# Chapter IV

# Results

## Item Elimination

The data are based on the responses of 2,000 examinees to the 77 item DRP. Item discrimination indices, item difficulty, item reliability and the proportion of examinees selecting each response are presented in Appendix I.

Examination of Appendix I reveals a questionable item. Item 75 has a negative discriminative value. This indicates that more low scoring examinees passed item 75 than did high scoring examinees. Its difficulty level (p= .14) appears to be atypical when compared to the difficulty level of the other items on the test and is below the chance level. In examining the distribution of responses, we see that the responses are more or less distributed among four response options, indicating that most examinees appear to have been guessing from among these four choices. The clustering of responses on incorrect options 2 and 4 suggests a superficial attractiveness of these options. The relatively large number of omits suggests that some examinees may have been confused about what was being asked. Ambiguity in the item stem may cause a larger proportion of examinees to be attracted to incorrect responses.

Items that are negative discriminators are considered flawed items. These items are generally eliminated from the item pool or rewritten completely. On this basis, item 75 was excluded from further analyses.

Unidimensionality

The unidimensionality assumption was examined by determining if the first dimension of the test data set accounted for a substantial proportion of the matrix variance. Table 1 presents the results of the principal components and principal factor solutions of the interitem matrix of phi correlations.

Table 1.

Summary of the Principal Factor and Principal Components Solutions Of the Interitem Matrix of Phi Correlations (N=2000)

| Principal Factor Analysis | | | |
|---|---|---|---|
| Factor | Eigenvalue | Percentage of Covariance | Cum. Pct. |
| 1 | 13.66 | .72 | .72 |
| 2 | 2.58 | .14 | .86 |
| 3 | .79 | .04 | .90 |
| 4 | .67 | .04 | .94 |
| 5 | .48 | .03 | .97 |
| **Principal Component Analysis** | | | |
| Factor | Eigenvalue | Percentage of Variance | Cum. Pct. |
| 1 | 14.38 | .19 | .19 |
| 2 | 3.33 | .04 | .23 |
| 3 | 1.55 | .02 | .25 |
| 4 | 1.41 | .02 | .27 |
| 5 | 1.25 | .02 | .29 |

As can be seen from Table 1, the percentage of total variance accounted for by the first factor is substantially greater than successive factors. The factor solution finds 72% percent of the common variance accounted for by the first factor. The principal component solution finds 19% of the total variance is accounted for by the first component. The table shows that the assumption of a dominant dimension underlying the DRP is well founded because the first factor accounted for considerable more variance than any other factor for both the principal components and principal factor solutions. As approximately 20% of the variance is accounted for by the first factor, the DRP would be also be considered unidimensional under the Reckase (1979) criterion.

Degree of Speededness

Table 2 contains basic data on the degree of speededness based on the Swineford/ETS measures of speededness. For the DRP, nearly 92% of the examinees attempted all 77 items on the test. To complete 75% of the test, an examinee must have reached item 57. Nearly all examinees reached the 57$^{th}$ item. Using the ETS rule-of thumb, if "virtually all" of the examinees reach at least three-quarters of the items and if all of the items are reached by at least 80% of the examinees, the test may be considered unspeeded. Using the ETS criteria, the DRP may be considered essentially unspeeded.

Table 2.

Speededness of the DRP

| Total Population | |
|---|---|
| Percentage completing test | 91.7 |
| Percentage completing 75% of the test | 99.5 |
| Number of items reached by 80% Of the examinees | 77 |
| Total number of items | 77 |

The relationship between speededness and ability was determined by comparing the completion rates of examinees classified into 3 ability groups (See Table 3). A comparison of number of items marked across ability groups shows that the number of items attempted is a function of group membership. The lowest percentage of marked response occured among the lowest ability group whereas the high ability group had the highest completion rate.

Table 3.

Percentage of Attempts by Ability Group

| Low Ability Students | |
|---|---|
| Percentage completing test | 86.1 |
| Percentage completing 75% of the test | 98.7 |

| Middle Ability Students | |
|---|---|
| Percentage completing test | 89.9 |
| Percentage completing 75% of the test | 98.9 |

| High Ability Students | |
|---|---|
| Percentage completing test | 96.3 |
| Percentage completing 75% of the test | 100.0 |

Approximately 86% of low ability examinees completed the test, whereas 90% of middle ability and 96% of high ability examinees attempted 77 items. The percentage of examinees completing 75% of the test was extremely high for all ability groups, ranging from 98.7% for low ability examinees to 100% for high ability examinees. In all cases the completion rates satisfy the ETS criteria.

Rapid Guessing Behavior. When the results conform to the ETS criteria, no major problem of speededness is likely. However, a problem with this criterion is that it does not account for some examinees who randomly fill in answers in the hope of getting some of the items correct by chance. This will result in fewer or no unreached items and the test will not appear to be speeded for these examinees. A test is speeded for examinees who engage in rapid guessing behavior (Schnipke, 1995). Therefore, to obtain an accurate measure of speededness, an assessment of rapid guessing behavior must be made.

A method that can be used to check for the existence of rapid guessing is to look for inconsistent response patterns. To accomplish this, the DRP was broken down into thirds ($T_1$, $T_2$, $T_3$) based on item number. Percentage correct scores are computed for $T_1$, the first 25 questions, $T_2$, questions 26 through 52 and $T_3$, questions 53 to 77. $T_2$ and $T_3$ contain 27 and 25 questions respectively. An inconsistent response pattern is defined as one in which a higher percentage correct score is obtained on the harder items. Since the DRP is ordered by difficulty, the expectation is the highest percent correct would occur with $T_1$, the easiest items, and the smallest percent correct would occur with $T_3$ the hardest items. If it can be shown that there is a significant difference in the percentage correct rate of $T_1$ and $T_3$ then this may be an indication that some examinees engaged in rapid guessing.

Frequencies of percentages correct are shown in Tables 4, 5 and 6 across ability groups. Inconsistent response patterns were found for 75, 112, and 306 low, middle and high ability examinees respectively. Of high achieving

students with inconsistent response patterns, similar percent correct responses across comparisons of thirds were observed. Differences in mean percent correct rates ranged from .012 to .015 for high ability students for the $T_2$-$T_3$, $T_1$-$T_3$ and $T_1$-$T_2$ comparisons. The sample sizes are 2, 3 and 301 respectively. For middle ability students, only one student obtained an inconsistent pattern for the $T_1$-$T_3$ comparison and two middle ability students had inconsistent response pattern in the $T_2$-$T_3$ comparison. The difference in mean percent correct for the $T_1$-$T_2$ comparison is .03 for 109 middle ability students. No substantive indications of rapid guessing were found for high or middle ability examinees. The aforementioned results are considered insignificant either due to inadequate sample size or negligible mean percent correct difference.

Table 4.

Percentage Correct for Low Ability Students (N=667)

| Frequencies for $T_1$ and $T_3$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_1$ | 663 | 99.4 |
| Higher percent correct in $T_3$ | 4 | .6 |

| Frequencies for $T_1$ and $T_2$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_1$ | 634 | 95.1 |
| Higher percent correct in $T_2$ | 33 | 4.9 |

| Frequencies for $T_2$ and $T_3$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_2$ | 629 | 94.3 |
| Higher percent correct in $T_3$ | 38 | 5.7 |

Table 5.

Percentage Correct for Middle Ability Students (N=666)

| Frequencies for $T_1$ and $T_3$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_1$ | 665 | 99.8 |
| Higher percent correct in $T_3$ | 1 | .2 |

| Frequencies for $T_1$ and $T_2$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_1$ | 557 | 83.6 |
| Higher percent correct in $T_2$ | 109 | 16.4 |

| Frequencies for $T_2$ and $T_3$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_2$ | 664 | 99.7 |
| Higher percent correct in $T_3$ | 2 | .3 |

Table 6.

Percentage Correct for High Ability Students(N=667)

| Frequencies for $T_1$ and $T_3$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_1$ | 664 | 99.6 |
| Higher percent correct in $T_3$ | 3 | .4 |

| Frequencies for $T_1$ and $T_2$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_1$ | 366 | 54.9 |
| Higher percent correct in $T_2$ | 301 | 45.1 |

| Frequencies for $T_2$ and $T_3$ | Frequency | Percent |
|---|---|---|
| Higher percent correct in $T_2$ | 665 | 99.7 |
| Higher percent correct in $T_3$ | 2 | .3 |

For low ability students, statistically significant differences in percent correct scores were observed for the $T_1$-$T_2$ and $T_2$-$T_3$ comparisons of the DRP (see Table 7). An example of two examinees with dramatic differences in percentage correct is an examinee with a total correct score of 18 who correctly responded to only one question in $T_2$ and scored eight correct questions in $T_3$. Another examinee answered two questions correctly in $T_2$ and 10 right in $T_3$. This examinee has a total correct score of 19. The difference in percent correct is .28 and .36 respectively for these two examinees.

Table 7.

Dependent Samples t-test Analysis on Percent Correct for Low Ability Examinees with Inconsistent Response Patterns

$T_1$-$T_2$ Comparison

| | $T_1$ | | $T_2$ | | Difference | | | |
|---|---|---|---|---|---|---|---|---|
| N | Mean | Std. | Mean | Std. | Mean | Std. | t | p-value |
| 33 | .70 | .16 | .64 | .18 | .06 | .05 | 5.91 | .0001 |

$T_2$-$T_3$ Comparison

| | $T_2$ | | $T_3$ | | Difference | | | |
|---|---|---|---|---|---|---|---|---|
| N | Mean | Std. | Mean | Std. | Mean | Std. | t | p-value |
| 38 | .26 | .07 | .19 | .07 | .07 | .08 | 5.91 | .0001 |

Nonresponses. Eight percent (160) of the test takers omitted or did not reach items toward the end of the test. Of those, 145 examinees had omitted responses, 30 had not -reached responses and 9 examinees engaged in both types of behaviors. The number of omits ranged from 1 to 21. The number of not-reached items ranged from 1 to 42.

Item Skipping. In order to distinguish between examinees who use item skipping as a strategy to complete as many items as possible and those who skip items with no apparent test taking strategy in mind, the examinees were divided into two groups. Based on Nagy (1986), examinees who have skipped items are defined in terms of the number of items the test taker has skipped. Low skippers (n=123) are defined as examinees with one or two omits. Examinees who have omitted three to 20 items are considered moderate skippers (n=21). High skippers (n=1) have omitted more than 20 of the items.

Of those who have omitted responses, 84% tend to omit at most two items. This low skipping behavior is more or less evenly distributed across ability levels. As expected, low ability examinees tend to omit more items than middle and high ability examinees. The range of omits for low ability examinees is 21, while it is 4 for the most able students. If the proportion of omitted responses for an item is greater than .15, that item is considered to have a high omitted nonresponse rate. No item has an omitted nonresponse rate greater than .0005.

Not-reached Items. Slightly more low than middle achievers did not respond to a string of items at the end of the test. A not-reached nonresponse rate is calculated as the number of not-reached responses for a particular item divided by the total sample size. An item is considered to have a high nonresponse rate if 15% or more of the test takers did not reach that item. As not-reached nonresponse rates ranged from 0 to .01, no item was considered to have a high not-reached nonresponse rate.

The DRP is considered to be essentially unspeeded. However, the existence of rapid guessing cannot be overlooked, especially for low ability examinees. Its existence is not enough to characterize the DRP as speeded, but the significant dependent samples t-test provides the first indication that guessing may have adverse effects on model specification.

## Equal Discrimination Indices

Uniformity in discrimination is quite an important assumption for the Rasch model. If there are serious departures from this assumption, an alternate model must be applied. To investigate this assumption, the distribution of biserial correlations is examined.

A careful examination of Figure 4 reveals a substantial variation in the levels of the item discrimination as measured by the biserial correlations. The biserial correlations ranged from .16 to .925. If a large percent of biserial correlations fall outside $\pm$ .15 of the mean biserial correlation (.606), the

assumption of equal discrimination has been violated (Hambleton and Swaminathan, 1985). Since 36.8% of the biserial correlations fall outside this range, the assumption of equal discrimination is not likely.

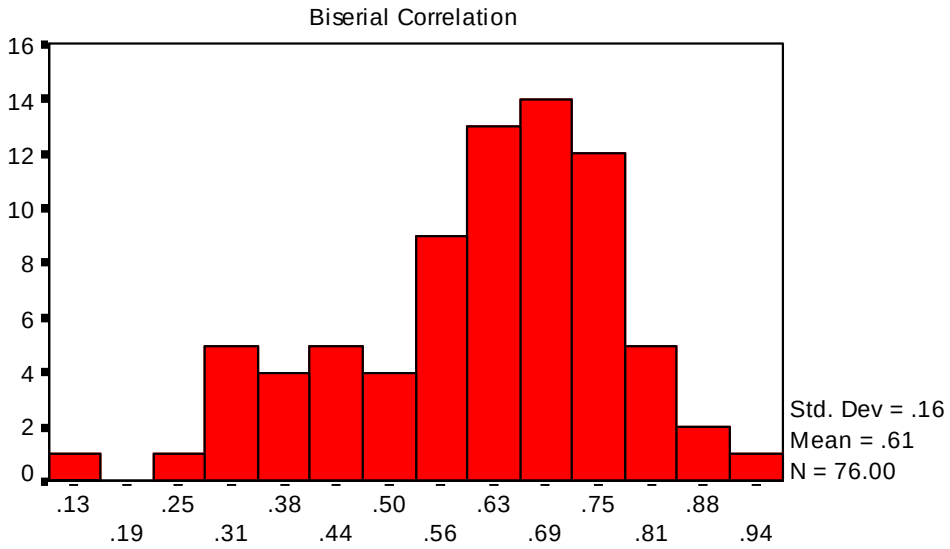



Figure 4. Histogram of the biserial correlations for 76 items.

This assumption may be evaluated from another viewpoint. The discrimination indices from the two- and three-parameter BILOG calibration may be tallied into a frequency distribution and then plotted in a histogram as depicted in Figures 5 and 6 respectively. We can observe that the distributions of discrimination indices fail to form a leptokurtic distribution to a degree sufficient to demonstrate uniformity in discrimination.

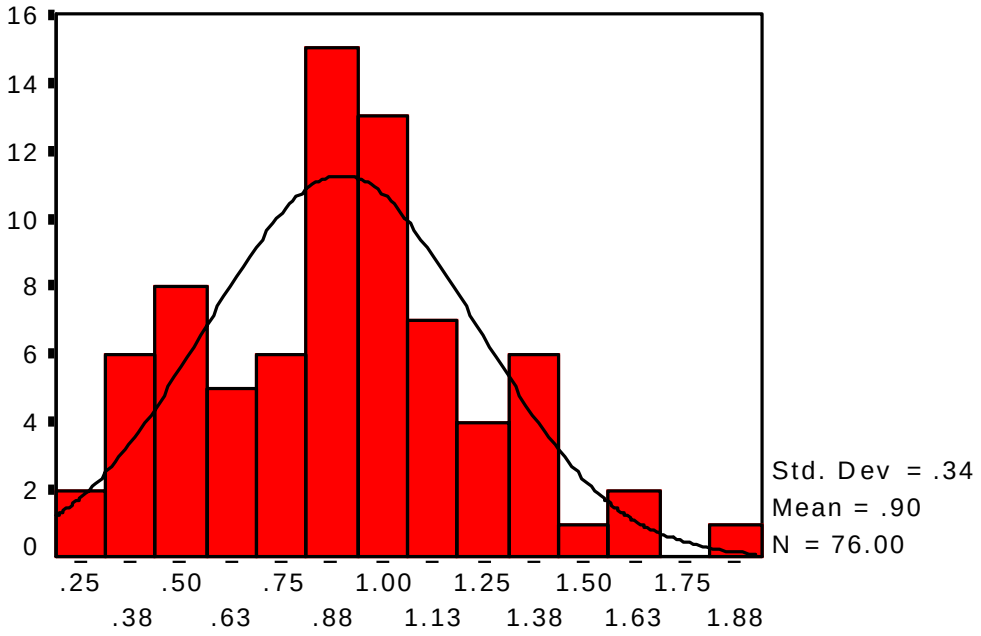



Figure 5. Histogram of discrimination indices for the two-parameter model BILOG calibration

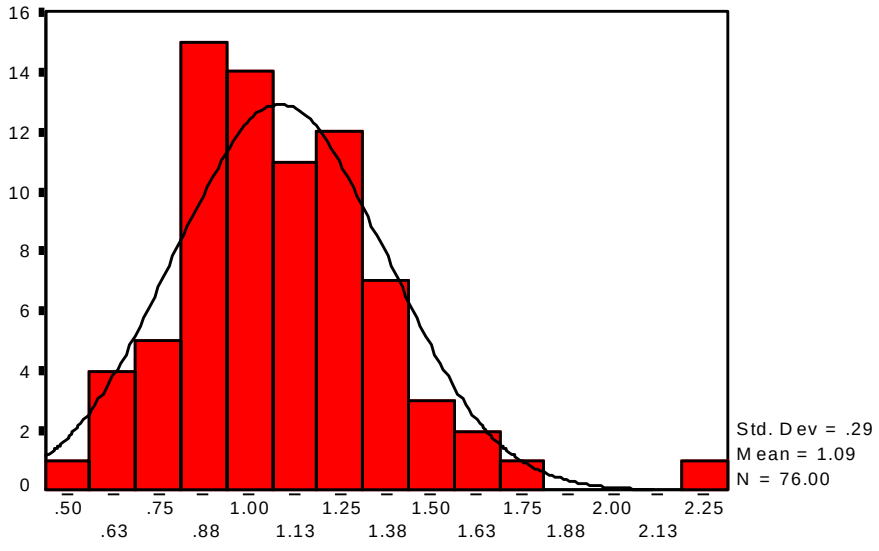



Figure 6. Histogram of discrimination indices for the three-parameter model BILOG calibration

Based on the Keifer criterion, an item's discrimination index ±3 standard errors must fall within the range of .8 to 1.2 to satisfy this assumption. The two-parameter BILOG calibration resulted in 34% of the items whose upper or lower discrimination parameter limits fell outside the appropriate limits. Most items classified as being uniform in discrimination were easy items. The three-parameter BILOG calibration resulted in 70% of the items drawing confidence intervals inconsistent with the assumption that respective discrimination indices are one.

