

Using the SAS® System to Examine the Agreement between Two Programs That Score Surveys Using Samejima's Graded Response Model

Jim Penny, Center for Creative Leadership, Greensboro, NC
S. Bartholomew Craig, Center for Creative Leadership, Greensboro, NC

ABSTRACT

In the Spring of 1998, researchers at the Center for Creative Leadership chose to use the Graded Response Model by Samejima as part of a project to assess the behavioral changes that are seen in the participants in the leadership development programs of the Center. It was determined that the existing DOS and Windows software would not function with sufficient reliability within the extant computing network of the Center, and a contract was awarded to a vendor for the development of scoring program that would operate reliably at the Center. As a part of the contract specifications, the Item Response Theory parameter estimates of ability and standard errors that were produced by the new computing module had to agree with those estimates produced by the existing software. As one might expect, it was difficult to demonstrate exact convergence of the two sets of estimates. This paper presents the statistical process by which the case for sufficient convergence was made.

INTRODUCTION

In the summer of 1998, the Center for Creative Leadership chose to extend its survey item analyses using the Graded Response Model by Samejima (Samejima, 1969 & 1979) for the purpose of scoring an instrument known as *Reflections* that is designed to measure behavioral changes in those managers who participate in the leadership development programs that are provided by the Center. Preliminary studies had shown that the model, as implemented by the software MultiLog (Thissen, 1991), provided important information regarding the characteristics of the survey items. However, it was also determined that MultiLog would not integrate well into the Center's extant computing network, and that any effort to accomplish such an integration, although certainly do-able, could result in a scoring system that would likely compromise the integrity of the network, resulting in not only the on-going expense of an unreliable computing system but also the liability of statistical analyses that could be in error.

For such reasons, the Center choose to contract with an independent software developer to create an Item Response Theory (IRT) (Lord, 1980) module that would score subject responses using the Samejima Graded Response Model and that would also function seamlessly within the existing computing network of the Center. The Center wrote into the contract with the vendor the following requirements that define the level of congruency that would be needed between (1) the estimates of IRT latent trait (theta) and (2) the estimates of the standard errors, as computed by both MultiLog and the new module. First, for any estimate of theta or standard error, an absolute difference of 0.02 or less was acceptable without explanation. Second, absolute differences between 0.02 and 0.05 may be acceptable at the discretion of the Center. Such discretion would be based on explanations such as improved numerical methods or better informed a priori assumptions, all of which were to be provided by the vendor. Third, absolute differences greater than 0.05 required compelling justification by the vendor.

A BRIEF DESCRIPTION OF ITEM RESPONSE THEORY

A discussion of the essential features of Item Response Theory (IRT) (Lord, 1980) is likely to facilitate the reading of this paper. IRT is a form of non-linear factor analysis that attempts to uncover an underlying latent trait from a set of item responses. In this brief discussion, we will assume dichotomous items; that is, items on which the responses are 0

and 1. Item response theory posits that an examinee's response is a probabilistic function of the examinee's ability in conjunction with the discrimination, difficulty, and guess-ability of the item, and that the values of the item parameters are independent of the respondent. Using the logistic ogive, we can model this function as

$$P(\theta) = c + \frac{1 - c}{1 + e^{(-1.7a(\theta - b))}}$$

where θ is the ability of the examinee, a is the discrimination index of the item, b is the difficulty parameter of the item, and c is called the pseudo-guessing parameter. The value 1.7 is a scaling parameter that renders the logistic ogive numerically similar to the normal ogive. This model is also known as the 3-parameter logistic model, or 3PL.

The distribution of θ in the population is often considered a normal deviate (Warm, 1978), and the distributions of a , b , and c are chosen to build a test of given characteristics. For instance, one would likely want values of the item difficulty parameter that span a particular range of subject ability. This is possible because the values of θ and b are linked in that they are considered to be on the same metric. Thissen and Wainer (1982) indicate that the sample size requirements for the 3-parameter model can be substantial owing to the reduced density of subjects at lower end of the ability distribution where the c -parameter is an important part of the probability function.

Figure 1 shows this model for $a=1$, $b=0$, and $c=0.20$, which are reasonable item parameter values for a 5-part multiple choice item of average difficulty and average discrimination.

