Based on the recommendations from some experts of measurement, the following criteria to determine the quality of test items have been agreed:

Table 1 Criteria of Test Item Quality

Prop Correct (Level of Difficulty – p)	
0.000 – 0.250 0.251 – 0.750 0.751 – 1.000	Difficult Average Easy
Point Biserial (Discriminating Power – D)	
0.199 – 0.200 – 0.299 0.300 – 0.399 0.400	Very low ≤ D Low Average High ≥ D
Prop Endorsing (proportion of the answers)	
0.000 – 0.010 0.011 – 0.050 0.051 – 1.000	Low Sufficient Good
Alpha (test item reliability)	
0.000 – 0.400 0.401 – 0.700 0.701 – 1.000	Low Average High

Besides the criteria above, the following guideline is considered important to classify which test items need revising or dropping.

Table 2. Criteria to classify the quality of test items

Level of Difficulty (p)	
0.000 – 0.099 0.100 – 0.299 0.300 – 0.700 0.701 – 0.900 0.901 – 1.000	Very difficult/needs total revising Difficult/needs revising Average/good Easy/needs revising Very easy/ needs dropping or total revising
Point Biserial (Discriminating Power – D)	
0.199 – 0.200 – 0.299 0.300 – 0.399 0.400	Very low ≤ D/needs dropping or total revising Low/needs revising Quite average/without revision High ≥ D/very good

Prop Endorsing (proportion of the answers)	
0.000 – 0.010	Least/drop, or needs revising
0.011 – 0.050	Sufficient/good enough
0.051 – 1.000	Very Good
Alpha (reliability of test item)	
0.000 – 0.400	Low/not sufficient
0.401 – 0.700	Average/sufficient
0.701 – 1.000	High/Good

In line with the the resukts of the item analysis in the current study, with the total number of items 60, and the total number of test takers 37, the following classifications can be determined:

...

FINDINGS AND CONCLUSIONS

Based on the results of the data analysis, it was found that

1. Although the tests were adapted from TOEFL (test of English as a foreign language), the quality of their items should be re-assessed, i.e., to make sure the validity, reliability, discriminating power and level of difficulty. This is given that the participants have different backgrounds, cultures, and experiences which may affect the quality of the tests.
2. Some of the tests items were found to be good enough; some need revision, and even some required total revision due to wrong answer.
3. As a whole, the total test items of the reading test is good and can be applied to measure the students' comprehension ability because it has the alpha index of 0.798.

BIBLIOGRAPHY

ASC. 1989-2006. *User's Manual for the ITEMAN™ Conventional Item Analysis Program*. St. Paul, Minnesota: Assessment Systems Corporation

Crocker, L., and Algina, J. 1986. *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Gulon, M.R. 2002. Validity and reliability. In Adam, G. & Berzonsky, M. (eds). 2002. *Blackwell handbook of adolescence*. Malden: Blackwell Publishing Co.

Hatch, E. & Farhady, H. 1982. *Research design and statistics for applied linguistics*. Rowley, Massachusetts: Newbury House Publishers, Inc.

Lyman, H.B. 1971. *Test scores and what they mean*. Englewood Cliffs, New Jersy: Prentice Hall, Inc.

Matlock-Herzel, S. 1997. *Basic Concepts in Item and Test Analysis*. Texas: A&M University

Miriam, K.L. 1996. *An introduction to psychological tests and scales*. London: UCL Press.

Power, T. 2012. *Methods of assessment*. http://www.tedpower.co.uk/esl0706.html (Accessed: 09 May 2013)