Development of Prototype Software for Item Analysis and Item-Banking System for Likert Scales

Hector John T. Manaligod

Introduction

Before a researcher distributes a survey to measure certain attributes of a sample, it is necessary that the survey undergoes the process of item analysis to determine the effectiveness of each item and to build quality item pool for later use. The process of filtering item statements for reliability characteristics entails computing statistics that is tedious especially when done by hand.

When using commercially available statistical software's, the scale constructor has to know how to run the software and find the appropriate command for each statistics for item analysis. Most of these statistical software's are made for a wide-range of purposes but there are only a few software's tailored solely for item analysis. Hogan (2007) pointed out that "Availability of item analysis program is somewhat erratic. They tend to come and go. More exactly, the companies that produce them tend to come and go." Again, the problem that may arise from this issue is the stability of the company that supplies the software to support their clients when technical failure occurs or when the clients need assistance.

Another alternative is the use of electronic spreadsheet for item analysis. Setting-up the worksheet and the formula also requires time and experience. This includes testing the result for accuracy in the worksheet with another computation reference.

Test constructors may opt to purchase software license but the cost of ownership is relatively expensive. Engaging the services of a statistician to do the item analysis is another choice, but again it entails expense.

Thus, there is a need to develop software solely for item-analysis and item-banking that is more practical, less-costly, and compact.

Objectives of the Study

The following objectives served as guide in developing the computer-aided item analysis and item banking software: (a) to develop a compact and integrated software for attitude measurement based on Likert Scale, (b) to help researchers select item statements with acceptable psychometric property, (c) to provide facility for researchers to create, store, and retrieve item statements with proven reliabilities, and (d) to set pioneering effort and generate interest by making locally-made software for measurement and evaluation for Asian countries.

Methodology

There are six phases involved in the development of this software which took approximately 10 weeks to accomplish (See Table 1).

Table 1. Phases of Software Development

	Phase	Duration
1.	Problem Analysis 1. Study Attitude Scale 2. Select Statistical Procedure	3 weeks
2..	Preliminary Design 1. Design General Systems Flow 2. Design Functional Modules	2 weeks
3.	Detailed Systems Design 1. Design Input/Output Format 2. Prepare Programming Specifications 3. Prepare Test Cases	2 weeks
4.	Programming 1. Write Programs 2. Test Unit	2 weeks
5.	Systems Integration 1. Check Interfacing of Modules/Programs 2. Prepare Pilot Data 3. Test Run System	1 week
6.	Presentation and Acceptance	1 day

Problem Analysis

This phase deals with the survey and review of available literature that contributed to the understanding of attitude measurement. Since the software dealt with the construction of item analysis for attitude scales, the study cited several literatures which explained the meaning and characteristics of attitude by Oppenheim (1966), Fishbein and Ajzen (1975), and Anderson (1981).

The basis for the measurement of attitudes are inferences from observable indicators which can be made in view of responses to a series of sentences called 'scales' or adjectives, observing overt behavior, and physiologic responses. Different attitude measurements and their differences cited from Anderson (1981) were Thurstone, Guttman, Semantic Differential, and Likert. In terms of format, the Semantic

Differential uses bipolar evaluation adjectives (e.g., good-bad, nice-awful), while Likert, Thurstone, and Guttman use sentences. The placement of sentences along the continuum also is different for Likert, Thurstone, and Guttman. Likert scales sentences are written only at the two ends of the continuum. In contrast, Thurstone and Guttman sentences are written to represent points in the continuum. The scales also show the extent to which they use the concept of differentiation. Gutman scales are cumulative i.e., positive responses to a sentence positioned somewhere along the continuum, which called for a positive response to all those in the left of that statement on the continuum.

The methodologies chosen to develop computer-aided item analysis and item banking were adopted from Summated Rating technique by Likert and Analysis of Variance (ANOVA) to determine reliability of the scale. Various handbooks for attitude measurement from Oppenheim (1986) and Mueller (1989) were adopted.

Likert Method

Likert in 1932 presented a technique which extracts responses from a group of subjects indicating their own attitudes toward a certain statement by assigning arbitrary values (e.g., the five-point continuum 1,2,3,4,5 or 1-5) to the degree of their agreement or disagreement (Ferguson, 1941 cited online 2009). The Likert method was developed to do away with the selection of items as in the case of Thurstone technique.

