Using SAS for Classical Item Analysis and Option Analysis

Chong Ho Yu, Ph.D., Arizona State University, Tempe, AZ
Josephine Wai-chi Wong, Arizona State University, Tempe, AZ

ABSTRACT

In spite of the growing popularity of the item response theory (IRT), classical item analysis (CIA) is still frequently employed by psychometricians and teachers for its conceptual and computational simplicity. This article will introduce how SAS can be applied to CIA such as computing *p*-values, discriminations, point biserial correlations, and logits. In addition, option analysis, which is helpful to both IRT and classical analysis, will be discussed. The purpose of option analysis is to examine clarity and plausibility of distracters in multiple-choice items.

INTRODUCTION

Although today the item response theory (IRT) is arguably the pre-dominant measurement model, classical item analysis (CIA) is still frequently employed by psychometricians, test developers, and teachers for a number of reasons. First, concepts of CIA are simpler than that of its IRT counterpart. Users without a strong statistical background could easily interpret the results without going through a steep learning curve. Second, CIA could be computed by many popular statistical software programs, including SAS, while IRT necessitates use of specialized software packages such as *Bilog*, *Winsteps*, *Multilog*, *RUMM*, *Parscale*, and *Conquest*. Several software packages on the market, such as *Iteman* and *Bilog* (Phase 1 output) are capable of computing CIA; nevertheless, SAS could also be used for producing comparable output. In this article, CIA will be explained conceptually and procedurally. In addition, option analysis, which could be helpful to both IRT and classical analysis, will be discussed. The purpose of option analysis is to examine clarity and plausibility of distracters in multiple-choice items.

WHAT IS CLASSICAL ITEM ANALYSIS?

Classical Item Analysis, also known as classical test theory (Novick, 1966; Lord & Novick, 1968), has been employed by researchers for several decades. Like most other classical statistics, CIA aims to make inferences from a sample to a hypothetical population such as estimating the true parameter in that population. In addition, CIA is based on the true score theory, which views the observed score as a combination of the true score and error. The true score reflects what the examinee actually knows, but it is always contaminated by different sources of errors. In this sense, test reliability is expressed as a ratio between the true score variance and the observed score variance. Since all sample statistics from CIA are estimates of population parameters, CIA tends to be sample-dependent. In other words, item attributes may depend on examinee attributes, and vice versa. Discussion of concepts and computational procedures of test reliability could be found in Yu (2001). In this article the focus is placed on item difficulty, item discrimination, Point-biserial, and logit.

ITEM DIFFICULTY AND ITEM DISCRIMINATION

One of the major statistics in CIA is the item difficulty, which is expressed in terms of the pass rate. If the score is dichotomous, the possible values of the pass rate will range from 0 to 1. This pass rate is also known as the *p*-value. In SAS, PROC MEANS or PROC SUMMARY could be employed to compute the pass rate for each item, depending upon how the data set is structured. For example, if the data set is organized as an N * P matrix, where N is the subject dimension and P is the item dimension, PROC MEANS is definitely appropriate. If the scores are structured in one dimension so that a single variable contains the score for each item by each subject, as shown in Table 1, then PROC SUMMARY is a better way of computing the pass rate.

Table 1. Scores structured in one dimension

Subject	Item	Score
Subject 1	Item 1	1
Subject 1	Item 2	0
Subject 1	Item 3	1
Subject 2	Item 1	1
Subject 2	Item 2	1
Subject 2	Item 3	0

The following is an example of the usage of PROC SUMMARY:

```
proc summary; class item; var score;
    output out=filename mean=passrate std=std
n=samplesize;
```