If one chooses to eliminate guessing from the model, then the above equation becomes

$$P(\theta) = \frac{1}{1 + e^{(-1.7a(\theta - b))}}$$

which is known as the 2-parameter logistic model, or 2PL. This model can be reduced further if one has reason to assume that the item can be modeled with $a=1$. Such a model can be written as

$$P(\theta) = \frac{1}{1 + e^{(-1.7(\theta - b))}}$$

which is known as the 1-parameter model and also the Rasch model (Rasch, 1960). The sample size requirements for the Rasch model as substantially reduced from the 2- and 3-PL, and are generally on the order of 1000, depending on exactly how far from 0 the value of the difficulty parameter is.

ASSUMPTIONS OF IRT

There are four basic assumption in item response theory. The first of these is that if an examinee knows the correct answer, then she will so respond. Absent such an assumption, there is little reason for testing. The second assumption is that the item response function follows the normal ogive which is modeled historically as the logistic ogive and that the resulting function accurately describes the monotone cognitive response of

the examinee. This assumption delayed the theoretical development of IRT for some 10 years owing to the lack of computing power in the 1950s and 1960s. Lord (1965) finally demonstrated the validity of this assumption.

The third assumption of item response theory is of local independence. Hambleton, Swaminathan, and Rogers (1991) define local independence as "...when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent. In other words, after taking examinees' abilities into account, no relationship exists between examinees' responses to test items. This set of abilities represents the complete latent space." (p.10). Formally, local independence can be represented mathematically by

$$\Pr ob(U_1, U_2, \dots, U_n | \theta) = P(U_1)P(U_2) \dots P(U_n) \\ = \prod_{i=1}^n P(U_i | \theta)$$

where U_i represents the response to item i and θ represents the ability of the examinee. The assumption of local independence may be violated in reading comprehension tests where sets of items are keyed to the same passages. Local independence is also violated when tests are speeded in that the latter items are not attempted because of the existence of the prior items.

The fourth assumption of item response theory is unidimensionality which means that the items on the test measure only one ability, attribute or area of knowledge. It should be noted that unidimensionality is a sufficient condition for local independence, but that local independence is not a sufficient condition for unidimensionality (Warm, 1978; Hambleton, et al, 1991).

SAMEJIMA'S GRADED RESPONSE MODEL

The Graded Response Model developed by Samejima (1969) to describe item for which a respondent chooses one of an ordered set of responses. For instance, a Likert-type item using a 5-point scale of 1=Strongly Disagree to 5=Strongly Agree could be so modeled, assuming that the four basic assumptions of IRT are met. The Graded Response Model posits an item response function for each point on the scale according to

$$P(x = k) = \frac{1}{1 + e^{(-a(\theta - b_{k-1}))}} - \frac{1}{1 + e^{(-a(\theta - b_k))}} \\ = P^*(k) - P^*(k + 1)$$

in which a is the discrimination parameter and b_{k-1} is the threshold parameter. $P^*(k)$ is the item response function that describes the probability that a response is in category k . In this model, the discrimination parameter is considered constant for all categories of k , and the threshold parameter, b_{k-1} , which is analogous to the item difficulty parameter described above, is the point on the θ -axis where the probability exceeds 50 percent that the response is in the next category. The scaling constant of 1.7 is not generally used in this model, though one could do so with little or no consequence.

A BRIEF DESCRIPTION OF CCL

The Center for Creative Leadership (CCL) is an international, nonprofit educational institution founded in 1970 in Greensboro, NC, with the study of leadership as its single focus. Some of the issues embedded in the on-going study of leadership include (1) The role of leadership in the success of organizations, (2) The factors that contribute to executive success, (3) The transformation of leadership from an individual activity into an organization-wide process and (4) The challenges and conflicts arising from societal and economic trends.

Research at CCL is definitional and fundamental in the sense that it delves into the nature of leadership and into the behaviors that define it. Moreover, CCL research seeks to understand how changes in individuals and organizations can result in a higher quality of, and greater capacity for, leadership. CCL research can also be characterized as applied in that methods are sought that transform research findings into practical techniques that can lead to measurable results for CCL clients. Although the CCL mandate is non-proprietary education, it differs from many academic institutions in that the research is practitioner oriented, focusing on practicing managers and functioning organizations. Action research and longitudinal follow-up that promote learning are emphasized.