These preliminary findings suggest that a model that accounts for item discrimination is likely to provide a better fit to the DRP. The comparison of the average absolute-value standardized residuals with the two- and three-parameter BILOG discrimination indices in the residual analysis section supports this finding.

## Guessing

In order to assess the possibility of an examinee obtaining the correct answer independent of ability but simply by means of a lucky random guess, the difference between observed item difficulty and item difficulty adjusted for guessing is used. The observed item difficulty is the classical p value. Item difficulty adjusted for guessing is calculated as the difference between the proportion of examinees who attempted the item and missed it divided by the number of alternatives minus one and the proportion of examinees who correctly

responded to an item. These differences are reported in Table 8. In the absence of guessing, the expected value of these differences are zero. Because the differences between observed and adjusted item difficulties range between .019 and .142, the lower asymptote may not be zero for all items. This is particularly true for the less able group where mean difference (.142) and the range of these differences (.201) are greatest.

Table 8.

Mean Difference Between Observed and Adjusted Difficulties

| Ability Grou | Mean Difficulty Item | Adjusted | Difference |
|---|---|---|---|
| Low | .430 | .288 | .142 |
| Middle | .732 | .665 | .067 |
| High | .922 | .903 | .019 |

Note. Ability based on number correct score. Mean difference = observed minus adjusted.

Another way to establish whether guessing is prevalent is to inspect the lower asymptote values derived from the three-parameter BILOG calibration. If these values are close to zero then there is no need for a lower asymptote. Figure 7 presents a histogram of these c-values.

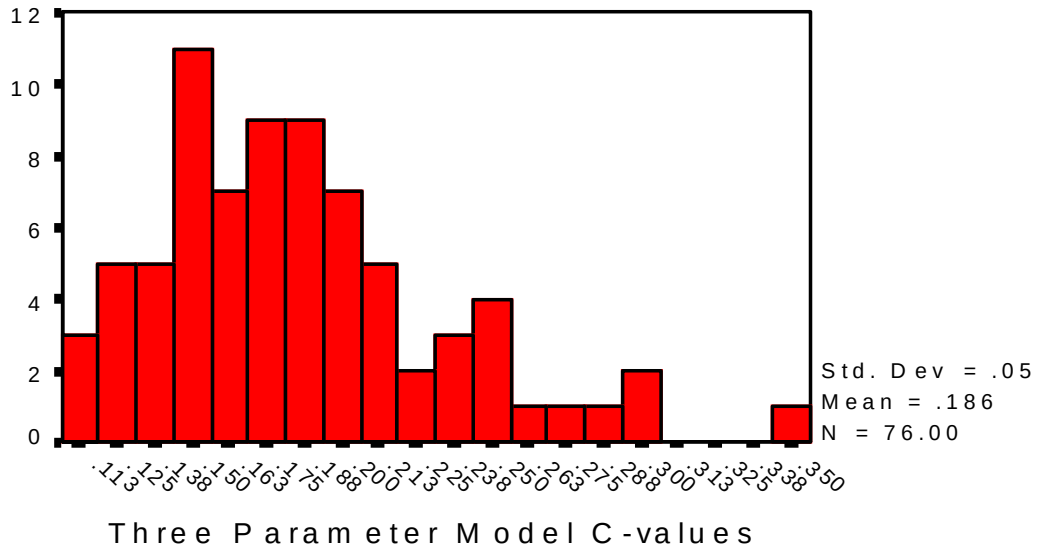



Figure 7. Histogram of lower asymptote values.

These c-values ranging in value from .11 to .35 have relatively small standard errors. The mean standard error is .058. As can be seen, the mean is much greater than zero, establishing the need for a lower asymptote. More than thirty percent of the items have estimated lower asymptotes greater than .2, the probability of correctly responding to an item through a random guess.

The D'Costa measure of surprise was used to determine whether guessing played a role in the response behavior of the examinees. This index was computed using the BWSINDEX program (D'Costa, 1994). The program was written to provide within (concern) and beyond ability (surprise) caution indices for up to 200 examinees and 200 questions. In order to perform the analysis, the data was divided into ten samples of size 200. The surprise (B)

index measures the extent to which an examinee is obtaining correct responses to items above his or her ability level. The B index is a function of proportions and will not be calculated using the total sample. However, it is reasonable to assume that similar indices would result as the samples are relatively homogenous (see Table 9). Median values for each sample approximate 56, the median of the entire sample.

Examination of the B indices across the 10 samples reveals a large number of students were obtaining correct answers to items outside of their ability level. Of 2,000 examinees, 32.6% had significant B indices (> .45). This percentage represents students across the range of ability. Three percent of inconsistent response patterns came from individuals who answered less than half (39) of the items correct on the test. Of 215 examinees characterized as low achievers, 32% of the examinees had a significant b index. Appendix II presents the results for low achievers who had significant B indices. These findings seem to suggest that guessing is prevalent across the distribution of ability.

The presence of guessing is established in a number of ways. The possibility of nonzero lower asymptote values is raised due to nonzero differences between observed and adjusted item difficulties. The mean BILOG calibrated three-parameter c-value of .19 strengthens the premise that lower asymptotes are necessary. Approximately one-third of the examinee population exhibited inconsistent response patterns as measured by the D'Costa B index. These patterns were found among examinees of high, middle and low ability.

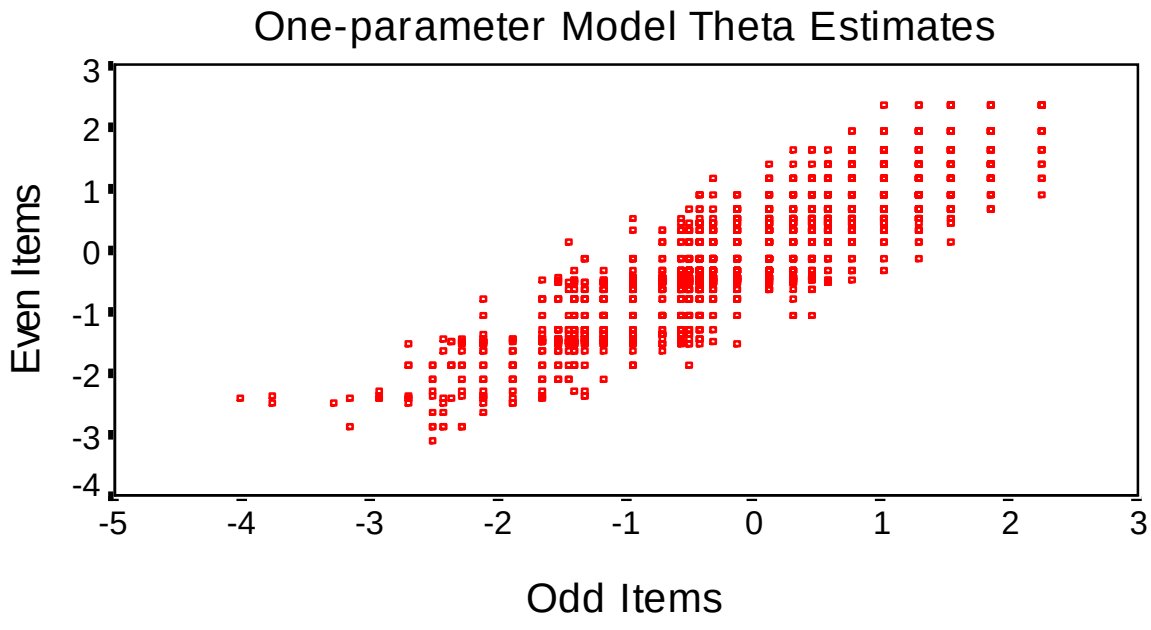
Table 9.

Descriptive Statistics for Total Correct Scores for 10 Samples (n=200)

| Sample | Minimum | Maximum | Range | Median |
|---|---|---|---|---|
| 1 | 11 | 76 | 65 | 56 |
| 2 | 26 | 76 | 50 | 62 |
| 3 | 24 | 77 | 53 | 60 |
| 4 | 22 | 76 | 54 | 56 |
| 5 | 11 | 74 | 63 | 58 |
| 6 | 19 | 76 | 57 | 57.5 |
| 7 | 11 | 73 | 62 | 55 |
| 8 | 16 | 74 | 58 | 51 |
| 9 | 14 | 74 | 60 | 56 |
| 10 | 14 | 75 | 62 | 56 |

Invariance of Ability Estimates

In order to determine if the invariance of ability estimates has been established, Bayesian estimates of ability are obtained on the easiest and the hardest 38 items on the test. If correlations of ability estimates obtained from the two halves of the test are comparable to correlations of baseline odd versus even items, the invariance of ability estimates has been established.



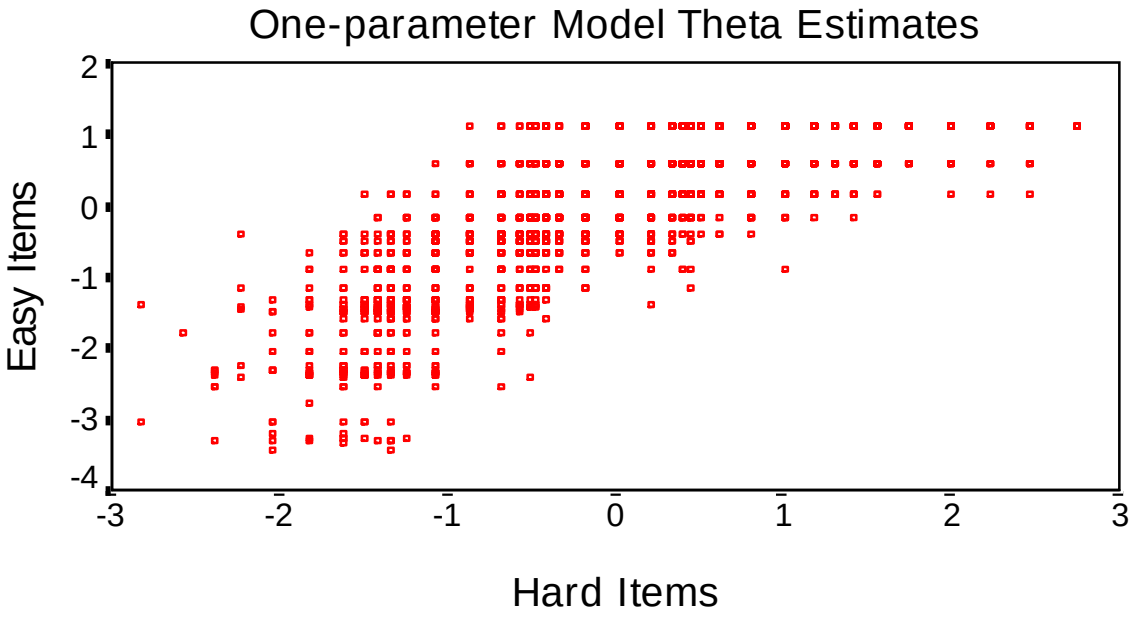




Figure 8. Bayesian estimates of ability for odd VS, even items (r=.88, N=2000) and Bayesian estimates of ability for hard VS. even items (r=.77, N= 2000) respectively for the one-parameter model.

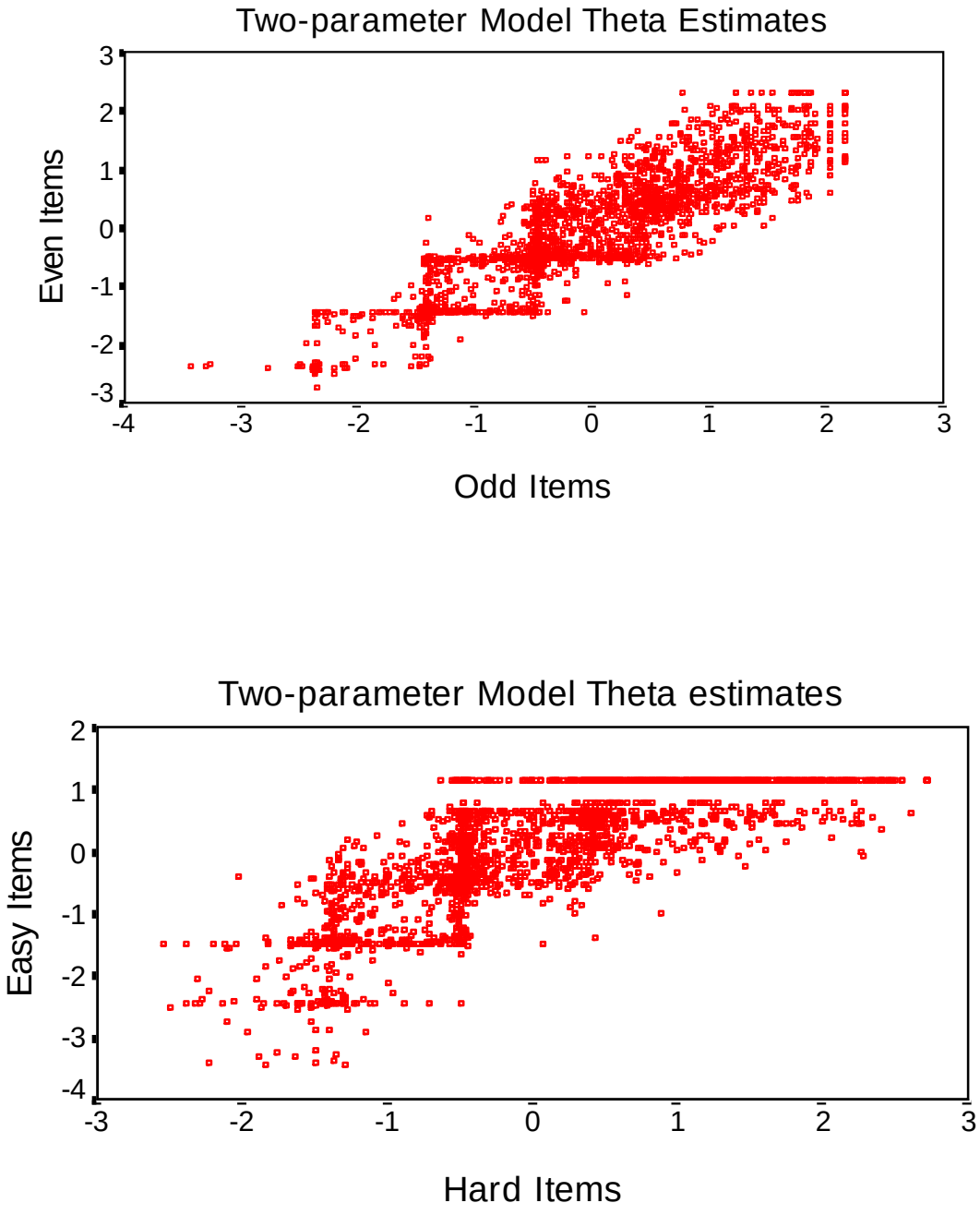



Figure 9. Bayesian estimates of ability for odd VS, even items (r=.88, n=2000) and Bayesian estimates of ability for hard VS. even items (r=.77, n= 2000) respectively for the two-parameter model.

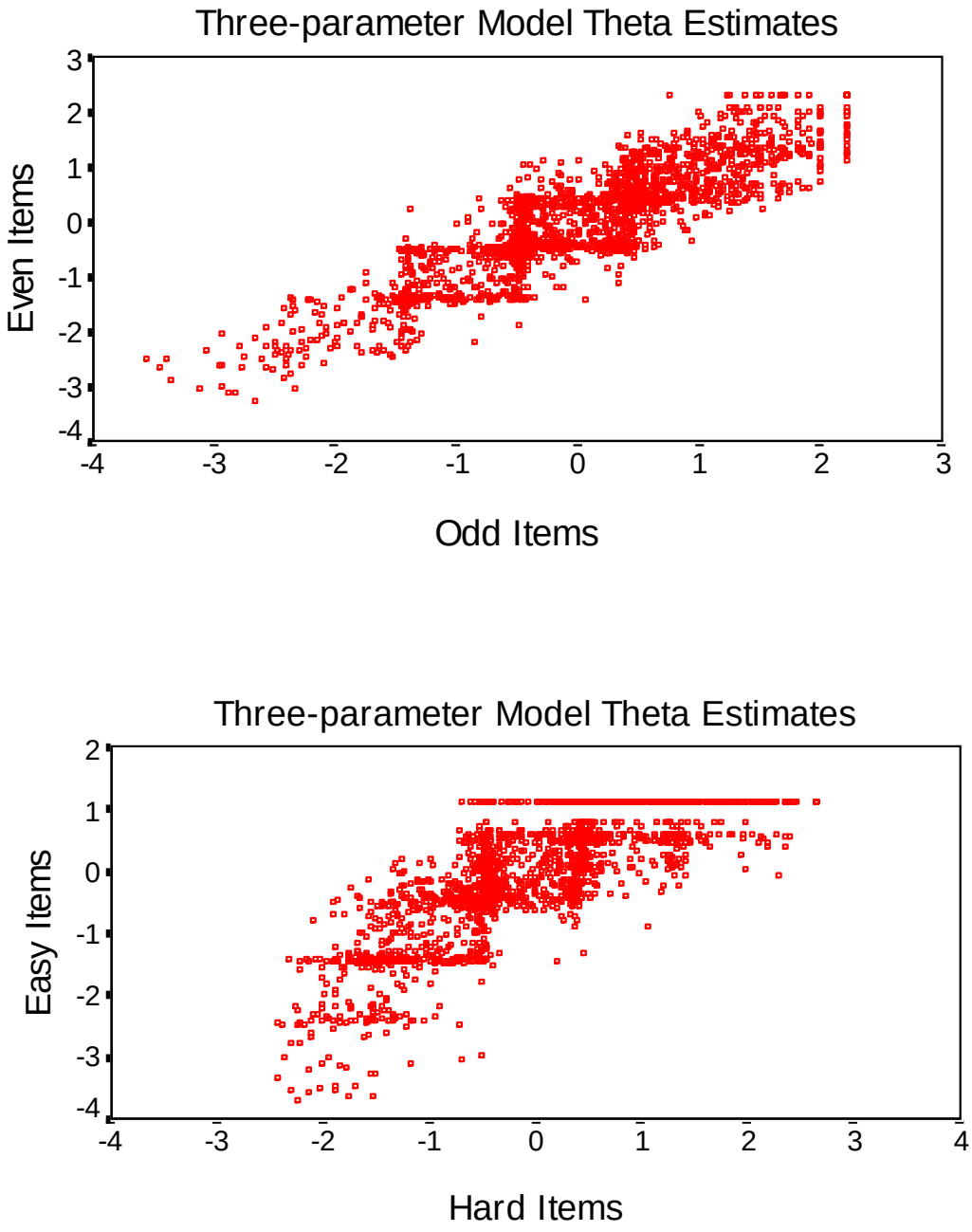



Figure 10. Bayesian estimates of ability for odd VS, even items (r=.89, n=2000) and Bayesian estimates of ability for hard VS. even items (r=.79, n= 2000) respectively for the three-parameter model.