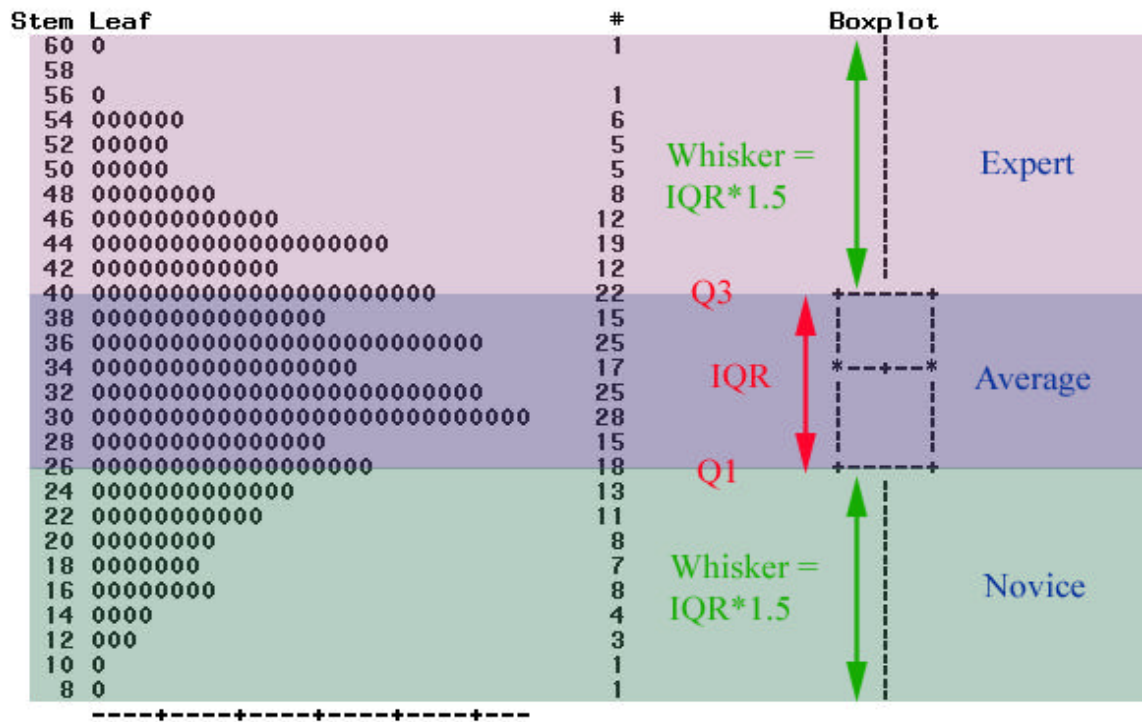The preceding procedure will return the pass rate of each item as depicted in Table 2.

Table 2. Item difficulty in terms of pass rate.

Item	Pass rate (p-value)	Item difficulty
Item 1	0.90	Easy
Item 2	0.50	Just right
Item 3	0.10	Difficult

However, the above item difficulty does not tell us how different types of examinees answered these questions. To be specific, if many people failed to answer particular items correctly, are those people novices or experts? Could those items discriminate examinees who have high proficiency of the subject matter from those who don't?

To obtain the information regarding discrimination, we must first classify examinees into three groups: novice, expert, and neither. There are numerous ways to perform this kind of classification, but none is universally accepted. For example, the software package *Iteman* considers the top 20 percent of subjects as experts and the lowest 20 percent as novices. Kelley (1939) suggest that using the upper and lower 27% is a robust way for computing discrimination. But some accepts the top and bottom 30%.

Figure 1. Stem/leaf plot and Boxplot.


```
Stem Leaf                                          #    Boxplot
60 0                                               1
58
56 0                                               1
54 000000                                          6     Whisker =        Expert
52 00000                                           5     IQR*1.5
50 00000                                           5
48 00000000                                        8
46 000000000000                                    12
44 0000000000000000000                             19
42 000000000000                                    12
40 0000000000000000000000                          22   Q3
38 000000000000000                                 15
36 0000000000000000000000000                       25
34 00000000000000000                               17   IQR          Average
32 0000000000000000000000000                       25
30 0000000000000000000000000000                     28
28 000000000000000                                 15
26 000000000000000000                              18   Q1
24 0000000000000                                   13
22 00000000000                                     11
20 00000000                                        8
18 0000000                                         7     Whisker =       Novice
16 00000000                                        8     IQR*1.5
14 0000                                            4
12 000                                             3
10 0                                               1
 8 0                                               1
   ----+----+----+----+----+----+---
```

I adopt the method of putting subjects above the third quartile (Q3) in the expert group while assigning subjects below the first quartile (Q1) to the novice group. Subjects within the Inter-Quartile Range (IQR = Q3-Q1) are treated as average (neither expert nor novice). In SAS, one can use PROC UNIVARIATE PLOT to get this information as shown in Figure 1.

Figure 1 shows a stem/leaf plot and a box/whister plot (Tukey, 1977), also known as boxplot. Basically, a stem/leaf plot is a horizontal histogram. This discussion will concentrate on the boxplot. In the boxplot, the "box" includes subjects who are between Q3 and Q1. This distance is known as the Inter-Quartile Range (IQR). In this analysis, examinees whose scores fall along this range are treated as average students. The upper edge of the box is Q3, and subjects whose score is above this line are treated as "experts." The lower edge of the box is called Q1, and examinees whose score below this line are regarded as "novices." The steam/leaf and box/whisker plots are helpful in visualizing the overall score distribution and detecting outliers. The two "tails" attached to the box are called "whiskers," which are constructed by multiplying IQR by 1.5. Scores located outside the whiskers are viewed as outliers. In this example no outliers are spotted. Although the stem/leaf plot and the box/ whisker plots are useful in visualization, it may be difficult to see the exact values of Q1 and Q3 from the plots.