A BRIEF DESCRIPTION OF LDP®

The Leadership Development Program® (LDP®) is designed to offer middle-level to upper-level managers an opportunity to stimulate a process of growth that is designed to help them become more successful and more productive in their work and in their personal lives as well as more effective in leading others to achieve higher professional and personal productivity. LDP is one of the effective and popular programs that CCL offers, and is designed from the Center's nearly 30 years of research and experience in individual development.

The training offered in LDP encourages the participants to explore three critical questions. These are (1) How do my co-workers perceive me as an individual, (2) Who am I in relation to the work groups in which I'm involved, and (3) Who am I in relation to my organization?

Post-program impact studies indicate that participants acquire a better understanding of their strengths and weaknesses, improve their ability to give and receive constructive feedback, and develop a personal leadership style. They also learn to see the connection between their individual impact and effectiveness and that of their work group and organization. Post-program feedback regarding behavior change is an integrated part of the program experience, and is designed to promote continued learning.

A BRIEF DESCRIPTION OF REFLECTIONS®

Reflections is a 360 survey retrospective measure of change that has been designed around the developmental changes that are frequently seen in the participants of the Center's educational programs. In this instance, the survey included scales that were appropriate for use with participants from the Leadership Development Program. Retrospective data for participants were collected from direct reports, peers, bosses, and other supervisors at approximately 3 months from the end of the program attended by the participant. Ordinarily, participants had a total of 10 or so raters each.

Raters were asked to rate the participant on 50 individual behaviors in 12 separate areas. Although it is unlikely that a participant would choose to change her behavior on all 12 areas, it is likely that the participant had chosen as many as three areas for special developmental attention. Moreover, the behaviors of all 12 areas are explicitly addressed in the LDP curriculum.

Raters were asked to evaluate the participant on each behavior. A rating of 1 to 9 was given to each behavior as currently exemplified by the participant. Simultaneously, each rater evaluated these behaviors by the participant in the time prior to her attendance in the CCL program. The difference in these two ratings is the perceived change in the behavior of the participant. The Reflections survey has an internal reliability in excess of 0.9 as measured using Cronbach's alpha. The reliability of the individual scales is generally in excess of 0.87.

DESCRIPTION OF THE TEST DATA

A small data set of 1433 subjects was identified by Center staff that was sufficiently large as to insure stable IRT subject ability estimates, yet being sufficiently small as to allow convenient manipulation and repetitive executions during the testing process. These data were chosen to be representative of

the data that are collected by the educational programs of the Center. However, no cases with estimates of the IRT latent trait that exceeded 2.0 in absolute value were included in the study, giving a sample of cases that spanned the approximate middle 95 percent of the expected distribution of the population.

THE LINEAR ASSOCIATION BETWEEN THE ESTIMATES OF IRT ABILITY

Figure 2 shows the scatter plots of the estimates of the IRT ability by the two computing packages. This plot was produced by the following SAS® code.

```
proc gplot data=eaps;
  title1 'Comparison of Latent Trait estimates';
  title2 'Multilog estimates on the Y-axis';
  title3 'New Module estimates on the X-
    axis';
  plot eap*theta;
run;
```

Visually, there appears to be a high degree of linearity between the two estimates. To assess this degree of linear association, the value computed by the new module was regressed onto the value from MultiLog. The code to perform this analysis is as follows.

```
proc glm data=eaps;
  model eap=theta;
  title1 'Regression of estimates of theta from
    New Module';
  title2 'onto those of Multilog';
run;
```

This linear model accounted for 99 percent of the variation between the two estimates ($F=100000$, $df=(1,1431)$, $p=.0001$). The estimate of the slope is .9879, statistically different from 0 ($T=323$, $p=.0001$), and very close to 1. The intercept is not statistically significantly different from 0 ($T=-0.11$, $p=.9121$).

It seems reasonable to infer, then, that the level of agreement between the estimates of the IRT latent trait computed by the new module and by MultiLog is acceptable high. However, when there is a difference, the magnitude will be such that the estimate by the new module exceeds that of MultiLog by 1 to 2 percent.