The examination of baseline scatter diagrams demonstrates strong positive correlations between the ability estimates computed from the odd and even items on the test for the one-(r=.88), two-(r=.88) and three-parameter (r=.89) models. The plotted points of Figures 8 through 10 (see these figures noting change in scale) suggest that a linear relationship may exist between theta estimates obtained by calibrating the easy and hard items on the test as well.

When ability estimates are invariant, the estimation of ability will be approximately the same regardless of the set of test items chosen. Based on this analysis, invariance cannot be refuted for any of the three logistic models.

The Test Standard Error Function. As a result of the difficulty of BILOG to provide estimates of ability for very low or high ability examinees, statements about the accuracy of ability on the test as a whole should be made. An estimate of the accuracy of ability is given by the standard error of this parameter. The test standard error function provides an indication of how accurately ability can be measured by the test. The most precise test measurements are represented by small differences between the function and the abscissa. Figures 11, 12, and 13 show the test standard error of measurement for the one-, two- and three-parameter models. Plots of the standard errors of all three models show a concave surface that increases at the extremes of the ability scale.

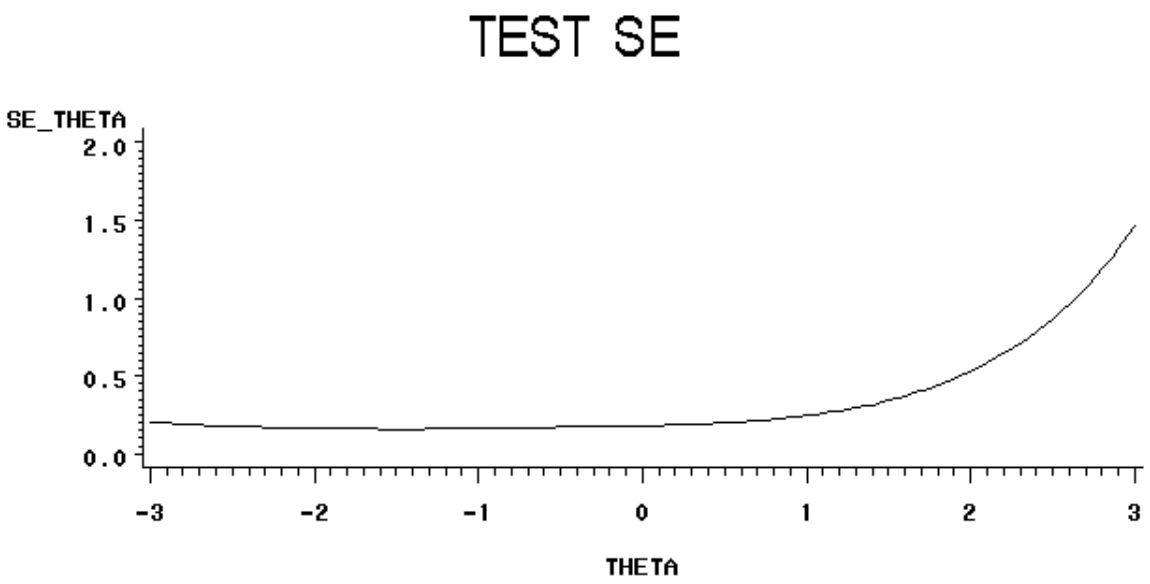



Figure 11. Test standard error function for the one-parameter model.

The salient characteristics of the three functions occur at the endpoints of the distribution of ability. The one-parameter model provides the worst estimates of ability for high ability students and the three-parameter model provides the best estimates of ability for these students.

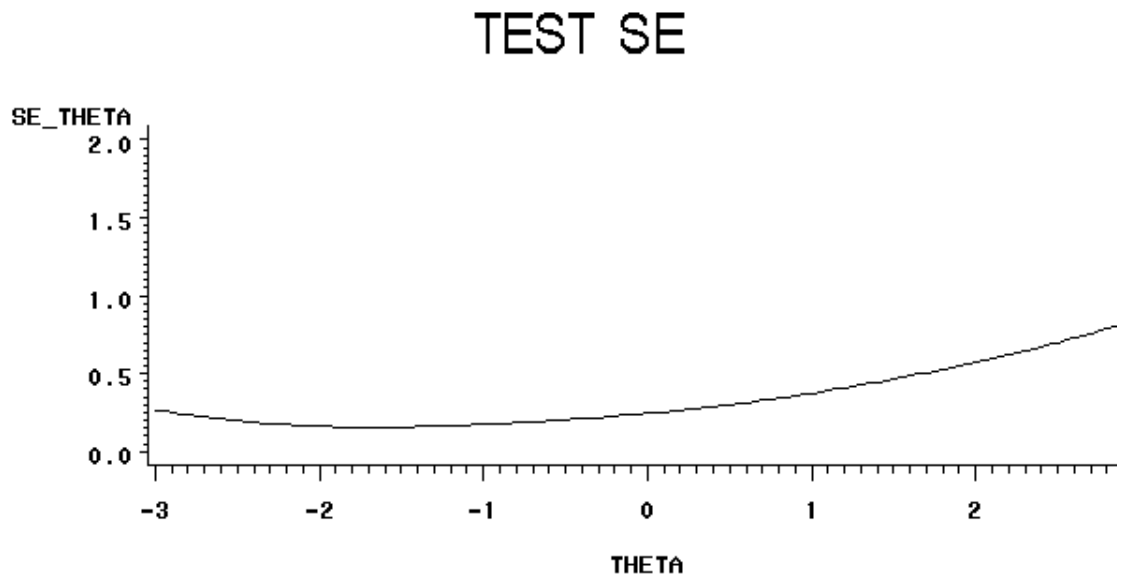



Figure 12. Test standard error function for the two-parameter model.

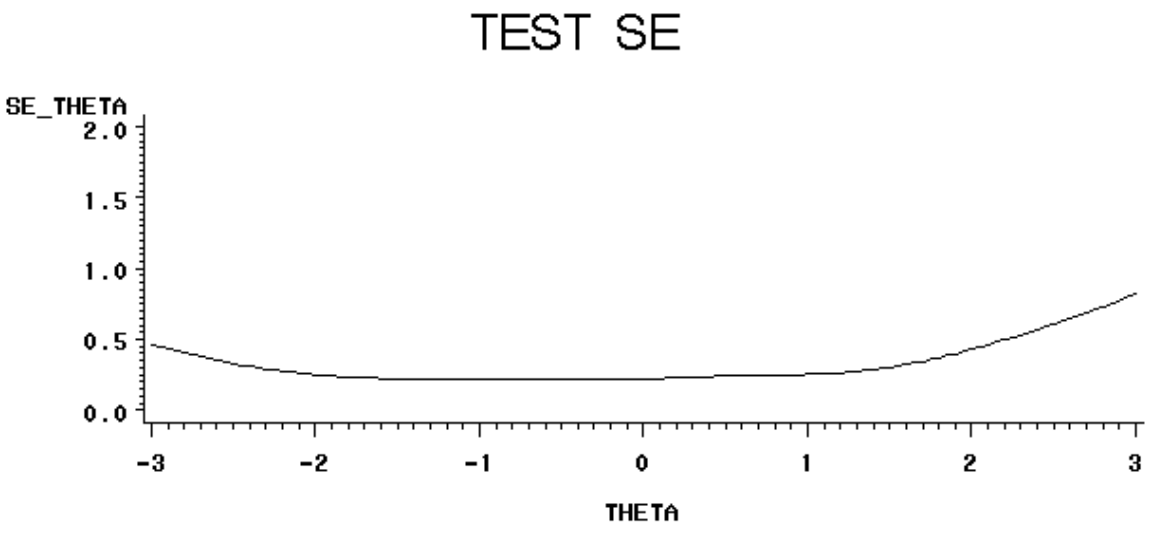



Figure 13. Test standard error function for the three-parameter model.

It is interesting to note that the three-parameter model, advocated for adjustments made to the probability of a correct response as ability declines makes only slightly worse estimates for the least able students as compared to the other models. The two-parameter model can be discounted as best estimating ability as errors in estimation increase steadily starting below the midpoint of the ability distribution. The one- and three-parameter models approximate ability for the middle range of ability with similar accuracy, but the three-parameter model provides the most error-free estimates of ability for the endpoints of the ability distribution.

Test Information. Inspection of the test information function (TIF) allows one to determine the range of the latent trait for which the test measures best. The TIF of the 76 item DRP are presented in Figures 14, 15 and 16 for the one-, two and

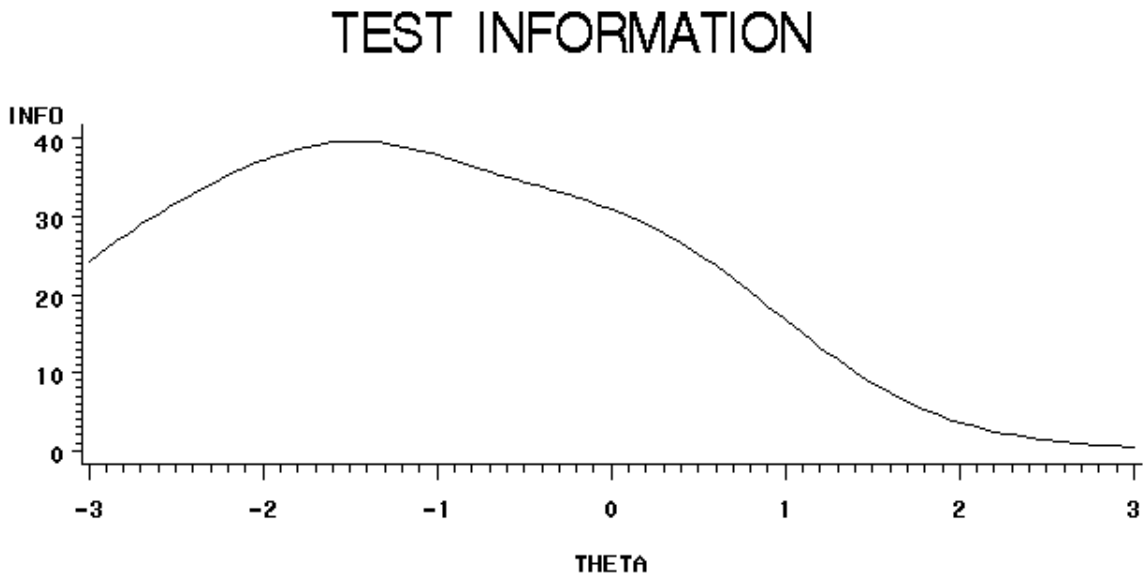



Figure 14. Test information function for the one-parameter model

three-parameter models respectively . Figure 14 provides the TIF for the one-parameter model.

An interesting point to consider is that the Rasch calibrated DRP is designed to provide information across a broad range of abilities enabling the differentiation between examinees with respect to their reading competency, however, the TIF is not as flat as one might expect.

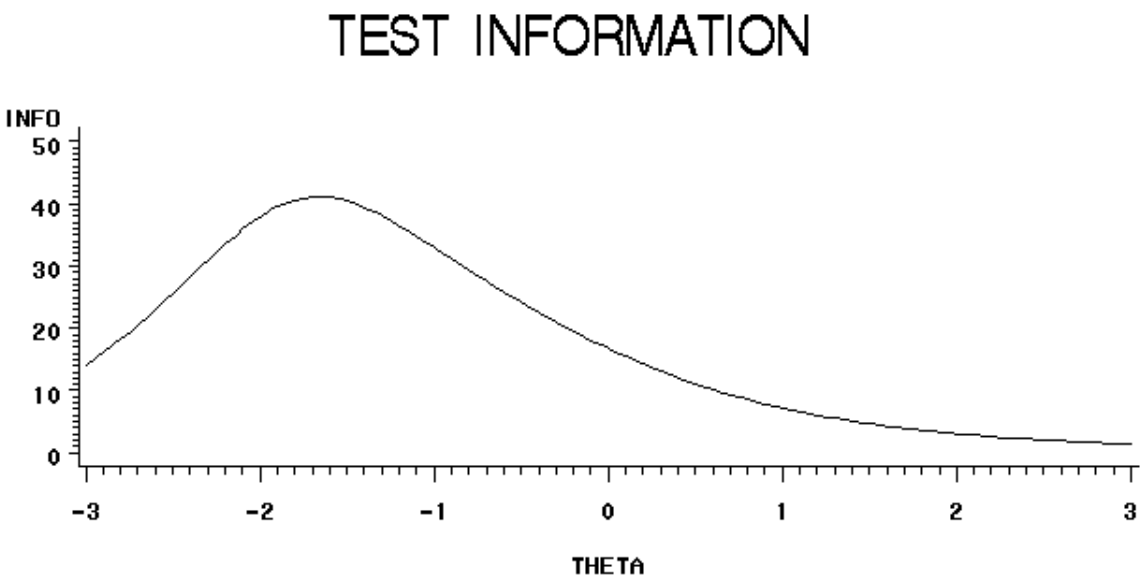



Figure 15. Test information function for the two-parameter model.

The two-parameter model resulted in a test information function that provided maximum information for abilities below the center of the ability distribution. Note that the function reaches its maximum near $\theta$=-1.5 and falls off sharply in both directions. The DRP as modeled by the two-parameter model can be useful for separating students into two categories: middle to high reading comprehension and low reading comprehension. The DRP provides little

information about the reading comprehension of students with ability greater than +.5 when calibrated by the two-parameter model.

The three-parameter model TIF plateaus between –2 and +1, making its use appropriate for middle ability students. The three-parameter model, however, provides more information at the lower end of the ability continuum than at the higher end of the ability distribution. Since the three-parameter model often provides a somewhat better fit to test data at the lower end of the ability continuum, it is surprising that this model is not more useful for predicting the ability for students of low ability.

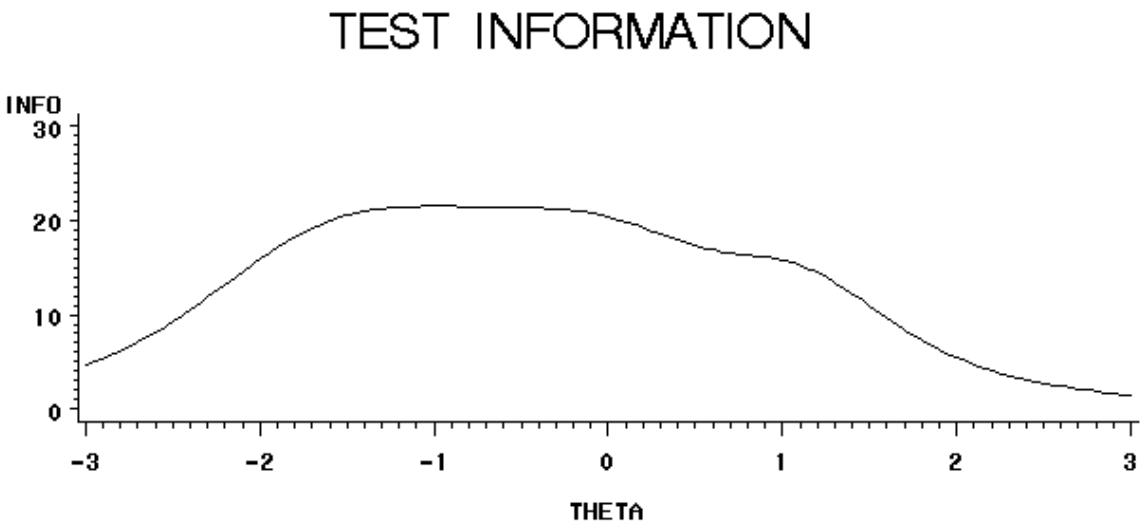



Figure 16. Test information function for the three-parameter model.