The Likert method item responses are categorized from "strongly agree" to "strongly disagree." Five categories are made as standard (strongly agree, agree, uncertain, disagree, and strongly disagree).

In scoring positively stated Likert items, "strongly agree" receives 5 points, "agree" 4 points, and so on. For negatively worded items, the scoring is reversed ("strongly agree" equals 1 point, "agree" equals 2, and so on). Thus, responses indicating a positive attitude toward an attitudinal object (agree responses to positive items, disagree responses to negative items) result in high scores. Responses indicating a negative attitude toward the attitude object result in low scores. In calculating the total scale score for each respondent, item scores are added.

Statistics Used

Four kinds of statistics are computed in order to refine the scale, namely, descriptive statistics (frequency distribution, mean, and standard deviation); discrimination index; correlation coefficient; and alpha coefficient (Cronbach alpha).

Descriptive Statistics includes frequency distribution, mean, and standard deviation of items that are necessary to determine the distribution and spread of the item scores. Items characterized by small standard deviation mean that they are not contributing to the reliability and are considered "factual" statements. Items spread across the response categories are better than those which are clustered in two or three response categories (Mueller, 1989).

Discrimination Index shows the extent to which each item discriminates amon respondents in the same manner as the total scale score. To do this, scores of each respondent are added. The subjects are then ranked from highest to lowest. Fifty cases from the ranked list are selected, 25% are from the lowest scoring subjects and another 25% from the highest scoring subjects. There are also other percentages in calculating the number of individuals comprising the two groups. The mean of high scorers are deducted from the mean of the low scorers item by item. If the difference is small, the particular item does not differentiate the high scorers and low scorers.

Correlation Coefficient is used to determine the extent to which the item relates with the total score. Having scored each item from 1-5 or 5-1, the item scores were added to obtain a total score. For instance, a subject answered 1,4,5 for items 1,2,3 respectively, the total score of the three items (1+4+5) would be 10. Then the total score is deducted from the item in question.

For example, subject A answered 5 in item 1 and the total score is 40, then 35 (i.e., 40-5) would be the new total score.
(See Table 2)

Table 2. Item Score and Total Score

Respon-dent	Total Score	Score on Item 5	Total Score-Item 5
A	40	5	35
B	42	5	37
C	35	4	31

Alpha Coefficient (or Cronbach Alpha) tells the reliability of the item by the proportion of error variance (Ve) to the total obtained variance (Vt). The formula used to arrive at the reliability of the scale (Kerlinger, 1973) is as follows:

$$r_{tt} = 1 - \left(\frac{V_e}{V_t}\right)$$

A hypothetical data is used to illustrate how the reliability coefficient (*Rtt*) was computed. (See Table 3)

Table 3. Analysis of Variance Summary

Source	Df	s.s	m.s.	F
Items	3	6.80	2.27	2.8(ns)
Individuals (Vt)	4	40.30	10.08	12.44 (.001)
Residual (Ve)	12	9.70	.81	
Total	19	56.80		

Source: Sample data in *Foundations of Behavioral Researcher* (1973) - Kerlinger p. 410

The analysis of variance yields the variances between the items, individuals, and residual. The variance for individuals is substituted for total variance (Vt) and residual for error variance (Ve), after which the reliability coefficient is computed using the equation as shown below:

$$r_{tt} = 1 - \left(\frac{V_e}{V_t}\right)$$

$$= 1 - \left(\frac{.81}{10.08}\right) \qquad = .92$$

Preliminary Design

The conceptual framework follows the Input-Process-Output (IPO) format in analyzing system requirements.