THE LINEAR ASSOCIATION BETWEEN THE ESTIMATES OF STANDARD ERROR

Although the level of agreement between the estimates of the IRT latent trait is encouraging, the agreement between the estimates of the standard error of the latent trait, plotted in Figure 3, is viewed with less enthusiasm. The SAS® code to create this plot is given as follows.

```
proc gplot data=eaps;
  title1 'Comparison of standard errors';
  title2 'MultiLog estimates on the Y-axis';
  title3 'RJ estimates on the x-axis';
  title4 'Complete dataset';
  plot se*se_b;
run;
```

Here, it is clear that the linearity that was seen with the estimates of ability does not exist in the same degree. Indeed, there appears to be a fanning of data points about the 45-degree line for standard errors that are smaller than 0.5 in magnitude. The results of linear regression indicate that only 35 percent of the variation between these two estimates of standard error is explained by the linear model ($F=756$, $df=(1,1431)$, $p=.0001$). Moreover, the slope is nearly 1.2 and statistically different from 0, or about 20 percent larger than the expected value of 1, whereas the intercept is also statistically different from 0, though it is close to 0. The SAS® code to produce this linear analysis is as follows.

```
proc glm data=eaps;
  model se=se_b;
  title1 'Regression of estimates of theta from
    New Module';
  title2 'onto those of Multilog';
run;
```

THE DEGREE OF ABSOLUTE AGREEMENT BETWEEN ESTIMATES OF IRT LATENT TRAIT

To examine the degree of absolute agreement in the estimates of the IRT ability, a difference was computed by subtracting the estimate provided by the new module from the estimate provided by MultiLog and then taking the absolute value of that difference. The bar chart shown in Figure 4 presents some alarming figures. That is, only 19 percent of the differences in estimates of the latent trait are acceptable without justification, only 21 percent are acceptable with appropriate justification, and an astounding 60 percent of the differences require compelling justification. The SAS® code to produce this bar chart is as follows.

```
proc gchart data=eaps;
  vbar adftheta /midpoints=0 to .25 by .01;
  title1 'Absolute difference in theta
    estimates';
  Title2 'between New Module and MultiLog';
run;
```

Similarly, the absolute difference of the two estimates of standard error were computed. Figure 5 presents a bar chart of these data, and it is clear that an alarming 65 percent of these differences require compelling justification from the vendor in order to be acceptable. The code to create this bar chart follows.

```
proc gchart data=eaps;
  vbar absdifse /midpoints=0 to .25 by .01;
  title1 'Absolute difference in SEs';
  title2 'between New Module and MultiLog';
run;
```

THE DEGREE OF AGREEMENT BETWEEN SAMPLES OF ESTIMATES

Given the integrity of the developer of the new module and the apparent disparity between our two sets of IRT estimates, we chose to explore these differences a bit further. Might it be appropriate to consider the mean of groups of estimates, rather than the individual values, especially in view that the exact definition of "difference" was not specified in the letter of agreement? The code to examine these two differences using a t-test follows.

```
proc means data=eaps mean std n stderr t
  prt;
  var difse diftheta;
  title1 'H0: Mean eq 0';
  title2 'Ha: Mean ne 0';
run;
```

The mean difference in the two estimates of the IRT latent trait is 0.01400 with a standard deviation of 0.09649, and the mean difference of the two estimates of the standard error is 0.006581 with a standard deviation of 0.08978. Both are statistically different from 0 ($T=5.49$, $p=.0001$; and, $T=2.7749$, $p=.0056$, respectively), a statistical result that should be taken with a grain of salt given that the sample size is 1433 for this computation. It is likely that this instance of statistical significance is more a matter of inflated power than it is a matter of practical effect.

THE EXAMINATION OF MEAN DIFFERENCES IN MULTIPLE SUB-SAMPLES

To address this problem of excessive statistical power, the group of 1433 cases was sampled 180 times, creating sub-samples of 35 cases each. The sub-sample size of 35 was in order to maintain a reasonable degree of power. For each sub-sample, the mean of both the difference in estimated IRT latent trait and the difference in the estimates of the standard error was compared to 0. In order to maintain an experiment-wise Type I error rate of .05, a test-wise value of .0165 was used to evaluate the p-values from the individual t-tests of the sample means. This value is computed using the equation

$$\alpha_E = 1 - (1 - \alpha_T)^n$$

where α_E represents the experiment-wise error rate, α_T represents the test-wise error rate, and n represents the number of tests. In the event that the population mean differences are zero, we can expect that approximately 5 percent, or 9, of these tests will exhibit statistical significance.