Item information functions tend to provide more information when it is assumed that guessing does not occur. The one-and two-parameter models having zero lower asymptotes will thereby produce item information functions that appear to provide more information than that of the three-parameter model. This in turn affects the information provided by a test at $\theta$ as the TIF is the sum of the item information at a given ability level. However, these item information functions and consequently the TIF are only valid when the data fits the model. If any two of the models fit the data, their TIFs would be identical. Since the three TIFs are vastly different, this is an indication that one and only one or that none of the three logistic models fit the data. In order to determine which, if any, of the models fit the data, an analysis must be made concerning the accuracy of the predictions of examinee performance. The model shown to exhibit the largest number of small residuals is likely to provide the most optimal fit to the DRP.

## Item Parameter Invariance

The invariance of item parameters is established when item parameter estimates are consistent across different subsamples, allowing only for examinee sampling error. The comparison of plots of item parameters obtained from subgroups that differ in ability can be used to determine if item parameters are invariant. A baseline for interpreting the plots of high vs. low achievers was obtained by comparing plots of item parameters for two randomly equivalent samples of high achievers and two randomly equivalent low ability samples.

The examinees are divided in half based on a median split of the number-right score creating two subgroups of 1,000 high and 1,000 low performing examinees. The high and low ability examinees were then divided at random into two subgroups. These four samples labeled as high ability students I and II and low ability students I and II each contain 500 cases. In the process of calibrating the items, the three-parameter model diverged creating the need to delete negative or low discriminating items. The most parsimonious sample consisted of 62 items. These 62 items were then used to obtain item parameter estimates for the one-, two, and three-parameter models.

If item parameters are invariant, plots of high vs. low ability comparisons should be linear with a slope of approximately one and should not differ significantly in scatter from the baseline plots. If the slope approximates one, item parameter estimates obtained from two different ability groups estimate the same parameter value. However, the verification of a positive slope is not enough to establish invariance. If it can also be shown that the expected difference in item parameter estimates between, high-low comparison samples, approximates zero or equivalently that the correlation of these differences approximates zero (Hambleton and Murray, 1986) then invariance holds. To assess invariance using the above-mentioned criteria, difference scores were obtained for four groups:

(1) The two samples of high ability students,

(2) The two samples of low ability students,

(3) Sample I high ability and sample I low ability and

(4) Sample II high ability and sample II low ability.

Small correlations between the first two groups and the last two groups will provide evidence of invariance. The expected difference in item parameter estimates between the two high and two low ability samples will serve as baseline scatter plots.

To assess the degree to which invariance is obtained across the three models, correlations between high-low samples are compared. Four possible correlations are available:

(1) Sample I high and sample I low ability students

(2) Sample I high and sample II low ability students

(3) Sample II high and sample I low ability students

(4) Sample II high and sample II low ability students.

The Fisher's z test was used to test the difference between two independent correlations and provided some insight into whether the degree of relationship between item parameter estimates is significantly higher in one model than those obtained from another model

For the one-parameter model, Figure 17 contains baseline plots of b-values in the two high and two low performing samples. There is a strong positive relationship between these baseline samples as evidenced by their correlations. For the high performing sample this correlation is .976 whereas it is .992 for the less able group. Figure 18 reveals that for the one-parameter model, item parameters do not correlate as strongly as the baseline samples. For sample I the correlation reduces to .918. For sample II, it is .925.

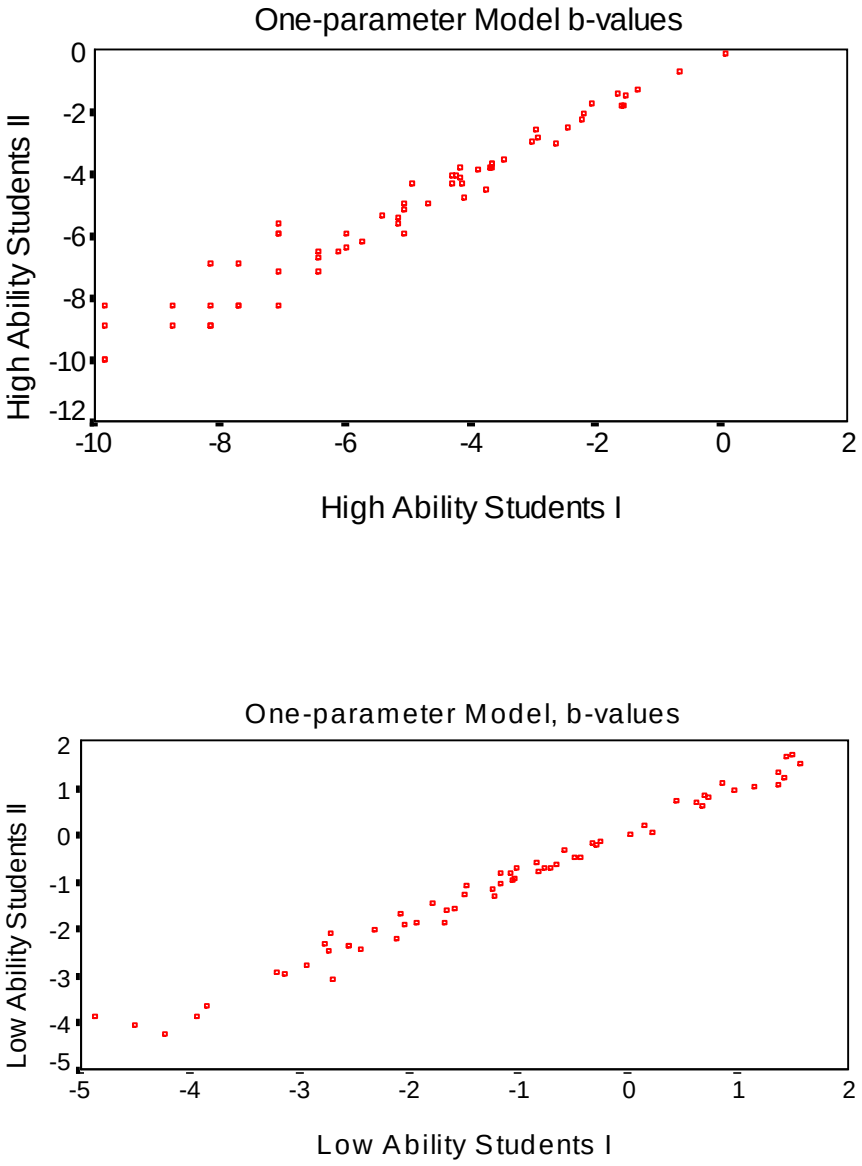
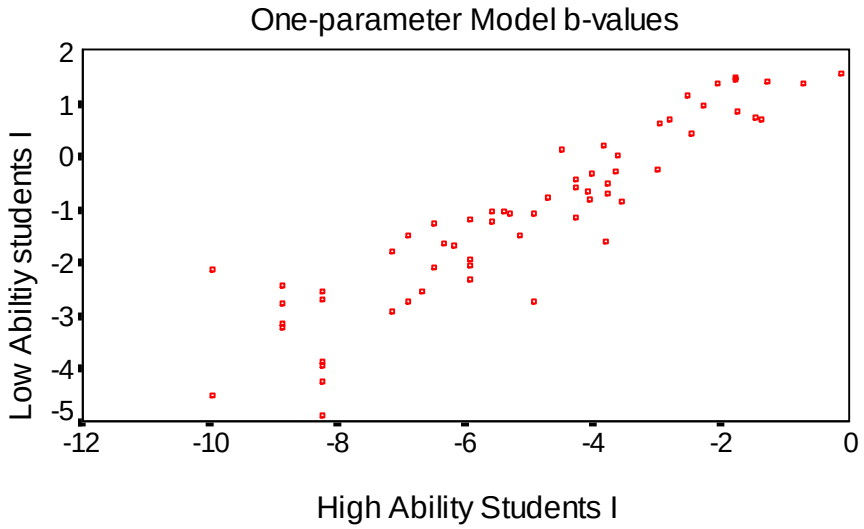



Figure 17. Plots of b-values for the one-parameter model obtained from two equivalent high performing students (N=500, r=.976) and two equivalent low performing students (N=500, r=.992).



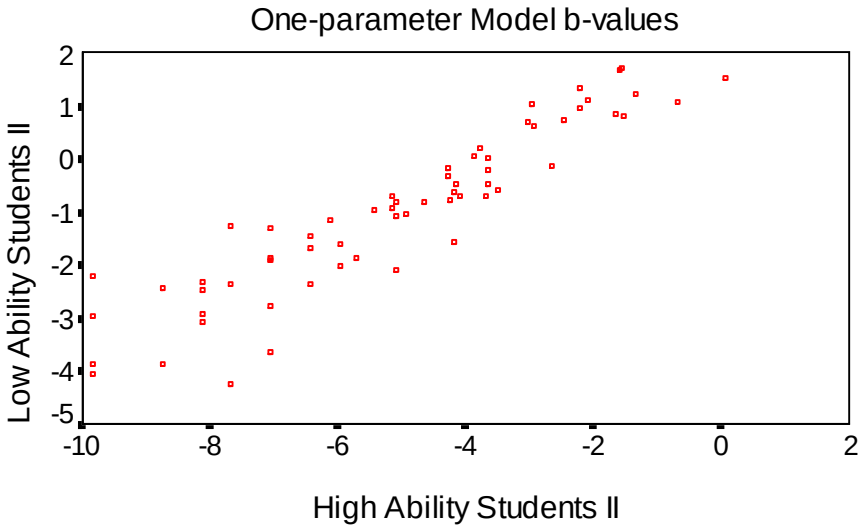
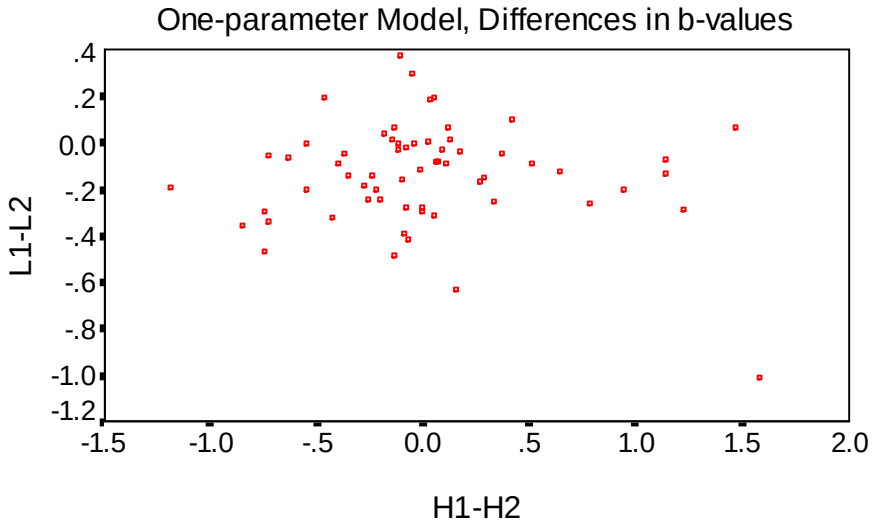




Figure 18. Plots of b-values for the one-parameter model obtained from two different high-low comparisons (N=500, r=.918 and r=.925 respectively).



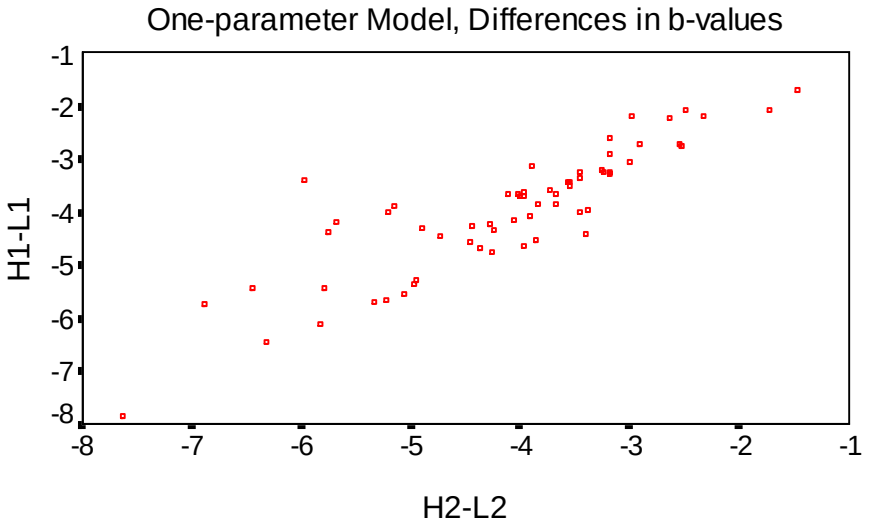




Figure 19. Plots of b-value differences, r=-.101 and r=.873 respectively.

As can be seen from Figure 19, the correlation between the differences in b-values for the baseline samples is very close to zero (-.101). The correlation between the high-low sample is .873 indicating that item difficulty estimates are not invariant across ability groups. Test items located at the bottom left-hand

corner of this figure provide the most inconsistent differences in item difficulty estimates in the two groups and should be reviewed.

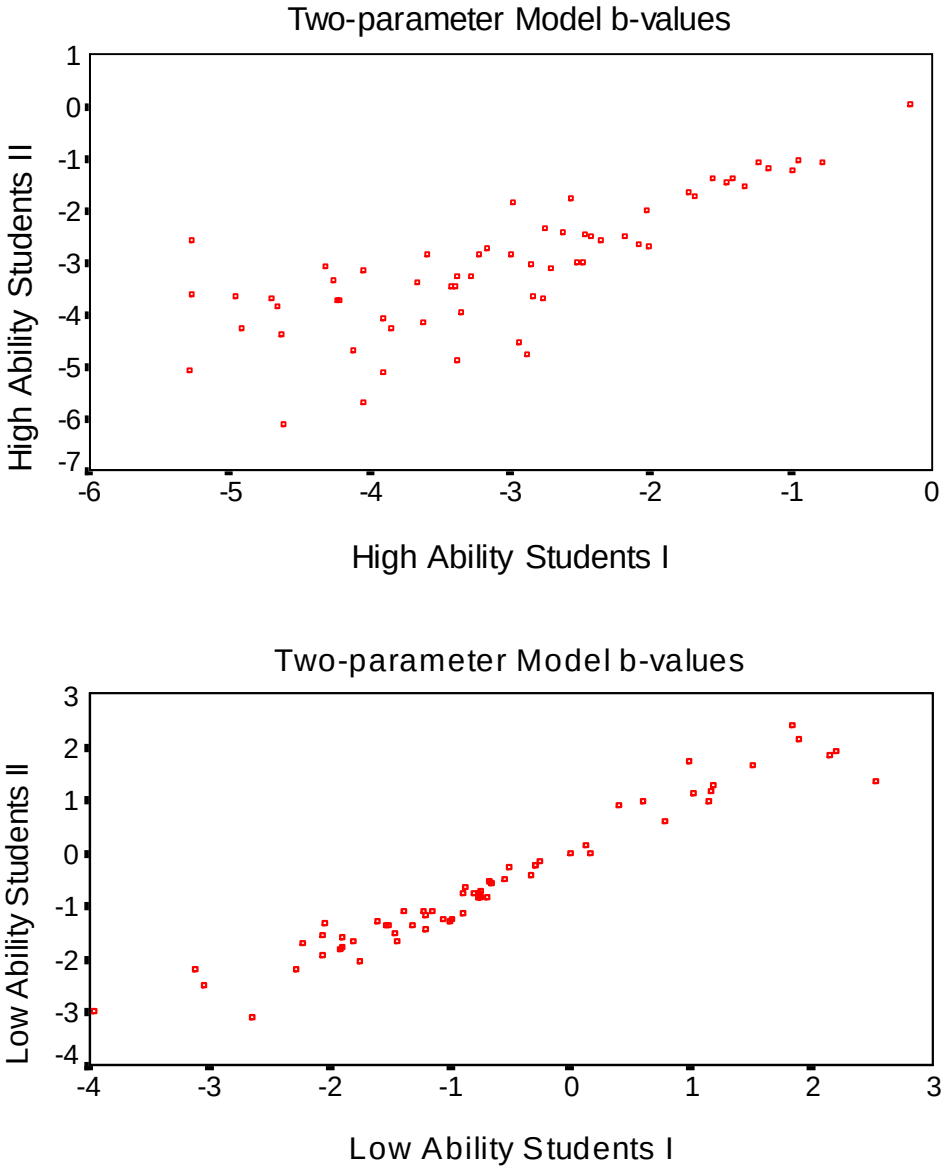
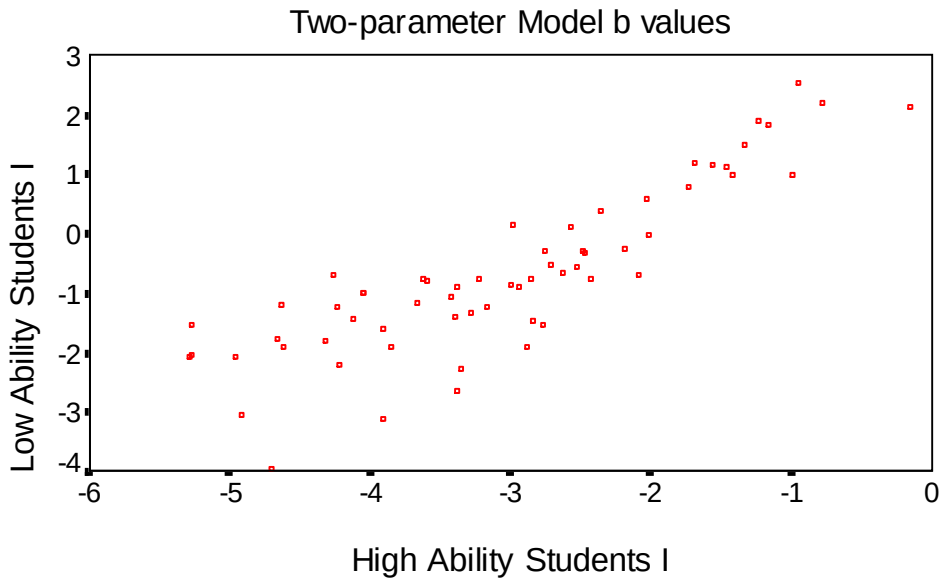



Figure 20. Plots of b-values for the two-parameter model obtained from two equivalent high performing students (N=500, r=.792) and two equivalent low performing students (N=500, r=.97).