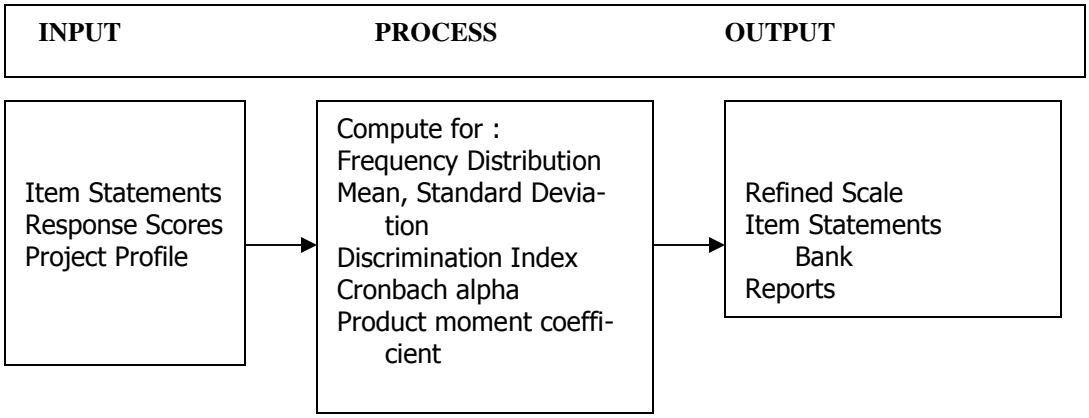

Figure 1. Input-Process-Output Schema

Detailed Systems Design

This phase deals with the design of forms, input and, output files that will be used to encode/edit data, calculate statistics, and generate reports. Here, the interconnections of files are also established. After the file formats are created and their connections made, the logic of the programs are formulated, usually written in English-like structured language called "pseudo language." At this stage, sample data to test if the program is working are created (the author used sample data from the books cited in this study).

Programming

At this stage, the logic to process the data is written in a particular language that can be understood by the computer. In this system, a popular and powerful language called *Visual Basic* was chosen because of its simple language format, syntax, and capability to connect with any databases.

Systems Integration

After all programs have been individually written and tested, the next phase is to test the system as a whole. This means that all programs contribute to the overall purpose of the system

Presentation and Acceptance

The final stage is where the intended users of the system actually "test-drive" the system. It is also at this point where the system is fine-tuned based on comments and suggestions made by the users. This is normal considering that there might be certain specifications that were left out during the design or certain requirements changed during the programming stage when the specifications had already been established.

Data Analysis Procedures

Attitudinal data analysis indicates reliability of the scale. One reliability approach is based on internal consistency. Internal consistency is established in several ways. First, the total score of the individual for all item statements is correlated with the response for a particular item. This method is known as *Summated Ratings*. Items are ranked from highest to lowest according to

The size of the reliability coefficient. Only positively correlated items are selected for the scale. Pearson Product-Moment Correlation is used to calculate the degree of relationship between *item score* (*X*) with *total score* (minus the item score) of each individual which is represented in the formula below as (*Y*).

$$r=\frac{N\sum XY-\sum X\sum Y}{\sqrt{\left[N\sum X^2-\left(\sum X\right)^2\right]\left[N\sum Y^2-\left(\sum Y\right)^2\right]}}$$

The correlation coefficient (Fox, et al, 2003) described the relationship as follows:

Table 4. Correlation Coefficient Evaluation

Correlation Coefficient	Relationship
+1	Perfect positive correlation
+.6	Strong positive correlation
+.3	Moderate positive correlation
+.1	Weak positive correlation
0	No correlation
-.1	Weak negative correlation
-.3	Moderate negative correlation
-.6	Strong negative correlation
-1	Perfect negative correlation

Items with zero or negative correlation coefficient are rejected or scrutinized for possible improvement.

Second, subjects are categorized as high-scoring and low-scoring groups. The means of each group are computed and compared item-by-item to get the significant differences of their means. Items that fail to discriminate the high-group from the low-group are eliminated from the item pool. This procedure is called *discrimination index*. The item discrimination index indicates the extent to which the item statement agrees with the scale as a whole (Hogan, 2007). Popham (2000) proposed criteria for item selection as shown in Table 5.