The following code was used to assign a random number from a uniform distribution to each observation. Then, the observations were sorted by that random number. After the observations were sorted, a group identification number of 1 through 35 was assigned using the integer result of the observation number divided by 35, resulting in 36 group numbering from 0 to 35. The data were then re-sorted by the group identification, and the t-tests of the mean difference in theta and standard errors was computed. This process was repeated 5 times to create 180 sub-samples of the data.

```
data eaps;
  set eaps;
  if _n_ eq 1 then seed = 123456789;
  retain seed;
  call ranuni(seed,key);
  drop seed;
run;
proc sort data=eaps;
  by key;
run;
data eaps;
  set eaps;
  group = int(_n_/35);
run;
proc sort data=eaps;
  by group;
run;
proc means data=eaps mean std n stderr t prt;
  var difse diftheta;
  by group;
  title1 'H0: Mean eq 0';
  title2 'Ha: Mean ne 0';
  title4 'Pass 1';
run;
```

Of the 180 samples, 5 had mean differences in the estimates of the standard error that were statistically different from 0; 6 samples had mean differences in the estimates of the IRT latent trait that were statistically significant. Of these proportions, neither the 5/180 nor the 6/180, is statistically different from the expected proportion of 5 percent ($Z = -1.37$, $p = .1713$ and $Z = -1.03$, $p = .3049$, respectively). Hence, one can argue that, at least at the sample level, the new code is sufficiently congruent with MultiLog as to require no explanation of the existing differences.

PROBLEMS WITH ABSOLUTE CONVERGENCE

The problems with absolute convergence between the two programs are discussed in some detail in the project report from the vendor. In particular, the report cites difficulties in determining (1) the computational algorithm that is used by MultiLog, (2) the operation of the scoring algorithm in areas such as the method of standardizing the estimates of the IRT latent trait that is used by MultiLog, (3) the number and location of quadrature points that are used by MultiLog, (4) the range of permissible values to be used with the estimates that is used by MultiLog, and (5) the procedure for handling estimates that exceed the permissible range that is used by MultiLog.

It seems reasonable, then, that the observed differences between MultiLog and the new IRT module are the result of some difference in the estimation algorithms. For instance, one scoring algorithm that is used by the new module is based on Bock & Mislevy (1982). Although of the same era as MultiLog, this procedure is based on a method of numerical

integration that leaves to the programmer the choice for the number of quadrature points. A choice of more points produces a more refined estimate, though at the expense of additional computing time. The second scoring algorithm that is implemented in the new module is based on Thissen & Steinberg (1984), again of the era of MultiLog. In this second scoring procedure, a histogram approach is used to evaluate the response profile that is presented. Again, the choice of how one constructs the histogram is left to the programmer, where a choice of smaller intervals results in more refined estimates and longer computing times.

DISCUSSION OF CONSEQUENCES

One might ask about the consequences of accepting the estimates from the new module in lieu of the estimates from MultiLog. That is, how does the use of the estimates that are provided by the new module contrast with the use of those estimates that are from MultiLog? What is the impact of this choice? The estimates of the IRT latent trait that are given by the new module exceed those from MultiLog by 1 or 2 percent, so using these larger estimates will make the changes in the behavior of participants appear a small bit greater. Moreover, the estimates of the standard error from the new module are about 20 percent smaller. Hence, the slightly larger change score in conjunction with the smaller standard error, both a result of using the new module, will likely produce more instances of statistically significant change in the behavior of the participants, resulting in measures of program impact that are greater than they would be had the IRT parameters been estimated using MultiLog.

REFERENCES

- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.
- Lord, F. M. (1965). An empirical study of item-test regression. *Psychometrika*, 30, 373-376.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institute.
- Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
- Samejima, F. (1979). *A new family of models for the multiple choice item*. Office of Naval Research Report 79-4. Knoxville, TN: University of Tennessee.
- Thissen, D. (1991). *MULTILOG User's Guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software, Inc.
- Thissen, D., & Steinberg, L. (1984). A model for multiple choice items. *Psychometrika*, 49, 501-519.
- Thissen, D., & Wainer, H. (1985). Some standard errors in item response theory. *Psychometrika*, 47, 397-412.
- Warm, T.A. (1978). *A primer of item response theory* (Tech. Rep. No. 941078). Washington, DC: U. S. Coast Guard.