Figures 20 and 21 provide comparisons between b-values obtained with the baseline and the two samples of high-low achievers respectively for the two-parameter model. The plot of the baseline sample of high achievers is represented by a large amount of scatter about a positive slope. The correlation between the sample of high achievers is .792.



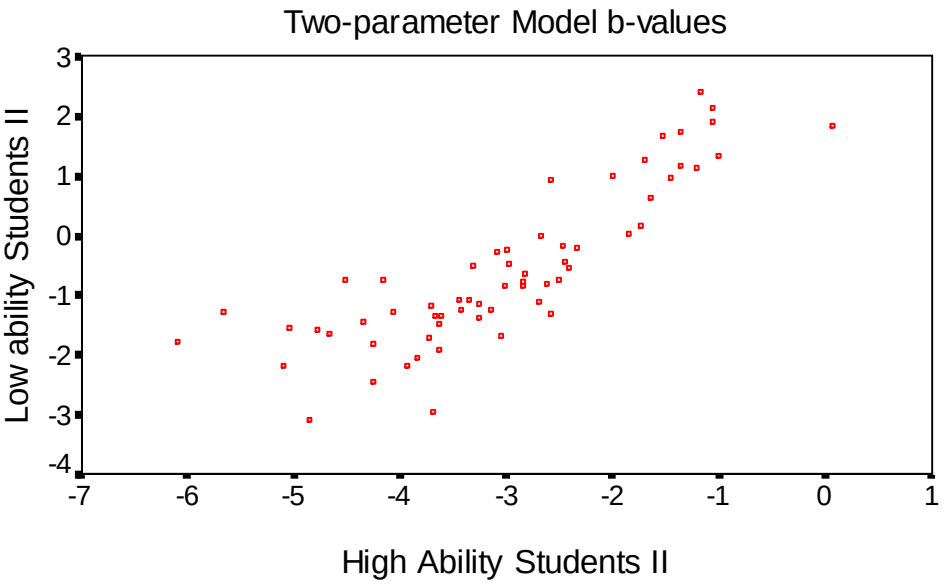
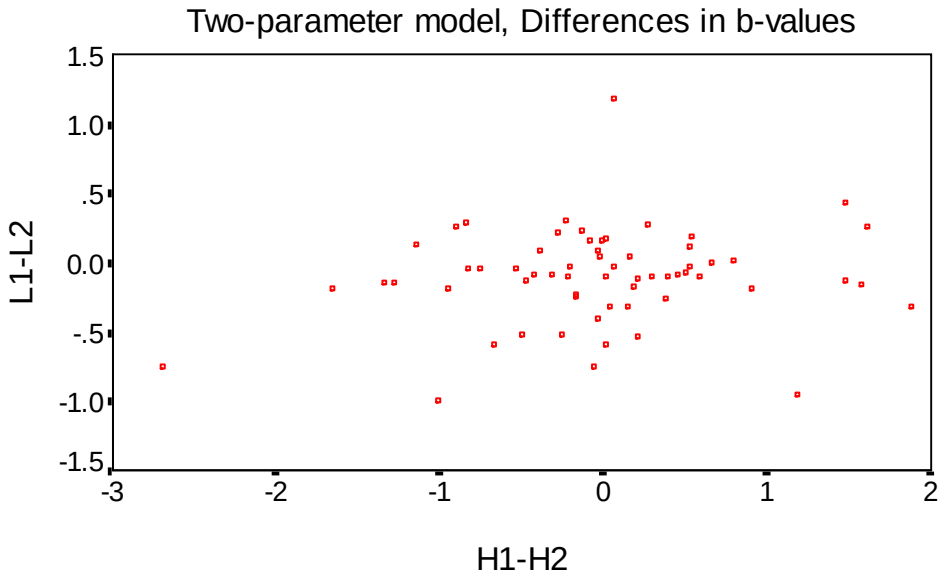




Figure 21. Plots of b-values for the two-parameter model obtained from two different high-low comparisons (N=500, r=.863 and r=.847 respectively).

There is a strong positive correlation (r=.97) between the b-values of the two samples of low ability students. The two groups of high-low comparisons both revealed positive correlations of approximately .85, attenuated possible due to inaccurate b-value estimation for high ability students.



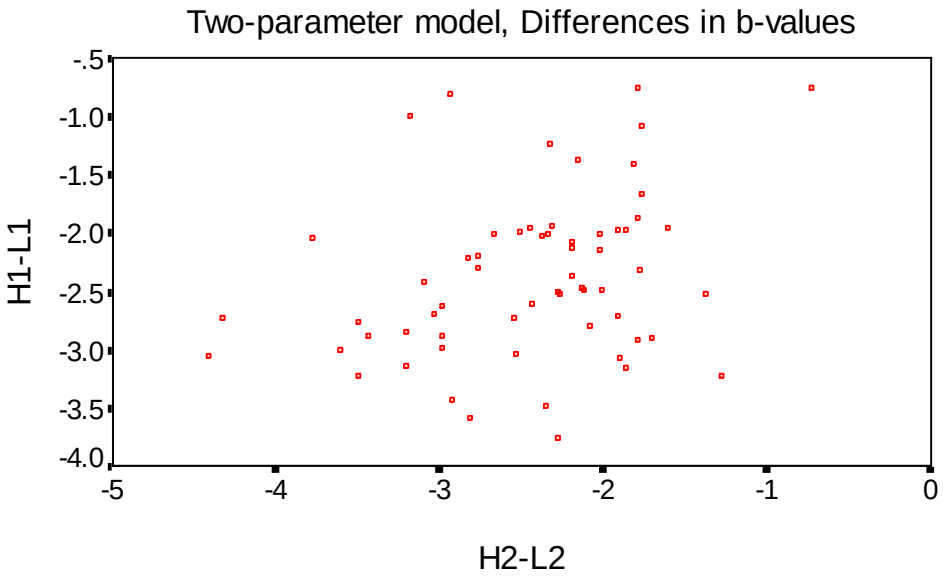
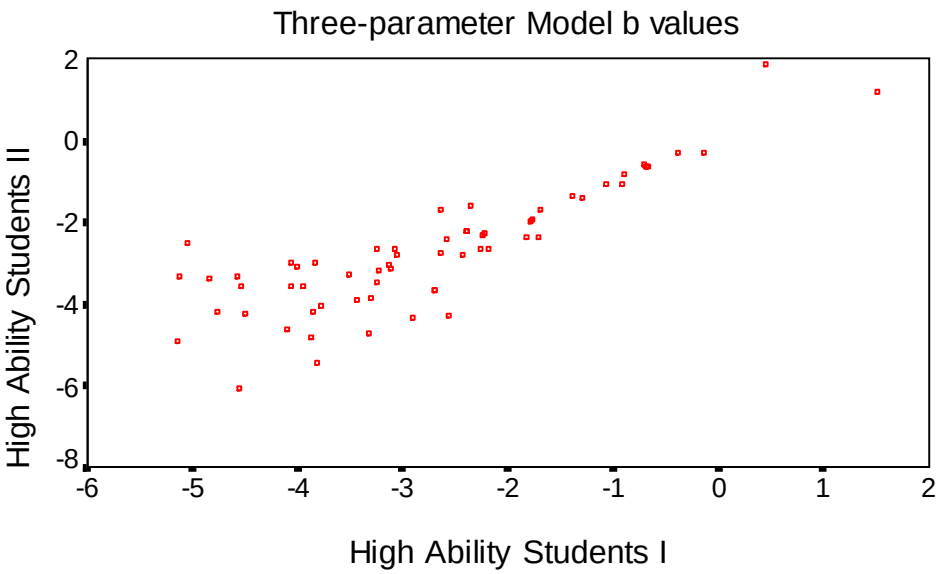




Figure 22. Plots of b-value differences, r=.148 and r=.307, respectively.



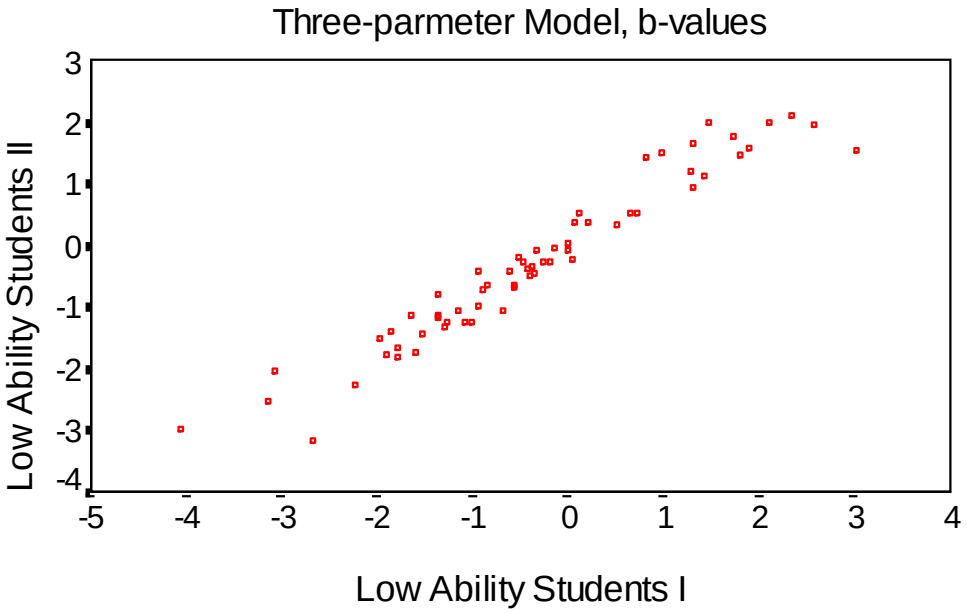




Figure 23. Plots of b-values for the three-parameter model obtained from two equivalent high performing students (n=500, r=.853) and two equivalent low performing students (n=500, r=.968).

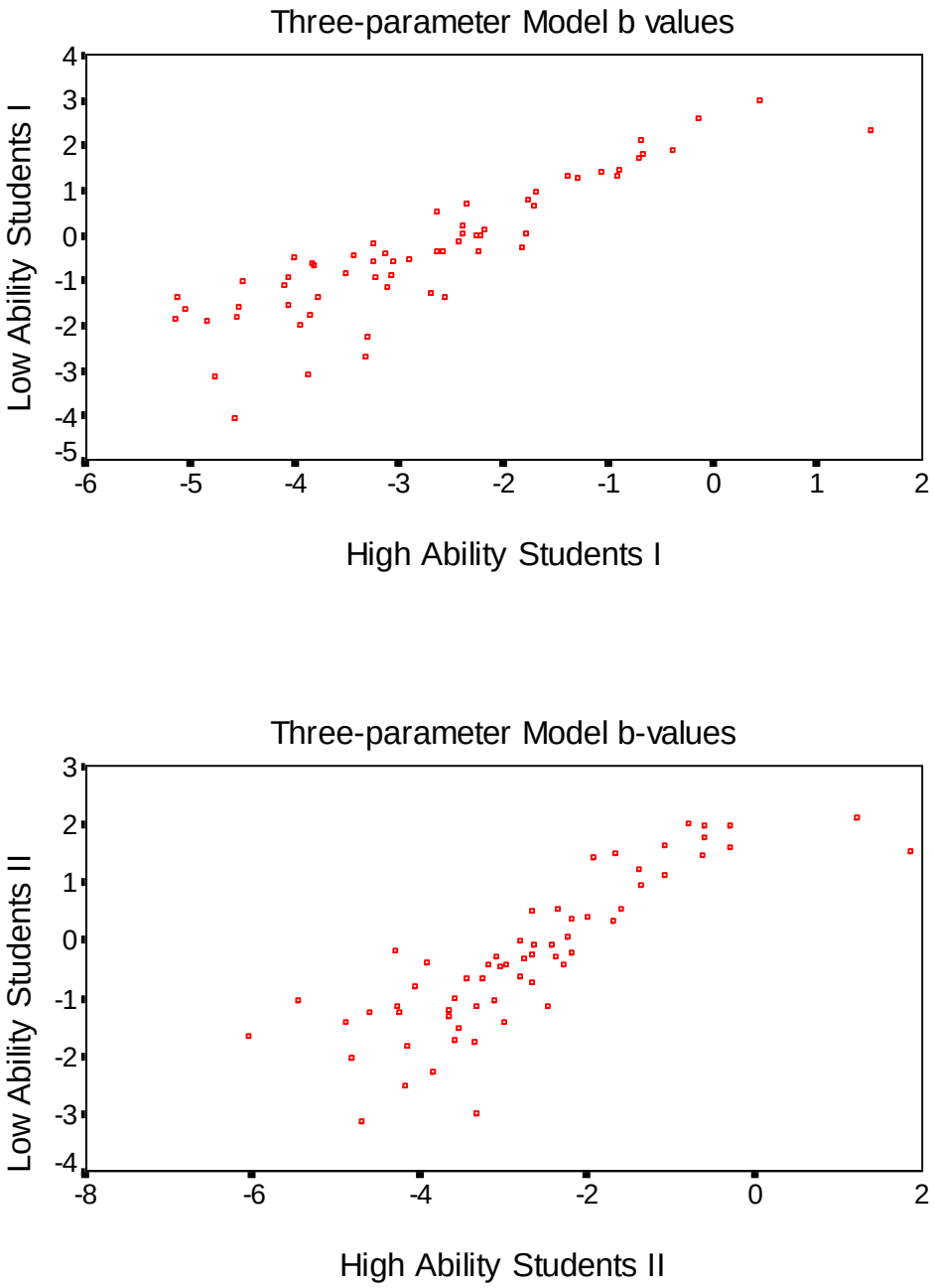



Figure 24. Plots of b-values for the three-parameter model obtained from two different high-low comparisons (n=500, r=.881 and r=.84 respectively).

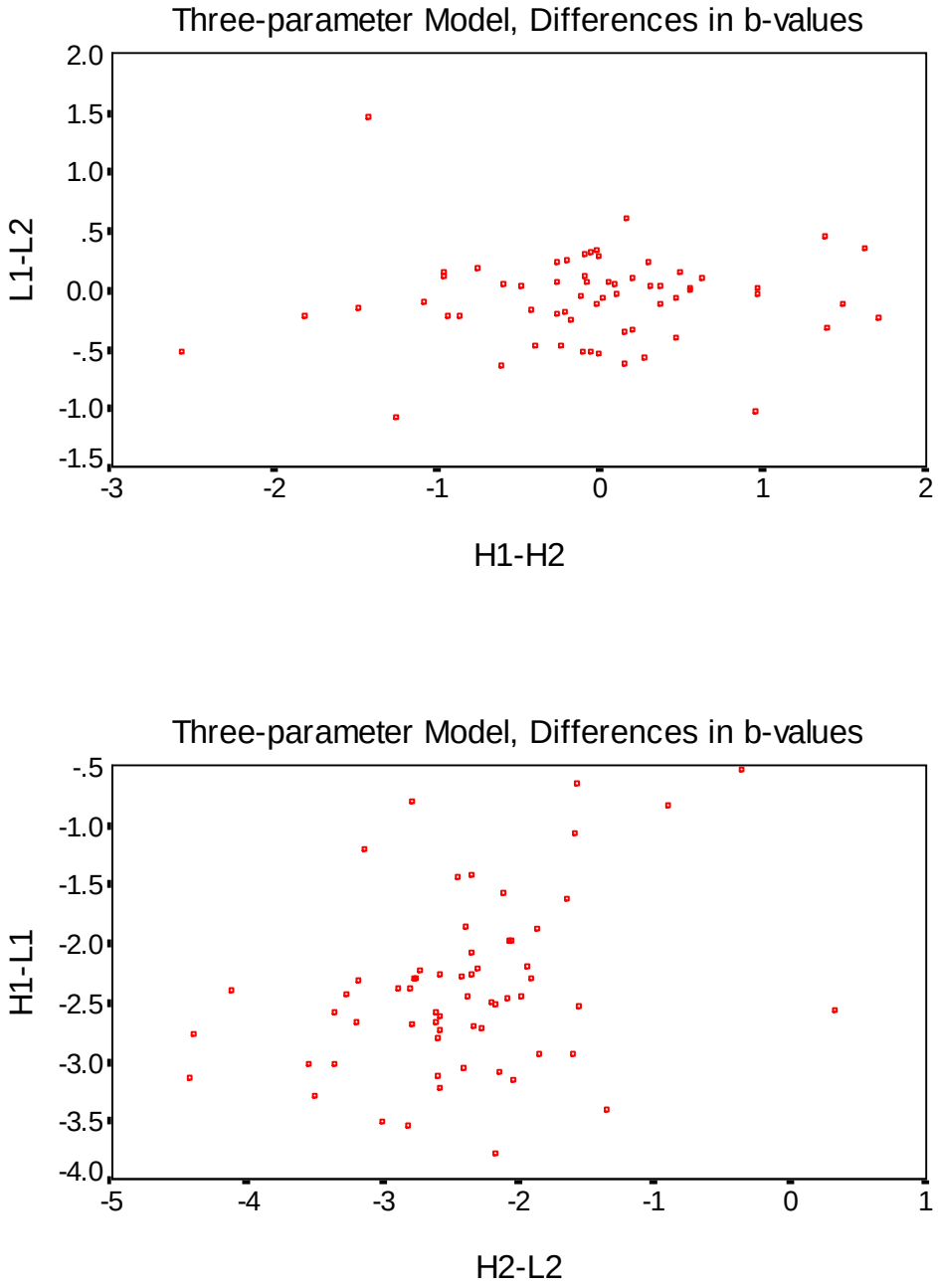



Figure 25. Plots of b-value differences, r=.041 and r=.34 respectively.