Table 5. Index of Discrimination

Discrimination Index	Item Evaluation
.40 or higher	Very good items
.30 to .39	Reasonably good, but possibly subjected to improvement
.20 to .29	Marginal items, usually needing and being subjected to improvement
.19 and below	Poor items, to be rejected or improved by revision

Another check on internal consistency of a scale is by computing the *alpha coefficient* (also known as Cronbach Alpha). This is used to analyze items or scales for attitude measurement or items that are not scored right or wrong. While Kuder-Richardson Formula 20 is used for dichotomous responses, the alpha coefficient is used for multiple responses like the response sets found in various attitude scales. Kerlinger (1973) outlined the use of ANOVA in connection with Cronbach Coefficient Alpha. To determine how the item contributes to the reliability of the scale, each item is successively removed, and the reliability coefficient is computed. An item that decreases or does not change the reliability coefficient is eliminated or examined for improvement.

The analysis of variance yields the variance between items, individuals, and residual. The variance for individuals is substituted for the total variance (Vt) and residual for error variance (Ve); after which the reliability coefficient (R_{tt}) is computed using the formula below:

$$r_{tt} = 1 - \left(\frac{V_e}{V_t}\right)$$

Accuracy check was performed by comparing the outputs generated using Statistical Package for Social Sciences (SPSS) and Microsoft Excel against the statistics produced by the item analysis software developed. Test set were from Kerlinger (1973) including the example for the computation of Cronbach's reliability coefficient.

RESULTS OF TESTING

The succeeding discussions compare the outputs generated by SPSS and the developed software. Comparisons are made in the mean and standard deviation, correlation coefficient, alpha/reliability coefficient, discrimination index, and frequency distribution.

Mean and Standard Deviation

Based on the output generated by SPSS as shown in Table 6, the mean for item 1 is 3.60; item 2, 3.80; item 3, 3.80; and item 4, 2.40. The output of the Item Analysis software registered the same values for item 1 that is 3.60, item 2 is 3.80, and so on. SPSS computed values for standard deviations, rounded to the nearest hundredths, for items 1 to 4 are 1.82, 2.28, 1.64, and 1.14. Again, the computed values for the item analysis software (see Table 7) are identical to that of the SPSS's.

Table 6. SPSS Output

Subject		Case Number	Item1	Item2	Item3	item4
	1	1	6	6	5	4
	2	2	4	6	5	3
	3	3	4	4	4	2
	4	4	3	1	4	2
	5	5	1	2	1	1
Total	N		5	5	5	5
	Mean		3.60	3.80	3.80	2.40
	Std. Deviation		1.817	2.280	1.643	1.140

Table 7. Item Analysis Software Output

Mean and Standard Deviation

Project	Code	Description	Mean	Std. Dev.
01	1	Boys prefer to play computer games	3.6	1.82
01	2	Girls prefer spend more time using Friendster	3.8	2.28
01	3	Friendster are only for young people	3.8	1.64
01	4	Internet chat is a waste of time	2.4	1.14

Item-Total Correlation

As shown in Table 8, the correlation coefficients for item 1, item 2, item 3, and item 4 are .92, .76, .83, and .94. Note that in Table 9, the values for the computed correlation coefficient (r) are exactly the same compared to Table 8.

Table 8. Item-Total Correlation Using SPSS

	Item-Total Correlation
item1	.920
item2	.761
item3	.829
item4	.942

Table 9 . Item-Total Correlation using Item Analysis Software

ItemMAN

PEARSON PRODUCT MOMENT

Project Code	Item Code	Description	rxy
01	1	Boys prefer to play computer games	0.92
01	2	Girls prefer spend more time using	0.76
01	3	Friendster are only for young people	0.83
01	4	Internet chat is a waste of time	0.94

Reliability Coefficient (Cronbach Alpha)

The computed value for Cronbach Alpha using SPSS is 0.92, while the value for item analysis is also 0.92 .

Table10. Reliability Coefficient Using SPSS

Cronbach's Alpha	Cronbach's Alpha Based on Standard-ized Items	N of Items
.920	.947	4

Table 11. Reliability Coefficient Using Item Analysis Software

ItemMan
Reliability Coefficient

Source	df	s.s	m.s
Product Code: 01			
Items	3	6.80	2.27
Individuals	4.00	40.30	10.08
Residuals	12.00	9.70	0.81

rtt = 0.92

Discrimination Index

The discrimination indices in Table 12 for items 1,2,3,4 were 3.0, 4.5, 2.5, and 2.0 respectively. The same values were generated in the item analysis (see Table 13) software developed.