Fortunately, PROC UNIVARIATE PLOT also produces text-based reports as shown in Table 3. Table 3 indicates that the cut-off for distinguishing expert from average is 41 and the cut-off for average and novice is 27.

Table 3. Quantile Information.

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	60
99%	55
95%	50
90%	47
75% Q3	41
50% Median	34
25% Q1	27
10%	21
5%	17
1%	12
0% Min	8

After assigning examinees into different groups according to their competency, we can compute the pass rate by group, as shown in the following example of SAS code. The item discrimination is defined as the *p*-value of the expert group subtracted from that of the novice group.

```
data two; set one;
    if totalscore => 41 then group = "expert";
    else if totalscore <=27 then  group ="novice";
    /* Insert codes here to compute the pass rate
of each item by group.It depends on how the data set
is structured */
    discrimination = highmean - lowmean;
```

The preceding procedure will yield results as shown in Table 4.

Table 4 Item discrimination table.

Item	Expert group pass rate	Novice group Pass rate	Item discrimination	Judgment
Item 4	0.90	0.10	+0.80	High
Item 5	0.70	0.60	+0.10	Low
Item 6	0.10	0.10	0.00	No
Item 7	0.90	0.90	0.00	No
Item 8	0.30	0.70	-0.40	Negative

Table 4 shows that Item 4 has a high discrimination while Item 5 has a low one. Both Item 5 and Item 6 have zero discrimination but the causes may be totally different. Item 6 seems to be extremely difficult and thus regardless of what the ability level is, the probability of giving the correct answer is low. Item 7 is exactly opposite. This question is extremely easy, and thus no matter how much or how little one knows, the probability of answering it correctly is very high. Item 8 is very problematic because experts tend to give the wrong answer while novices tend to give the right answer. There are a number of possible factors: (a) the key is incorrect, (b) the wording of the question and multiple-choice options is confusing, (c) the item is located near the end of a speed test, and the difference is due to a random fluctuation (guessing). The test developer could not rely on the numbers alone to determine the cause, and thus option analysis is necessary. Option analysis will be discussed in a later section.

Figure 2. Bar chart of item score by group.
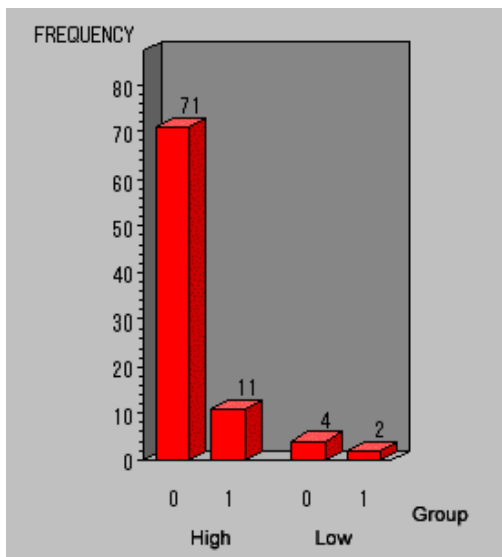
However, when the size of high group or the low group is small, the item discrimination should not be trusted without reservation. For example, even if in one item the high group mean is .13 and the low group mean is .33, and thus the item discrimination is -.20, it does not necessarily mean that this item favors novices. When one examines the bar chart by group, one could tell that this impression is misled by the samll sample size in the novice group (Figure 2). To avoid this type of misinterpretation, besides analyzing the numeric output, it is advised that the researcher also examinee the frequency of the two groups by a bar chart by group. Since the same SAS code will be reused for many items, writing the code as a macro is more efficient:

```
%macro chartbar(itemid);
PROC gchart;
  vbar3d &itemid / group=group discrete type=freq
freq;
run;
%mend chartbar;
```

LOGIT

In CIA, the logit is another common statistic. The purpose of using the logit is to avoid misinterpretation of results based upon raw percentages. The difference between two items in terms of difficulty near the midpoint of the test (e.g. 50% and 55%) does not equal to the gap as two items at the top (e.g. 95% and 100%) or at the bottom (5% and 10%). Take weight reduction as a metaphor: It is easier for me to reduce my weight from 150 lbs to 125 lbs, but it is much more difficult to trim my weight from 125 lbs to 100 lbs. However, people routinely misperceive that distances in raw percentages are comparable (Bond & Fox, 2001). As a remedy, the logit is used to convert the raw score to its natural logarithm. In this approach, distances from different ranges in the scale are comparable. Logit is the natural log of the odds ratio, which is the ratio of the probability of success and the probability of failure. For example, if 70% of the examinees answered the item correctly, the odd of getting the right answer is 7- to 30. In algebraic terms it is expressed as:

Logit = Log(Passrate/ 1 – Passrate)

In SAS the logit can be computed by the above equation since *log* is a built-in function in SAS. Some programs such as *Bilog* divide the logit by 1.7 in order to make the Logit model and the Probit model comparable. You may notice that the odds ratio could also be found in the results of a logistic regression (see Figure 3). Although the two contexts are different, the concepts are essentially the same. In logistic regression the researcher is interested in learning whether the regressors could predict a dichotomous outcome (e.g. pass/fail). In this case the odds ratio indicates the "odds" of passing and failing. By the same token, in item analysis the test developer is concerned with the odds of answering this item correctly and incorrectly. The logit tells us this information.

Figure 3. Odds ratio in logistic regression

```
                      The LOGISTIC Procedure

               Analysis of Maximum Likelihood Estimates

                Parameter Standard    Wald                 Standardized    Odds
  Variable   DF  Estimate   Error  Chi-Square Pr > ChiSq     Estimate      Ratio

  Intercept  1   -0.6931   0.4629    2.2421     0.1343
  male       1    1.5404   0.6726    5.2455     0.0220          0.4298      4.667


         Association of Predicted Probabilities and Observed Responses

            Percent Concordant    46.7    Somers' D    0.367
            Percent Discordant    10.0    Gamma        0.647
            Percent Tied          43.3    Tau-a        0.188
            Pairs                  420    c            0.683
```
.

POINT-BISERIAL

In CIA the test developer cares about not only individual items, but the test as a whole, and therefore item-total correlation is an important piece of information. To be specific, if the response pattern of the item does not conform to all other items, this question may be problematic. Besides the reliability measure in terms of internal consistency, the point-biserial correlation coefficient, which is the correlation coefficient between the item and the total, is also an indictor for this kind of diagnosis. Like the Pearson coefficient, the point-biserial is also a product-moment correlation coefficient. However, the Pearson coefficient is used for computing the relationship between two continuous-scaled variables, whereas the point-biserial is applicable to the relationship between one binary variable and one continuous-scaled variable. In the case of CIA, the individual item is a dichotomous variable, in which only 1 or 0 is a possible value, and the total is a continuous-scaled variable, in which scores of all items are summed.

Table 5. Frequency table showing selection of options.

Option	Label	Count	%	Bar
Item 7: Which group is the biggest threat to world peace?
A	Federation	33	41.25	
B	Vulcan	3	3.75	
C	The Borg	30	37.5	
D	Ferengi	11	13.75	
E	Q	3	3.75	

Usually values of the point-biserial lie between -1 and +1. But in CIA it is unlikely to exceed 0.75 or to fall below -0.10 (Wilnut, 1975). When the point-biserial is negative, it could be caused by using a wrong key or putting ambiguous words into the item. Since point-biserial is a type of product moment correlation coefficient, one can use PROC CORR to compute it.

It is important to note that biserial and point-biserial are conceptually and computationally different though their names look similar. Unlike the point-biserial, the biserial is not a product-moment correlation; it is less likely to be influenced by the item difficulty (du Toit, 2003). Moreover, the biserial correlation may be systematically larger than its point-biserial counterpart (Crocker & Algina, 1986).

OPTION ANALYSIS

The preceding statistics are necessary, but insufficient for diagnosing a test. The test developer must pay close attention to how examinees select different options in order to enhance the test. Let's look at Table 5. In Item 7 the correct answer is "C" and the pass rate is acceptable (0.375). Thus, by looking at the pass rate alone, one may not notice that this question needs revision. As you may notice, 41.25% examinees selected "A" as their answer and viewed the Federation as a bigger threat. Perhaps option A could arguably be an acceptable answer. To avoid confusion, the test developer might consider either to drop option A or to replace it with another distracter.

Although one can use PROC FREQ to obtain a frequency table of responses for each item, it is tedious to check each fequency table against the key especially when a test is composed of many questions. One way to make this task more efficient is to compare the pass rate and the percentage of the most often chosen option for each question. If the two numbers are different, the program will put a flag on that item. For example, in Item 7 the pass rate is 37.5% but the proportion of the most popular option is 41.25%. Thus this item will be marked for further examination. On the other hand, if the two numbers match, it means that the right answer is selected by most examinees. The SAS code in the next page is an example for this type of detection. It is assumed that the data set of the raw data is structured in a way that a single variable contains responses of each item by each subject.