Figure 22 shows that the high-low correlation between two-parameter model b-value differences is not zero, but is much closer to zero than the one parameter model. The smaller b-value difference score correlation indicates that the two-parameter model calibration provided more consistent estimates of item difficulty than did the one-parameter model.

Figure 24 reveals a larger amount of scatter in the three-parameter item difficulty estimates obtained from the high-low comparisons as compared to the baseline plots in Figure 23. The baseline plots of b-value differences between the two high and two low ability samples is .041.

Figure 25 shows that the correlation of the difference in difficulty estimates that exists between the high-low comparison is .34 suggesting that the test items are being calibrated at similar difficulty levels for these groups and the feature of item parameter invariance can not be ruled out.

Table 10 contains observed Fisher's z scores that compare the correlations of high-low comparisons. The correlation of difficulty estimates of sample I, high and sample I, low ability students (sample I) is tested against the correlation of b-values obtained from sample II high and sample II low ability students (sample II) for the one- two- and three-parameter models. For example, the test of the difference between two independent correlations reveals that the correlation of sample I one-parameter b values to sample II three-parameter difficulty estimates results in an insignificant observed Fisher's z-score of 1.93. The comparison of the correlation coefficients of the one-parameter, sample II to three-parameter sample I b-values is also insignificant.

Table 10.

Fisher's z-test for the difference between two independent correlations

| | | Sample II | | |
|---|---|---|---|---|
| | Model | 1-p | 2-p | 3-p |
| Sample I | 1-p | - | 1.8 | 1.93 |
| | 2-p | 1.27 | - | .91 |
| | 3-p | 1.32 | -.73 | - |

Note: Tabulated values represent observed Fisher's z-scores

For each pairwise high-low comparison, the Fisher's z test fails to reject the null hypothesis of the equality of population correlations. This result suggests that the degree of relationship between item difficulty estimates is no different among the one-, two- and three-parameter models.

Plots of high-low item difficulty differences provide the most revealing results. If item parameters are truly invariant then one would expect the estimation of item parameters to be fairly close for all subsamples. This is verified by correlating item parameters for high-high, low-low and low-high ability groupings. In each case, the correlations were relatively high, ranging from .79 to .99. This indicates that there is a relatively strong relationship between the estimates of the item parameters between the groups. Although the relationship is not perfect, large item parameter estimates in one group are associated with large item parameter estimates in the comparison group. Likewise, as the estimates of the item parameters decrease for one group, the comparison group is comprised of similarly decreasing item parameter estimates. Since the

relationship is direct, and item parameter estimates in one ability group are associated with item parameter estimates in a comparison ability group that are approximately the same, the difference between these item parameters should be very close to zero. If there is no systematic tendency for the item parameter estimates of one ability group to vary with the item parameter estimates of a comparison ability group, then the correlation of difference scores should also be approximately zero. Based on the correlation of b-value differences, one can conclude that test items are functioning quite differently among the three models.

Based on the correlation of one-parameter model b-values differences, this model cannot possibly fit the data. The correlation of the differences between b-values obtained in the two samples is not zero, but very close to one, giving support to the argument that the feature of invariance of item parameters is violated for this model. The extent to which invariance has been achieved is similar in the two- and three-parameter models. The correlation of item difficulty estimates among high ability students is lower than expected for the two- and three-parameter models (.792 and .853 respectively). As a result, the correlation of high-low comparisons is attenuated while differences in difficulty estimates approach zero. The two- and three-parameter models both estimate item parameters similarly and remain the most plausible examples of model data fit. These two models have acceptable correlations for high-low comparisons and near zero correlations of differences in item difficulty estimates.

## Residual Analysis

Sample Elimination. Item parameter estimates obtained through BILOG calibration are derived through marginal maximum likelihood estimation. BILOG allowed simultaneous estimation of all parameters. The scaling factor D=1.7 is employed to scale estimates in the normal metric for the two- and three-parameter models. As is customary, the logistic, rather than the normal metric, is selected for the one-parameter calibration.

Omits and not reached items are treated as wrong answers. While maximum likelihood ability estimates are not available for examinees with zero or perfect scores, they are available for examinees with very high or very low observed test scores. The standard error of measurement for these extreme scores is very large and hence provides little interpretative value in comparison to other examinees. Therefore, examinees with ability scores greater than 3 or less than -3 were deleted from further analyses. For the one-parameter model, this resulted in the exclusion of 21 examinees. As a result, the lowest ability group was deleted. For the two- and three-parameter model, 40 and 25 respondents were dropped, respectively.

Description of Residuals. The magnitude of misfit can be obtained using the absolute values of the raw residuals (AVRR) and the absolute values of the standardized residuals (AVSR) presented in Table 11.

Table 11

Average and Absolute Average Raw and Standardized Residuals at Twelve Ability Levels

| | -2.75 | -2.25 | -1.75 | -1.25 | -0.75 | -0.25 | 0.25 | 0.75 | 1.25 | 1.75 | 2.25 | 2.75 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | | | | | | | | | | | | |
| 1-p | 20 | 34 | 94 | 152 | 285 | 435 | 396 | 245 | 200 | 74 | 44 | 1 |
| 2-p | 2 | 31 | 78 | 205 | 325 | 434 | 332 | 248 | 136 | 94 | 49 | 26 |
| 3-p | 18 | 37 | 81 | 160 | 283 | 385 | 388 | 333 | 187 | 78 | 29 | 6 |
| Raw | | | | | | | | | | | | |
| 1-p | 0.015 | 0.011 | 0.001 | 0.006 | 0.014 | 0.022 | 0.03 | 0.025 | 0.03 | 0.025 | 0.21 | |
| 2-p | 0.613 | 0.024 | 0.021 | 0.008 | 0.003 | 0.004 | 0.003 | 0.007 | 0.003 | 0.001 | 0.003 | -0.002 |
| 3-p | 0.005 | 0.008 | -2 | 0.002 | 0.006 | 0.002 | 0.003 | 0.005 | 0.005 | 0.004 | 0.001 | 0.0002 |
| \|Raw\| | | | | | | | | | | | | |
| 1-p | 0.13 | 0.09 | 0.07 | 0.06 | 0.05 | 0.05 | 0.06 | 0.05 | 0.04 | 0.05 | 0.03 | |
| 2-p | 0.63 | 0.09 | 0.05 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 |
| 3-p | 0.08 | 0.07 | 0.04 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 |
| Std. | | | | | | | | | | | | |
| 1-p | 0.42 | 0.16 | 0.28 | 0.36 | 0.83 | 1.61 | 2.1 | 1.6 | 1.43 | 0.94 | 0.63 | |
| 2-p | -4.61 | 0.7 | 0.6 | 0.36 | 0.23 | -0.13 | -0.05 | -0.42 | -0.26 | -0.56 | -0.75 | -0.05 |
| 3-p | 0.06 | 0.11 | -0.05 | 0.08 | 0.3 | 0.18 | -0.18 | -0.36 | -0.34 | -0.44 | -0.34 | 0.03 |
| \|Std.\| | | | | | | | | | | | | |
| 1-p | 1.87 | 1.34 | 1.64 | 1.68 | 2.08 | 2.72 | 3.21 | 2.58 | 2.06 | 1.41 | 1.02 | |
| 2-p | 4.68 | 1.53 | 1.15 | 0.91 | 1.03 | 1.02 | 0.97 | 1.06 | 0.81 | 1.32 | 1.47 | 0.52 |
| 3-p | 0.83 | 0.84 | 0.76 | 0.73 | 0.86 | 0.82 | 0.87 | 0.92 | 0.87 | 0.95 | 0.88 | 0.24 |

The large values of the one-parameter model residuals compared to the two- and three-parameter models suggests the more general models provide a better fit to the data across ability groups in all but one case. For the two-parameter model in the -2.25 ability category, the AVSR provides higher measures of misfit than did the one-parameter model for the lowest ability group. Dismissing the lowest ability group due to inadequate sample size better fits were obtained with the more general models, regardless of ability level. For the one-parameter model, the average absolute-value raw and standardized residuals correlate .928 reflecting the fact that they describe fit in a similar way. For the more general models, these correlations reduce to .473 and .495, respectively. This reduction in correlation may be due to restriction of range.

Raw residuals are sensitive to the amount of misfit in both directions and thereby provide different measures of fit as compared to standardized residuals (Hambleton, 1985). By considering sample size and sampling errors associated with the average observed performance ($p_{ij}$), standardized residuals provide more accurate estimates of fit. Because sample sizes vary, sometimes significantly, across ability groups and across the three models, further analyses will concentrate on interpretations using standardized residuals.

Table 12 summarizes the number and percentage of absolute-value standardized residuals obtained from the Residual Analysis (RA) programs. Since the residuals are assumed to represent a sample from a standard normal distribution, we would expect approximately 95% of these points to be contained in the interval (-1.96, 1.96). For these data, the three-parameter model has

approximately 5% of absolute-value standardized residuals exceeding a value of 2 whereas the one- and two-parameter models have respectively 41.6% and 16.4% exceeding this value. Of the 45 three-parameter model standardized residuals greater than 2, 17 of these standardized residuals are estimated to be greater than two for the two-parameter model and the one-parameter model estimates 15 of then to be greater than 2. The three models are in agreement in only 10 cases for outliers greater than two.

Table 12.

Absolute-Value Standardized Residuals for the One-, Two- & Three-Parameter Models

| | One-Parameter Model | | Two-Parameter Model | | Three-Parameter Model | |
|---|---|---|---|---|---|---|
| | Frequency | Percentage | Frequency | Percentage | Frequency | Percentage |
| $0 < \|SR\| < 1$ | 263 | 31.1 | 534 | 57.8 | 655 | 70.9 |
| $1 < \|SR\| < 2$ | 232 | 27.4 | 238 | 25.8 | 224 | 24.2 |
| $2 < \|SR\| < 3$ | 170 | 20.1 | 77 | 8.3 | 31 | 3.4 |
| $\|SR\| \geq 3$ | 182 | 21.5 | 75 | 8.1 | 14 | 1.5 |

Note. For the one-parameter model, there are 836 standardized residuals (76 items and 11 ability levels)

For standardized residuals, the use of cutoffs produces some assurance of goodness-of-fit but numerical criteria are no substitute for graphical examination. Figures 26, 27and 28 show histograms with an overlaid normal curve, a box plot, confidence intervals for the mean, median and a table of statistics.

When comparing the mean, median and standard deviations, there are obvious differences in the standardized residuals for each logistic model. The mean standardized residual approaches zero as the number of parameters in the model increases. The confidence interval for the mean contains zero for the three-parameter model, whereas the upper limit is less than zero (-.245) for the two-parameter model and the lower limit is greater that zero (.792) for the one-parameter model. The distributions of standardized residuals for the two- and three-parameter models are characterized by a restriction of range about their medians of approximately zero with a moderate number of outliers resulting in peaked sample distributions. The distribution of standardized residuals for the one-parameter model is characterized by an abundance of outliers centered its median of one.

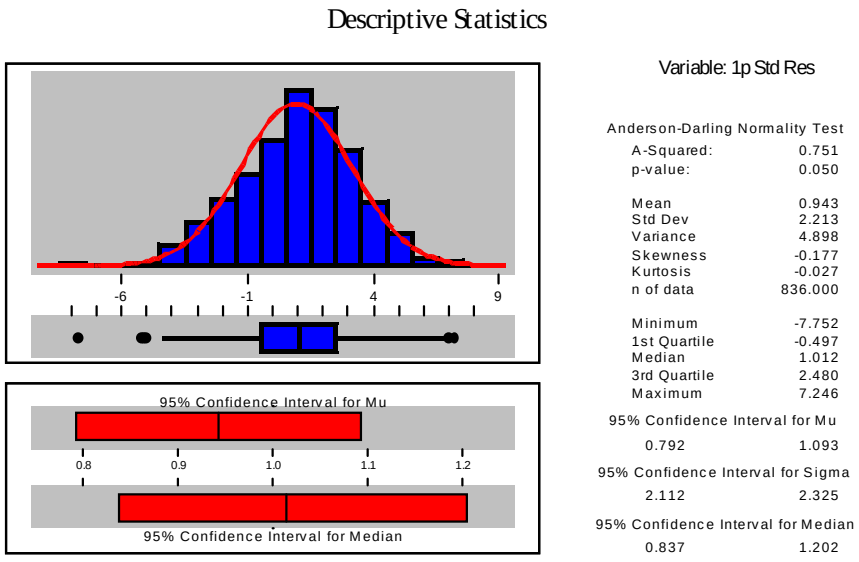



Figure 26. Descriptive statistics for the one-parameter model standardized residuals.

Measures of skewness indicate that all of the models tend to have a majority of standardized residuals that cluster to the right. This indicates that the three logistic models tend to underestimate performance. This tendency is largest for the Rasch model and smallest for the three-parameter model.

When the sample size is large, most any goodness-of-fit test will result in rejection of the null hypothesis. Since our sample size is large, a comparison of the observed significance level as well as the actual departure from normality must be considered. Based on small observed significance levels, the Anderson-Darling test and large kurtosis values, the hypothesis of normality of standardized residuals can be rejected for the two- and three-parameter models. Measures of kurtosis and skewness for the one-parameter model are representative of those from a sample that comes from a normally distributed population. A comparison of the histograms will support this view.

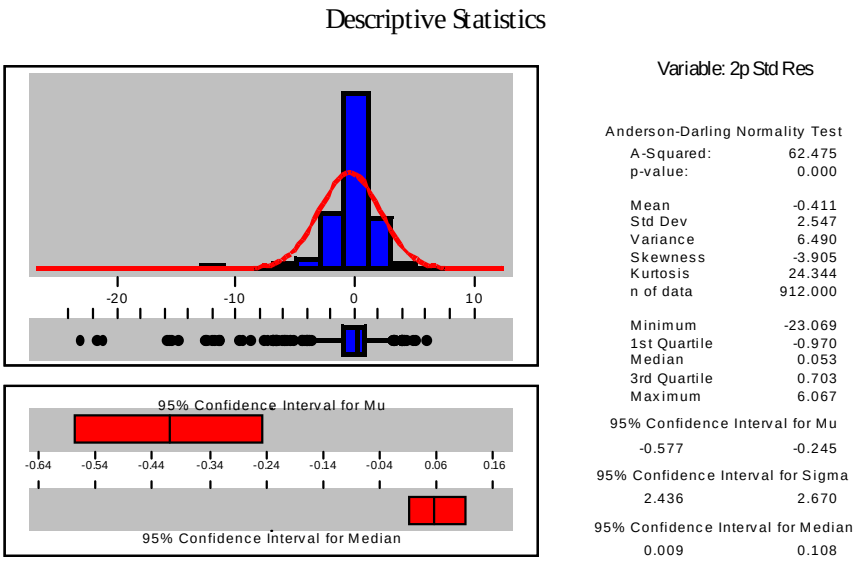



Figure 27. Descriptive statistics for the two-parameter model standardized residuals.

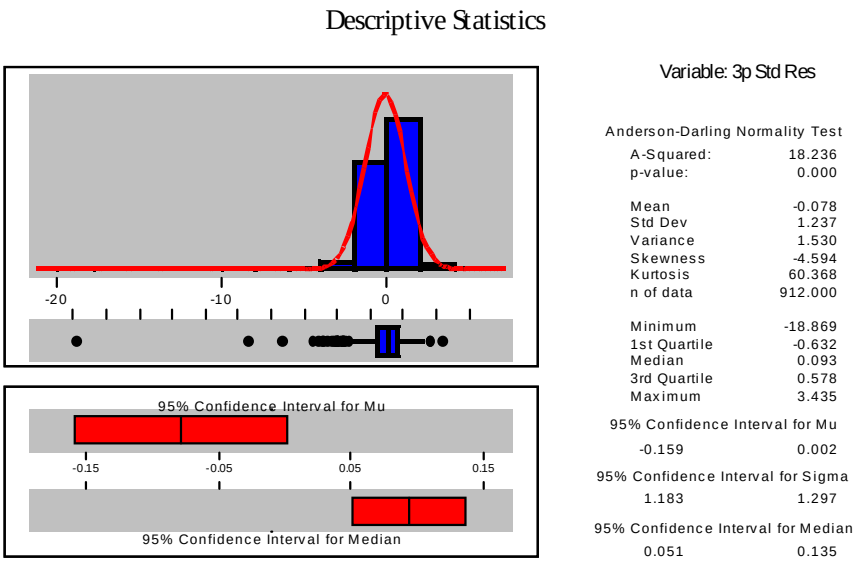



Figure 28. Descriptive statistics for the three-parameter model standardized residuals.

Test data that adequately fit a model are represented by standardized residuals that are small and randomly distributed about zero. The validity of the normality assumption for the standardized residuals is questionable for the two- and three-parameter models. However, this is not seen as evidence of misfit, because if a model is a perfect fit to some observed dataset all residuals are zero. It is encouraging that these models have so few outliers. In addition, during the development of the DRP, items generated are those which fit the Rasch Model. This item preference possibly increased the likelihood of normally distributed residuals for the Rasch model.

Equal Discrimination Indices. Average absolute-value standardized residuals are computed by averaging the absolute values of the standardized residuals

across all ability levels. The ith AAVSR represents the magnitude of misfit for the ith Item (i=1, ... ,76). The impact of the use of a discrimination parameter is highlighted by plotting average absolute-value standardized residuals (AAVSR) versus item discrimination as shown in Figure 29.