Table 12. Discrimination Index Using Excel

Discrimination Index

Subject	Item1	Item2	Item3	Item4	Total
A	6	6	5	4	21
B	4	6	5	3	18
C	4	4	4	2	14
D	3	1	4	2	10
E	1	2	1	1	5

Upper Group (33% Upper)	Subject	Item1	Item2	Item3	Item4
	A	6	6	5	4
	B	4	6	5	3
	Mean	5	6	5	3.5
Lower Group (33% Lower)					
	D	3	1	4	2
	E	1	2	1	1
	Mean	2	1.5	2.5	1.5
Discrimination Index		3	4.5	2.5	2

Table13. Discrimination Index Using Item Analysis Software

ItemMan

Discrimination Index

Project	Code	Description	Upper Group	Lower Group	Index
01	1	Boys prefer to play computer games	5.00	2.00	3.00
01	2	Girls prefer spend more time using	6.00	1.50	4.50
01	3	Friendster are only for young people	5.00	2.00	2.50
01	4	Internet chat is a waste of time	3.50	1.50	2.00

Frequency Distribution

The frequencies per item using SPSS and the software for item analysis were the same as shown in tables 14 and 15 respectively.

Table 14 . Frequency Distribution Using SPSS

Item 1		**2**		**3**		**4**	
Valid	Freq	Valid	Freq	Valid	Freq	Valid	Freq
1	1	1	1	1	1	1	1
3	1	2	1	4	2	2	2
4	2	4	1	5	2	3	1
6	1	6	2			4	1
Total	5		5		5		5

Table 15. Frequency Distribution Using Item Analysis Software

ItemMAN

Frequency Distribution

Project	Item Code	Response	Frequency
01	1	1	1.00
01	1	3	1.00
01	1	4	2.00
01	1	6	1.00
01	2	1	1.00
01	2	2	1.00
01	2	4	1.00
01	2	6	2.00
01	3	1	1.00
01	3	4	2.00
01	3	5	2.00
01	4	1	1.00
01	4	2	2.00
01	4	3	1.00
01	4	4	1.00

Detailed Systems Design

The system has three files to create, namely item, response, and project. The information contained in these files are checked for errors. Statistics computed is composed of total scores per respondents, mean, frequency distribution, standard deviation, correlation coefficient, discrimination index, and alpha coefficient for total scale.

The main interface consists of four modules. They are as follows:

Figure 1. Main Menu

The main program opens with the main menu where four icons are posted from left to right. The leftmost icon with the image of a man represents the module for encoding the respondents' data. The second icon with paper and bubbles image is where data for item statements are maintained. The notebook icon handles the data about the project. Lastly, the icon represented in main interface as cogs is where statistics are computed and reports are generated.

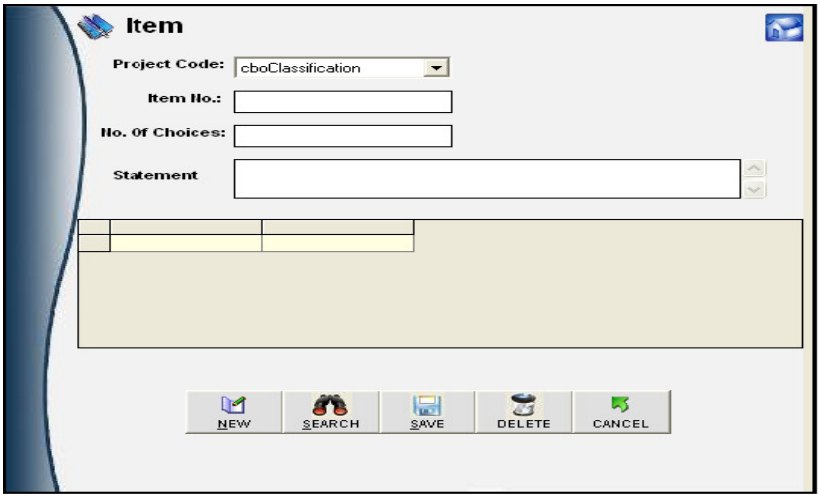

Figure 2. Item Statements Module

The first module (Figure 2) handles the data entry for item statements. It can create new item statements, delete, search and edit items. The data grid at the bottom of the interface shows all items when search is made for a particular project.