Table 6. Frequency table showing unused options.

Option	Label	Frequency	Percent	Bar
Item 8: Question: Which of the following vehicles is the most dependable?				
A *	Mercedes Benz	77	96.25	
B	Dodge Neon	3	3.75	
C	Kia	0	0.00	
D	Yugo	0	0.00	

```sas
/* Concatenate the itemid and the responses */
data rawdata; set rawdata;
  temp=item||resp;
  proc sort data=rawdata; by itemid;
run;
/* output the frequency tables of responses for all
items to a comma delimited file (csv) */
ods csvall file="freq.csv";
proc freq data=rawdata; tables temp;by itemid;run;
ods csvall close; quit;
/* import the csv file into SAS */
PROC IMPORT OUT= WORK.READCSV
  DATAFILE= "freq.csv"
  DBMS=CSV REPLACE;
  GETNAMES=NO;
  DATAROW=8;
RUN;
/* Read the imported file, separate the itemid from
the raw responses */
data freq; set readcsv;
  itemid = substr(var1,1,5);
  response=substr(var1,6,10);
  response=compress(response);

  itemid=compress(itemid);
  rename var2=frequency;
  rename var3=percent;
  rename var4=cf;
  rename var5=cp;
  drop var1;
run;
/* Clean up the blank lines */
data freq; set freq;
  if cf =" " then delete;
  if itemid = " " then delete;
  if response = " " then delete;
run;
/* find the most often chosen option */
data most; set freq;
  most=percent/100;
  proc sort nodupkey; by itemid descending most;
run;
/* Merge the dataset "most" with the dataset
carrying the pass rate, flag the item if the two
numbers do not match */
```

Now let's examine another type of problem that could be observed in option analysis (see Table 6). Item 8 appears to be an easy question because 96.25% of examinees chose the right answer. We can see that the problem is due to implausible distracters. No person picked "Kia" and "Yugo"; only 3.75% chose "Dodge Neon." This is probably because the test was administered in America and Germany (Mercedes Benz owns 50% of Chrysler). If Options BD are "Lexus," "BMW," and "Cadillac," it will require some knowledge to determine which choice is the correct answer. Needless to say, comparing a Mercedes with Neon, Kia and Yugo would definitely give the right answer away. Although one could use PROC FREQ to list all responses in a table for spotting unused options and implausible distracters, again, it is tedious when the test is long. A more efficient way is to let SAS count the number of used options for each item, and then find the difference between the number of available options and the number of used options. The following is an example of how to use SAS for counting used options:

```sas
/* Count how many options are used. Set the counter
to 0 when an unique itemid is found. Increment the
counter until seeing a new itemid */
data freq; set freq; by itemid;
  if first.itemid then
  f=0;
  f+1;
  proc sort; by descending f itemid;
  proc sort nodupkey; by itemid;
run;
```

SUMMARY

Although CIA is fairly simple in terms of conceptualization and computation, it is crucial to emphasize that the statistics, including item difficulty, item discrimination, logit, and point-biserial, are not sufficient in conducting test diagnosis. On some occasions an acceptable pass rate might mask the problem of a misleading option. Further, when some anomaly occurs, the number alone may not be informative enough to find the root cause of the problem. For example, if negative item discrimination or a negative point-biserial is found, it could be due to a wrong key or some other reasons. CIA and option analysis must be applied together for a thorough analysis. The SAS code introduced in this article may alleviate some repetitive steps.

REFERENCES

Bond, T. G. & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. New Jersey: Laerence Erlbaum Associates.

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

Du Tiot, M. (Ed.). (2003). IRT from SSI. IL: Scientific Software International.

Novick, M. R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, 1-18.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Kelley, T. L. (1939). Seletion for upper and lower groups for the validation of test items. Journal of

Educational Psychology, 30, 17-24.

    Tukey, J. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

    Wilmut, J. (1975). Objective test analysis: Some criteria for item selection. Research in Education, 13, 27-56.

    Yu, C. H. (2001). An Introduction to computing and interpreting Cronbach Coefficient Alpha in SAS. Proceedings of 26th SAS User Group International Conference.

CONTACT INFORMATION
Your comments and questions are valued and encouraged.

Contact the author at:

    Chong Ho Yu, Ph.D.
    Josephine Wong
    PO Box 612
    Tempe AZ 85280
    Email: asumain@yahoo.com.hk
    Website: http://seamonkey.ed.asu.edu/~alex/