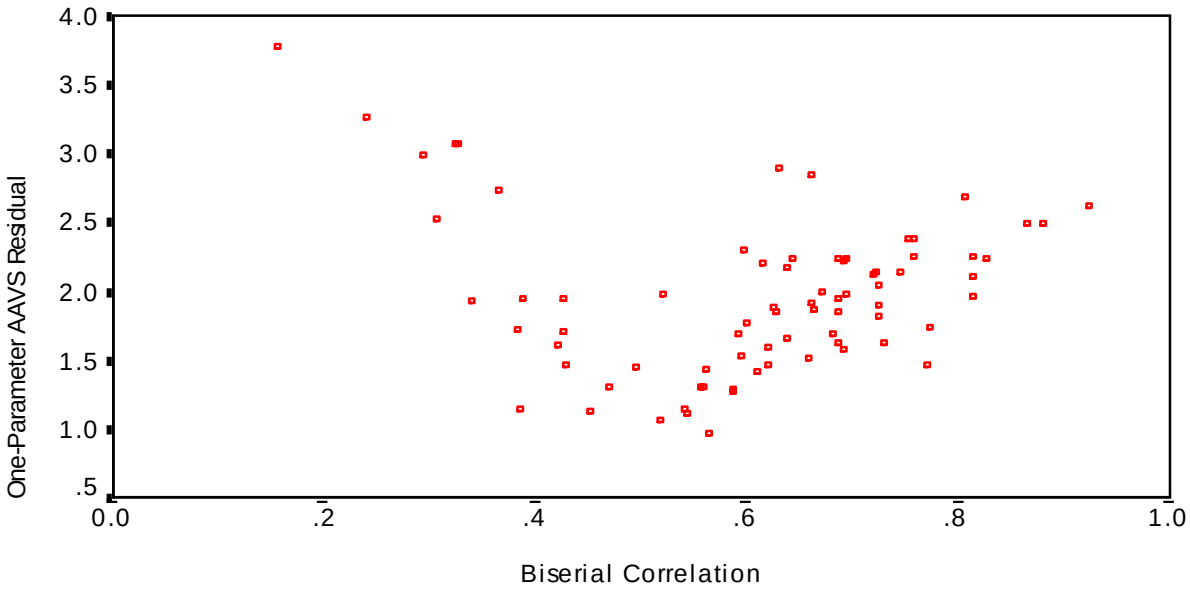



Figure 29. Scatterplot of one-parameter model average absolute-value standardized residual and biserial correlations.

The relationship between item discrimination and the one-parameter model AAVSR reveals a curvilinear relationship. Items with low and high bisreial correlations tend to have high AAVSRs. The one-parameter AAVSRs are large and have a greater variation for low discriminating items than for highly discriminating items. Similarly, Figures 30 and 31 display the same plots for the two- and three- parameter models. The curvilinear relationship that is apparent with the one-parameter model disappears when the two- and three-parameter models are fit to the data.

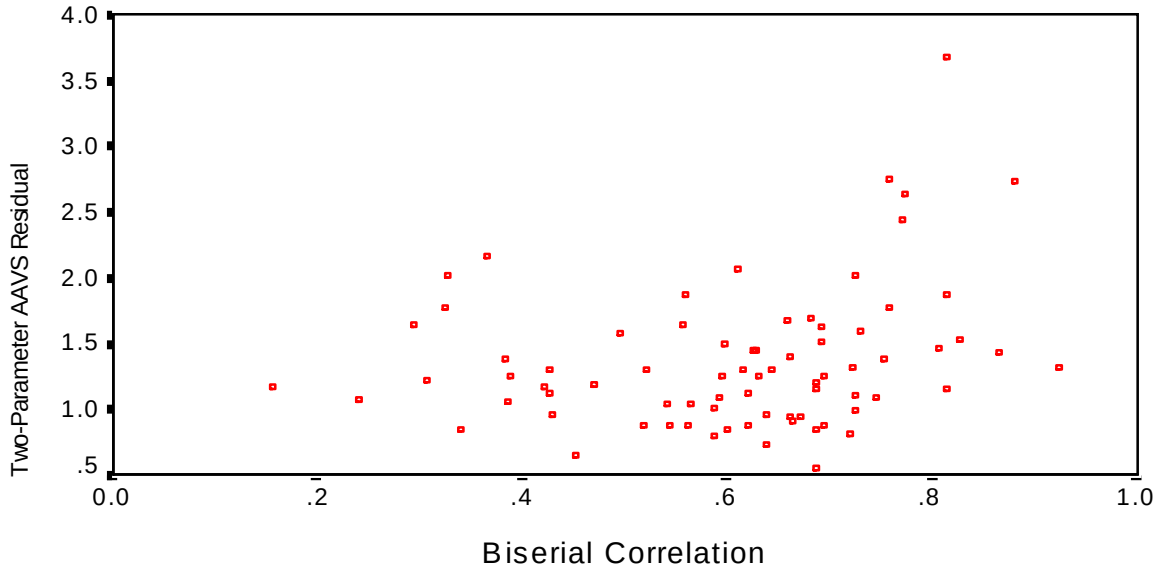



Figure 30. Scatterplot of two-parameter model average absolute-value standardized residual and biserial correlations.

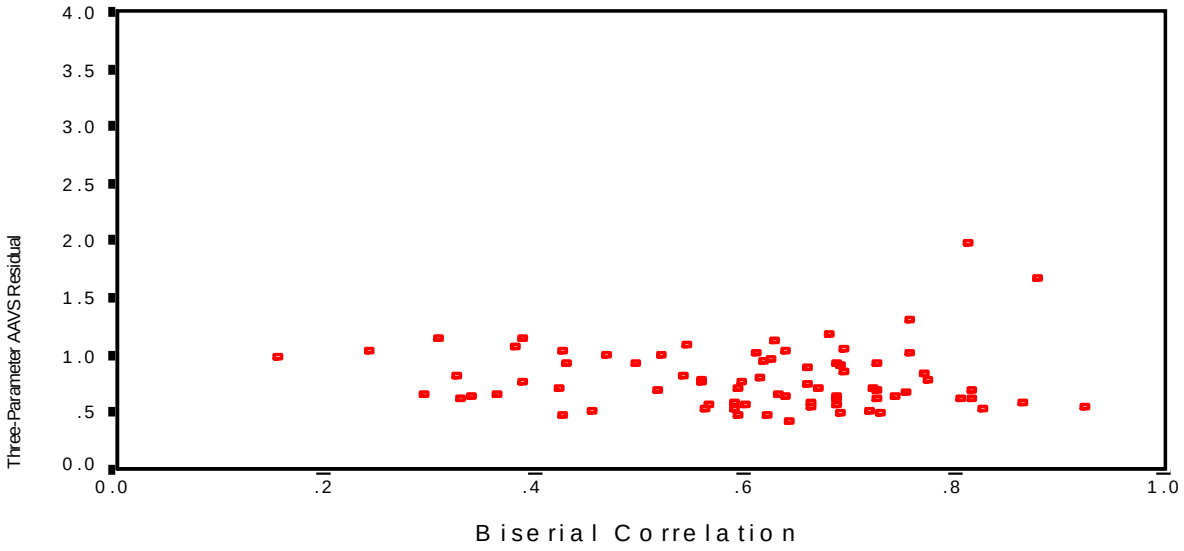



Figure 31. Scatterplot of three-parameter model average absolute-value standardized residual and biserial correlations.

The size of misfit is generally smaller for the three-parameter model as compared to the two-parameter model. The AAVSR of the two- and three-parameter models tend to be small and vary more or less homogeneously for both low and high discriminating items. For the two-parameter model a slightly wider variation of misfit was found among highly discriminating items as compared to the three-parameter model. This is easily confirmed upon examination of Table 13.

Table 13.

Relationship between Biserial Correlations and Averaged Absolute-value Standardized Residuals

| | | Discrimination Indices | | | |
|---|---|---|---|---|---|
| Model | AAVSR | (4)[a] | (14)[a] | (33)[a] | (25)[a] |
| | | 0-.25 | ,25-.50 | .50-.75 | .75-1 |
| | 0 to 1 | 0 | 0 | 1 | 0 |
| 1-parameter | 1.01 to 2 | 0 | 11 | 21 | 12 |
| | Over 2 | 2 | 5 | 9 | 15 |
| | 0 to 1 | 0 | 3 | 13 | 3 |
| 2-parameter | 1.01 to 2 | 4 | 9 | 19 | 16 |
| | Over 2 | 0 | 2 | 1 | 6 |
| | 0 to 1 | 2 | 9 | 27 | 21 |
| 3-parameter | 1.01 to 2 | 2 | 5 | 6 | 4 |
| | Over 2 | 0 | 0 | 0 | 0 |

Note. AAVRS= Averaged absolute-value standardized residuals. [a]Number of biserial correlations in the corresponding category.

The lack of homogeneity across items by the one-parameter model indicates that a model that accounts for the variation in the discrimination power of test items is more appropriate. As the two- and three-parameter models provide substantial improvement in fit over the Rasch model, the assumption of equal discrimination is untenable for these data.

Item Difficulty. Item difficulty ranged from .296 to .98 where 78% of the items have a p-value of .5 or greater. A relationship between the one-parameter model absolute-value standardized residuals (AVSR) and classical item difficulties is revealed through the inspection of Figure 32. Approximately 53% of AVSRs are associated with hard items ($p \leq .5$) greater than 2, whereas 35% of large AVSR are associated with easy items. This tendency for hard items to have high residuals is possibly due to examinee guessing.

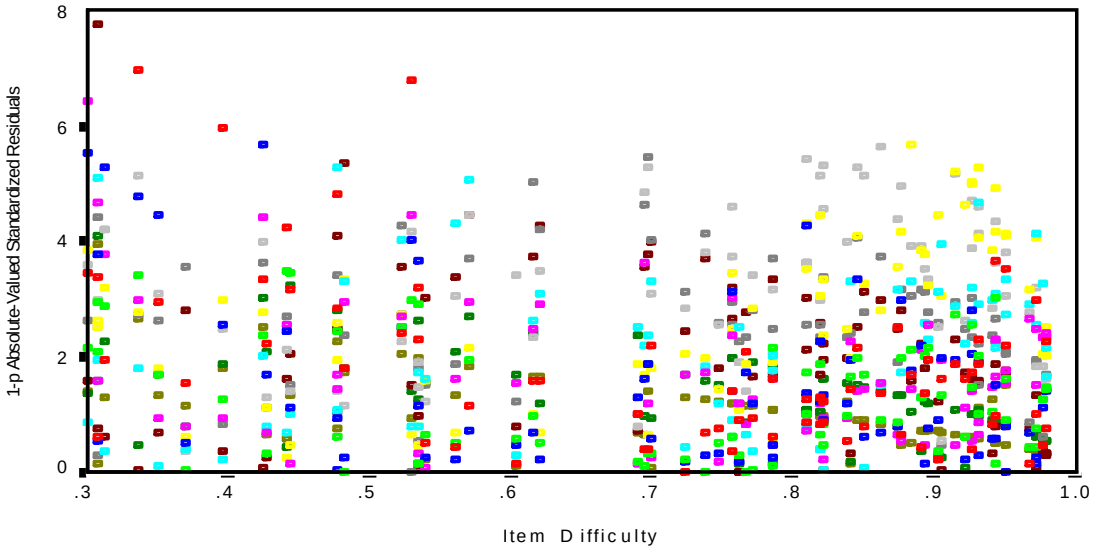



Figure 32. Scatterplot of one-parameter model absolute-value standardized residuals and item difficulty.

Figures 33 and 34 present scatterplots of AVSR plotted against classical item difficulty indices for the two- and three-parameter models. As can be seen in these scatterplots, the residuals are substantially smaller.

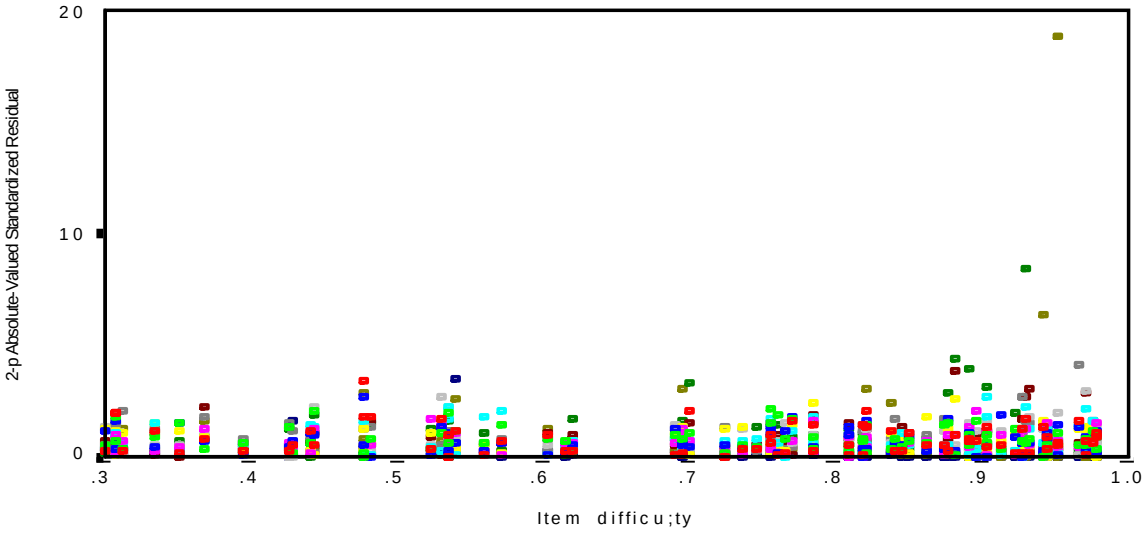



Figure 33. Scatterplot of two-parameter model absolute-value standardized residuals and item difficulty.

The two-parameter mode only has 4% of AVSR greater than 2. For hard items, 20% of the associated AVSR are greater than 2; 15% of AVSR have values greater than 2 for easy items. The three-parameter model has 8% of AVSR greater than 2, half of which account for hard items. Based on the reduction in the number of large AVSR obtained for difficult items, it appears that estimating item lower asymptote has been beneficial.

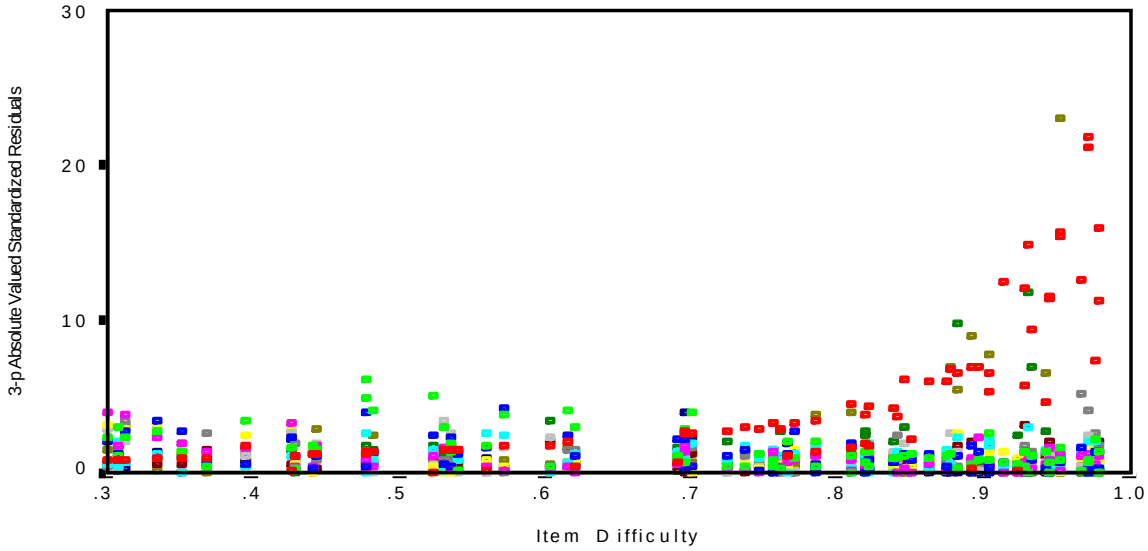



Figure 34. Scatterplot of three-parameter model absolute-value standardized residuals and item difficulty.

Percent correct scores indicate the average item performance is high (71.8% correct) suggesting that the 3-p model might be of little utility. However, 30% of the items have a percent correct rate of less than .6 implying the possibility of guessing by low ability students. In order to determine the usefulness of a lower asymptote, the AVSR are sorted by easy and hard items and reported for each model based on whether the AVSR is representative of fit (AVSR less than 1) or misfit (AVSR greater than 1). The one-parameter model provided 98.7% of fit indices greater than one regardless of the level of item difficulty (Table 14). Better fits were obtained when the two- and three-parameter models were fit to the test data.

Table 14.

Absolute-value Standardized Residual by Item

| AVSR | One Parameter Model | | Two Parameter Model | | Three Parameter Model | |
|---|---|---|---|---|---|---|
| | N | % | N | % | N | % |
| \|SR\|<1 | 0 | 0 | 2 | 2.6 | 11 | 14.5 |
| \|SR\|>1 | 16 | 21.1 | 14 | 18.4 | 5 | 6.6 |
| \|SR\|<1 | 1 | 1.3 | 17 | 23.4 | 48 | 63.2 |
| \|SR\|>1 | 59 | 77.6 | 43 | 56.6 | 12 | 15.8 |

The three-parameter model accounted for the largest percentage of AVSR less than or equal to one for the easy items (14.5%) as well as for the hard items (63.2%). This finding suggests that examinee guessing was an important factor with hard items. The one-parameter model was not able to account for this behavior resulting in large AVSR whereas the adjustment made by the three-parameter model resulted in substantially better fits.