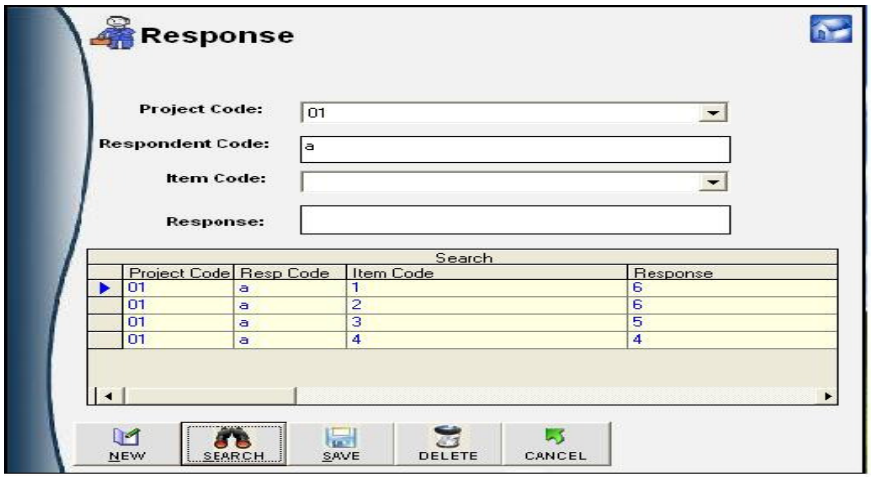

Figure 3. Respondent Module

Responses to each item statements are created, deleted, edited, or added in the second module, the respondent module (Figure 3). The search command button is to locate item for editing or deletion. Data that are needed in this module are the project code, the subject or respondent code, item code, and the answer to the particular item. All responses are displayed in the grid located at the bottom of the module. Data validations are made for project code, respondent code, and item code.

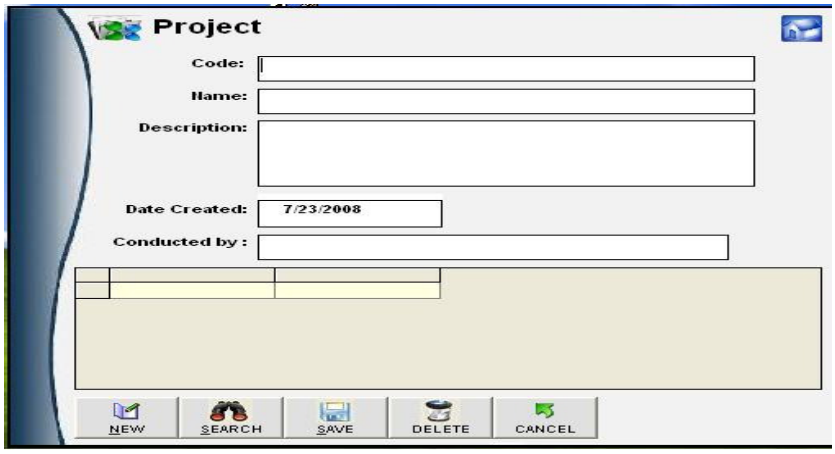

Figure 4. Project Module

Particulars about the project or he scale are managed in the project module. Data about the name of the project, description, date made, and who conducted it are found in the third module (Figure 4). The project profile is entered first before item statements are encoded and processed.