CONTACT INFORMATION

Jim Penny
Center for Creative Leadership
One Leadership Place
Greensboro, NC 27438-6300
Work: 336-286-4442
Fax: 336-286-4434
Email: pennyj@leaders.ccl.org

Figure 1

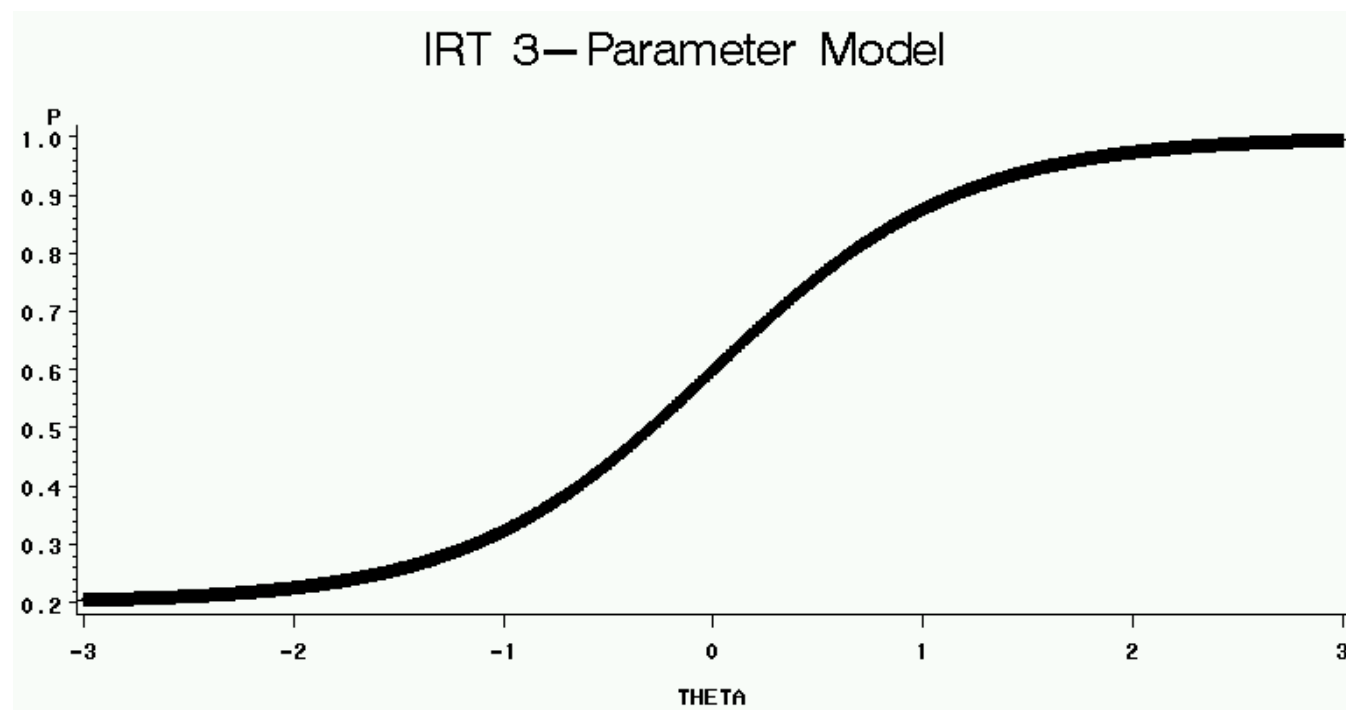


Figure 2

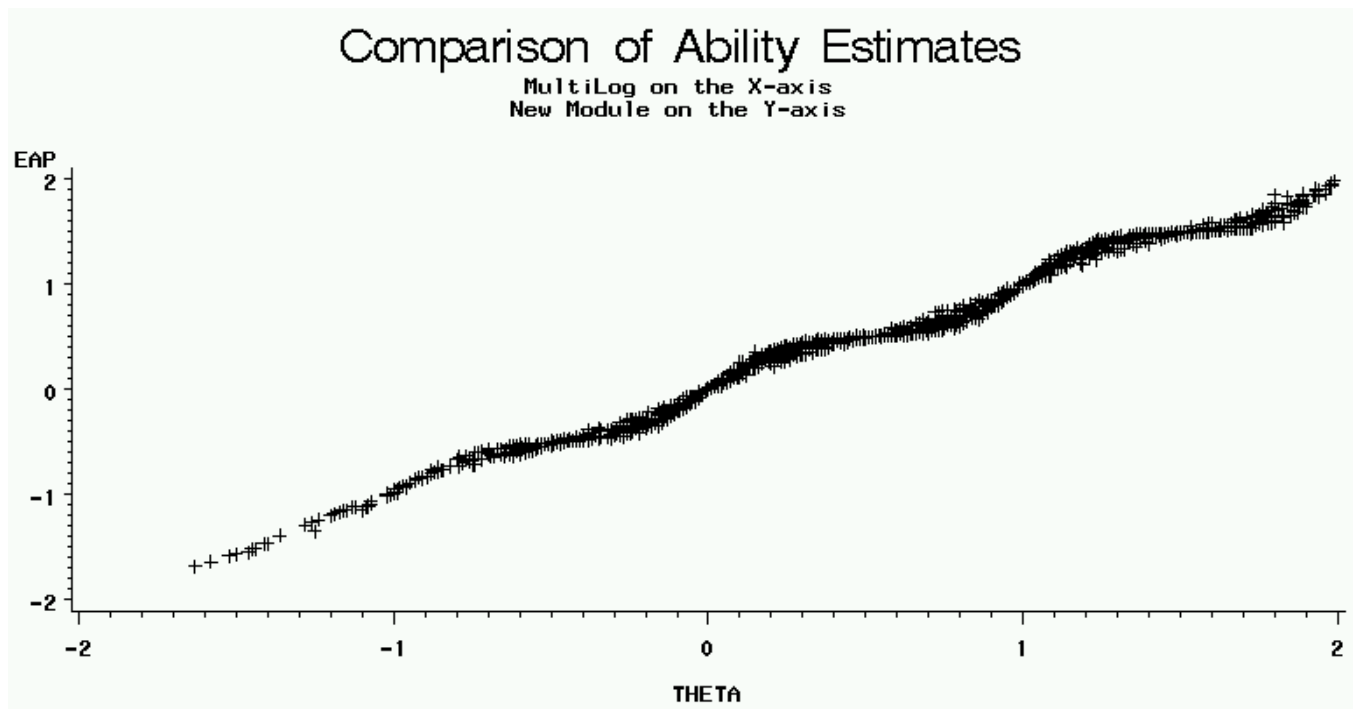