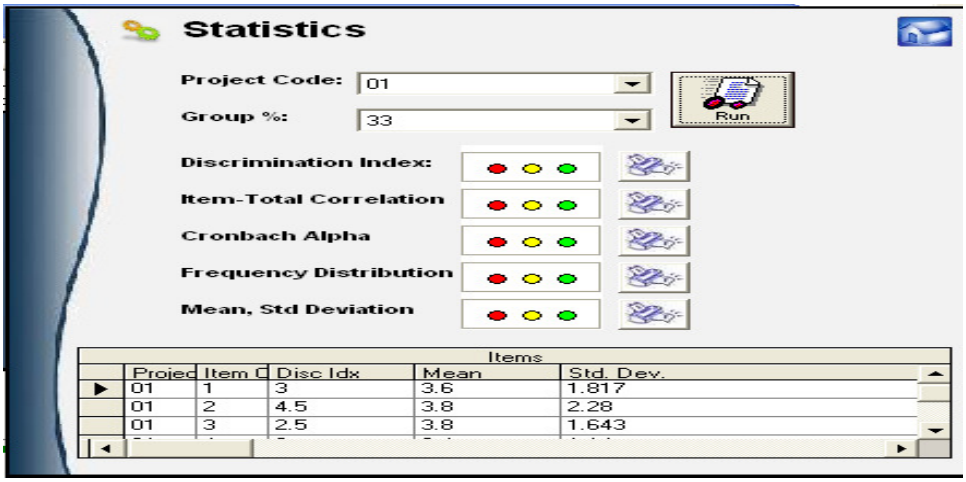

Figure 5. Statistics Module

The last module (Figure 5) is the statistics and report generation module. It computes and prints the various statistical tests for item analysis. Before the statistics are computed for the scale, percentage of the group, shown here as "Group %," is selected. The Group % field prompts for the number of cases to be selected for the high and low groupings for the computation of Discrimination Index. Indicator lights are flashed for each statistics to show that they are being processed. The box at the right of each statistic is for report generation. Below the indicator lights is a grid which shows the computed value of each item such as discrimination index, mean, and standard deviation. The Cronbach Alpha and the Frequency Distribution are shown separately by clicking their respective printer-icon boxes.

Summary

Item analysis is a necessary set of procedures for teachers and researchers. The software developed in this study on item analysis and item banking facilitates the creation and evaluation of items and generates item pool for educator to enhance their classroom practices. The valuable insights derived from the item analysis will enhance the delivery of educational content by way of effective and immediate feedback on the quality of the items and the scale as a whole.

The software is an ideal tool intended to help Filipino teachers and researchers generate quality education through consistent and reliable items to assess their students.

Endnote: Copies of the software can be downloaded with permission through this email address: hjmanaligod@yahoo.com

References

Andersen, L.W. (1981). *Assessing Affective Characteristics in the Schools.* Boston, MA : Allyn and Bacon.

Borich, G. and Kubiszyn, T. (2000). *Educational Testing and Measurement : Classroom Application & Practice* (Sixth ed.) New York: John Wiley & Sons, Inc.

Connel, J. (2003). *Beginning Visual Basic 6 Database Programming*. Canada: Apress.

Ferguson, G.A. (1981). Statistical Analysis in Psychology and Education. (Fifth ed.) New York: McGraw-Hill, Inc.

Ferguson, LW. (1941). *A Study of the Likert Technique of Attitude Scale Construction.* [Internet version]. Retrieved January 28, 2009, from www.brocku.ca/MeadProject/ /sup/ Ferguson_941.html.

Fishbein, M. and Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior : An Introduction to Theory and Research.* Reading, Massachussettes : Addison and Wesley.

Frankel, J.R. and Wallen, N.W. (1993). *How To Design and Evaluate Research in Education* (2nd Ed.) . New York: McGraw-Hill, Inc.

Fox, J.A., and Levin, J. (2003). *Elementary Statistics in Social Research* (9th Ed). Boston: Pearson Education Group.

Gronlund, N.E. and Linn, R.L. (1990). *Measurement and Evaluation in Teaching*. (6th Ed.).New York: MacMillan Publishing Company.

Hogan, T.P. (2007). *Educational Assessment : A Practical Introduction*. New York: John Wiley & Sons, Inc.

Kerlinger , F.N. (1973). *Foundations of Behavioral Researcher (2nd Ed)*. New York.

Mueller, D.J. (1989). Measuring *Social Attitudes : A Handbook for Researcher and Practitioners.* New York: Teacher College Press.

Oppenheim, A.N. (1986). *Questionnaire Design and Attitude Measurement*. London: Heineman Educational Books, Ltd.

Petroutsos, E. (2000). *Mastering Database Programming with Visual Basic 6.* California: Sybex.

Popham, W.J. Modern (2000). *Educational Measurement : Practical Guidelines for Educational Leaders* (3rd Ed). MA: Allyn & Bacon.