Figure 3

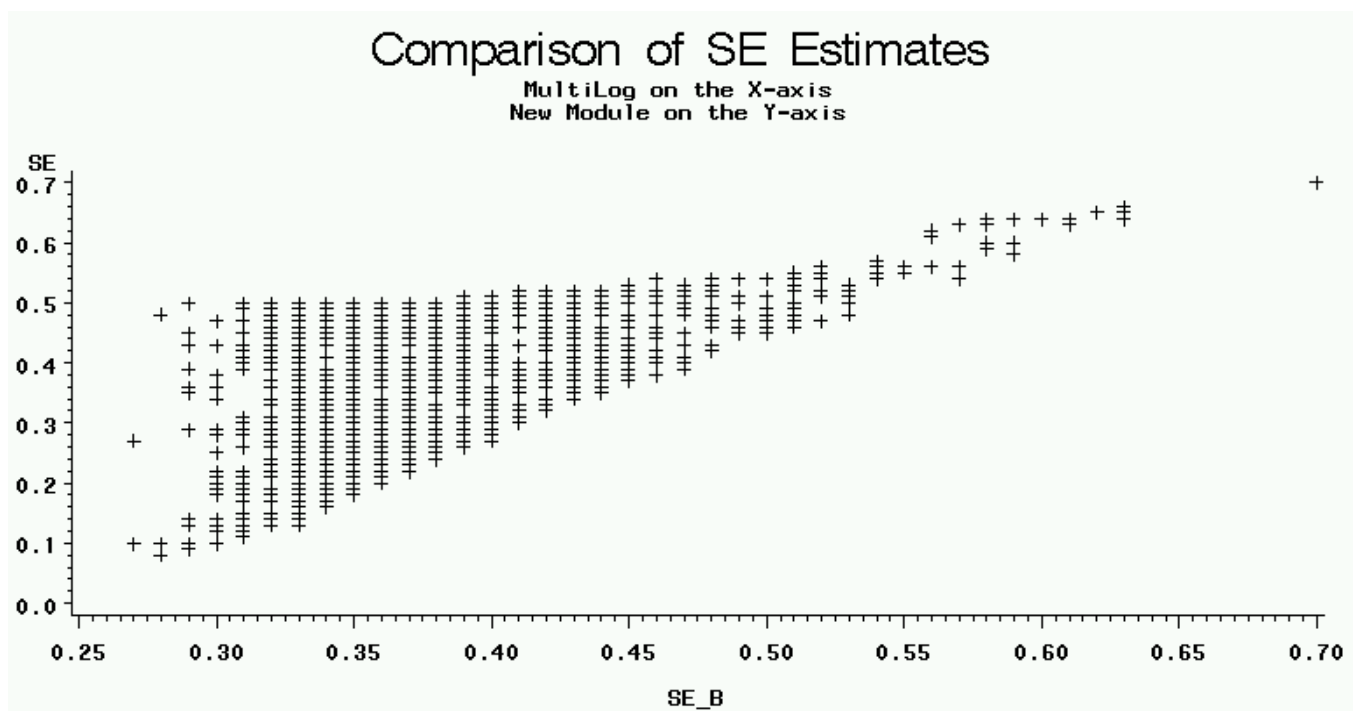


Figure 4

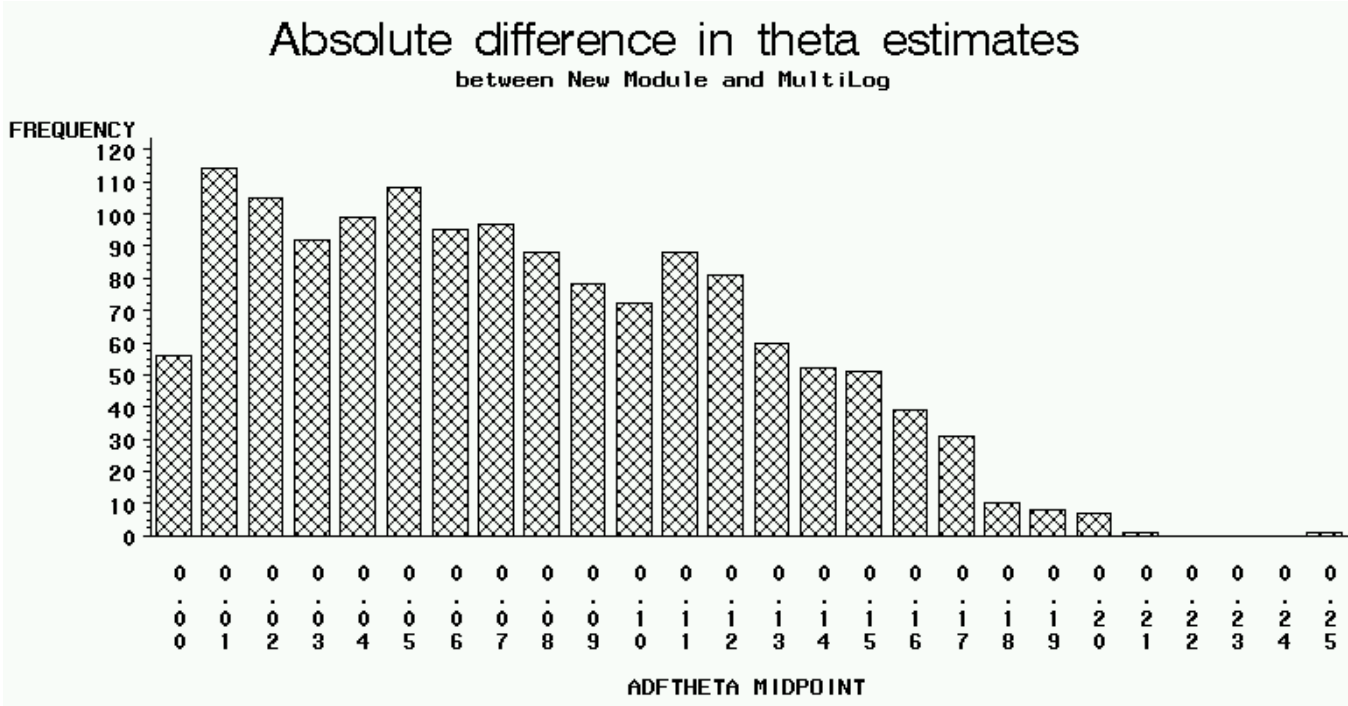


Figure 5