Inspection of standardized residuals (Figures 35, 36, and 37) allow for comparisons of the direction of prediction. The variation of the standardized residuals about zero for the two- and three-parameter models is fairly uniform across the item difficulty scale. These models tend to overestimate the performance on easy items. On the other hand, the Rasch model tends to underestimate examinee performance, especially for easy items. For the Rasch model, examinee performance is underestimated 68.7% of the time. Of underestimated residuals, 83.1% occurred for easy items.

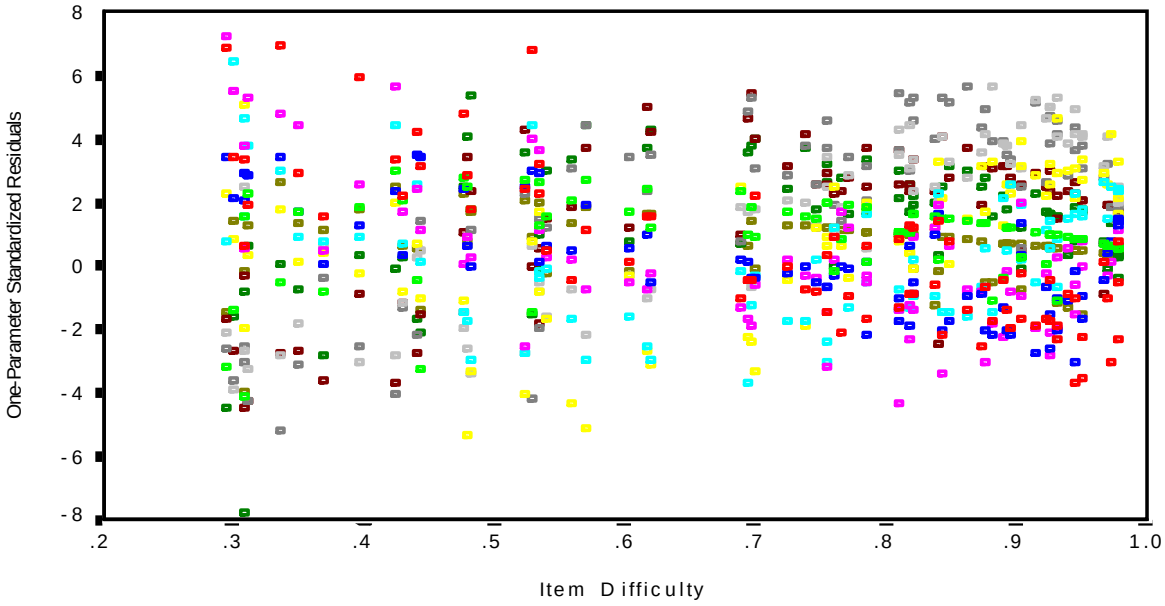



Figure 35. Scatterplot of one-paramter model standardized residuals and item difficulty.

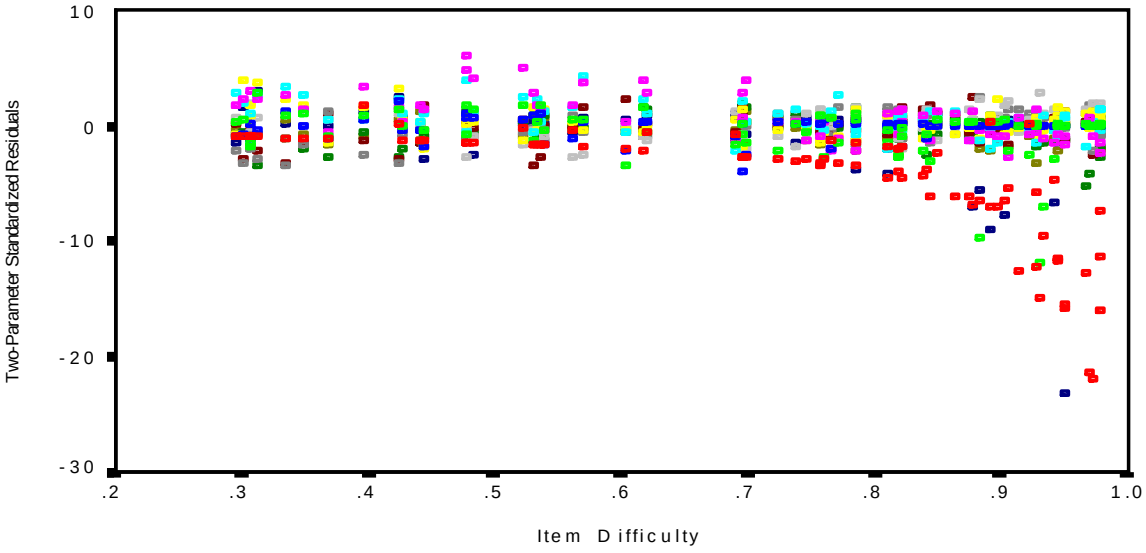



Figure 36. Scatterplot of two-paramter model standardized residuals and item difficulty.

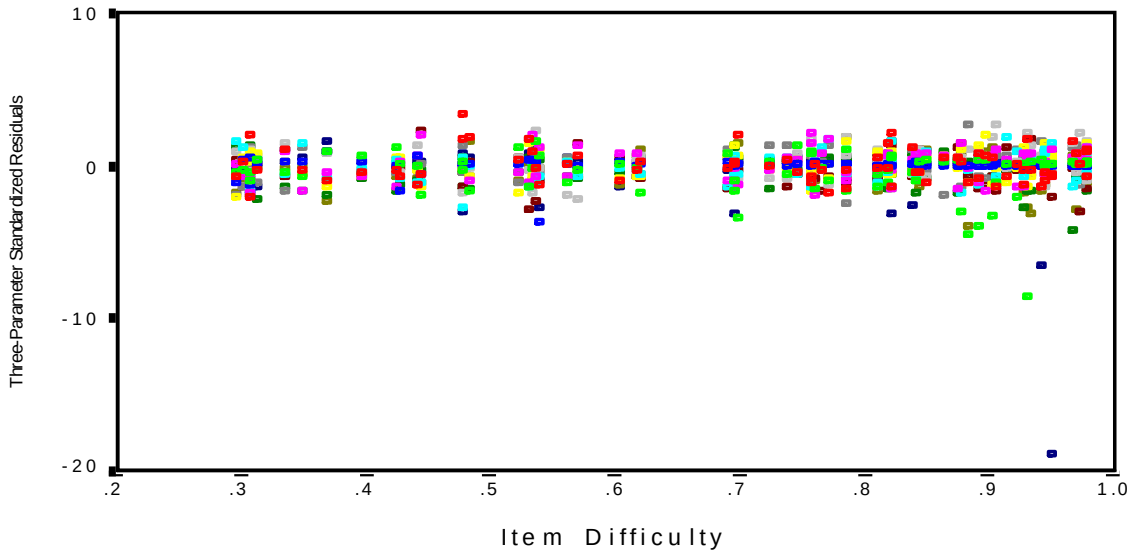



Figure 37. Scatterplot of three-paramter model standardized residuals and item difficulty.

Standardized Residual Plots. In order to determine why certain items fit or misfit each model, the SR for each item across ability groups was compared. A perfect model fit would lead to SR of zero and produce a horizontal line with an intercept and slope equal to zero. The plots of SRs for three typical items are presented in Figures 38 to 40. As can be seen from these figures, there are considerable differences in model fit.

Figure 38 is representative of easy items with low biserial correlations. The classical statistics show the item as having a low discrimination (r=.25) and being very easy (p=.98). Residual plots like those in Figure 39 were obtained for hard items with low biserial correlations. This item has a p-value of .3 and discrimination index of .26. Items with moderate difficulty and discrimination such as item 49 (p=.62, r=.53) had similar fit indices for the two- and three-

parameter models (see Figure 40). Because the two- and three-parameter models take varying item discrimination into account, these models provided a better fit than the one-parameter model.

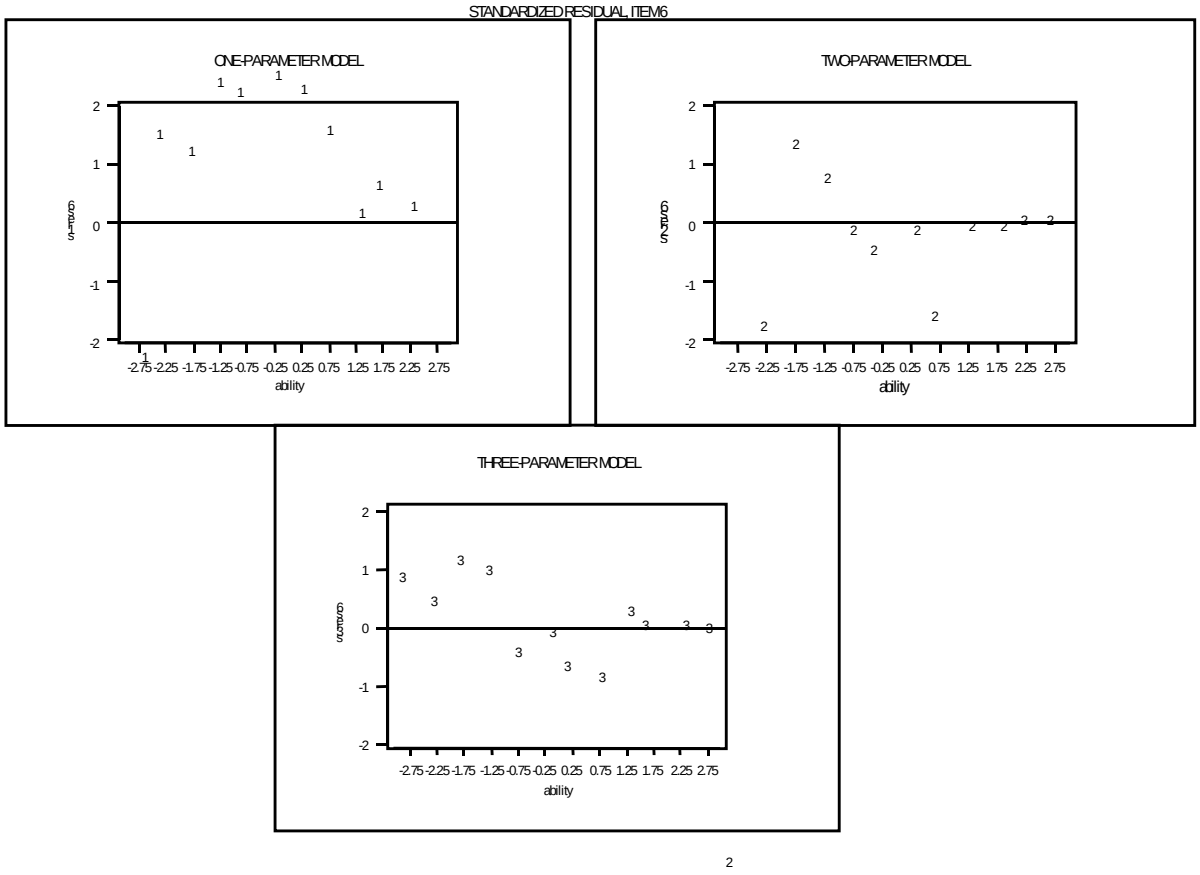



Figure 38. Scatterplots of the one-, two- and three-parameter model standardized residuals with ability for item 6 (r =.25, p=.98).

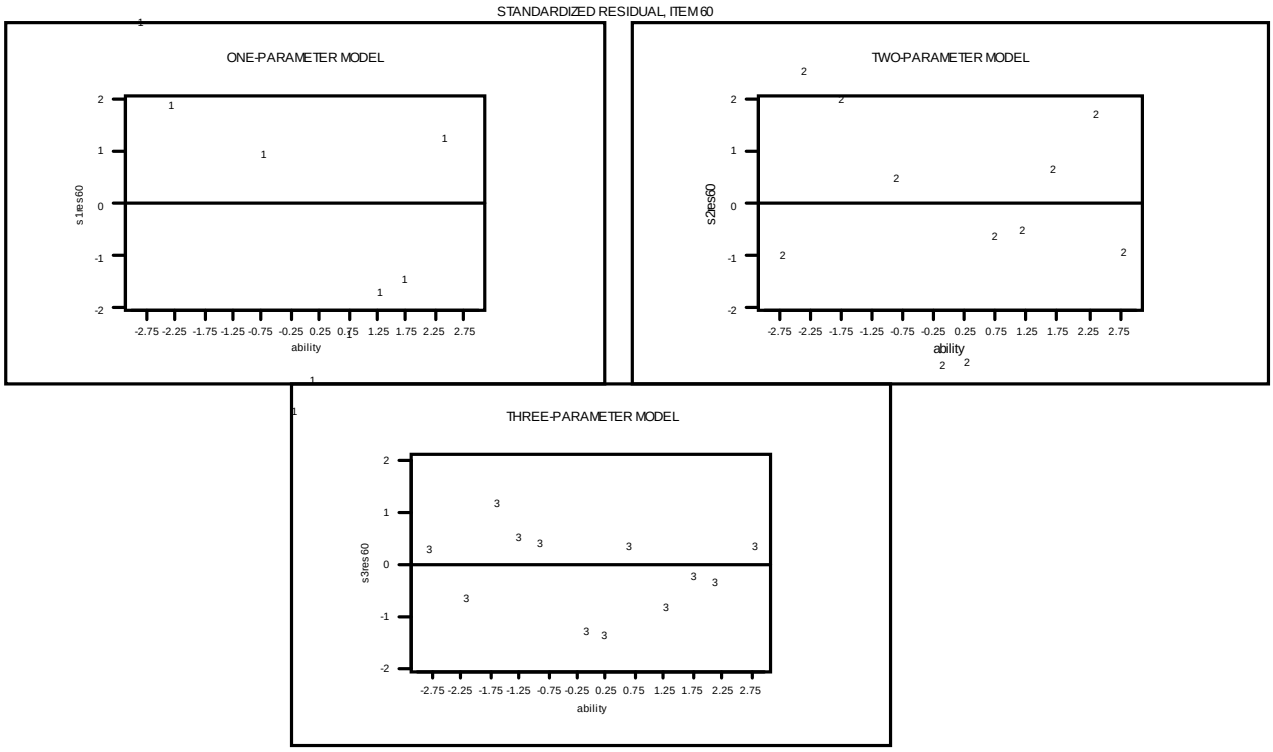



Figure 39. Scatterplots of the one-, two- and three-parameter model standardized residuals with ability for item 60 (r =.26, p=.30).

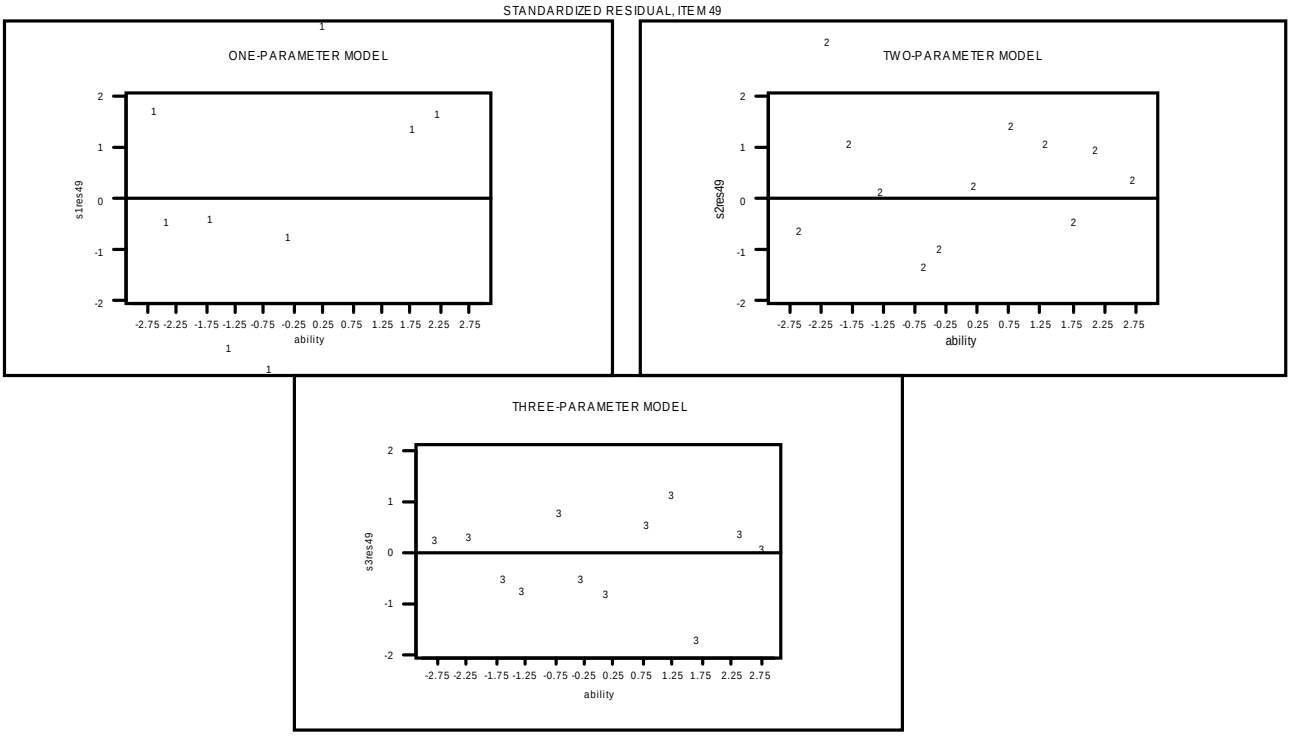



Figure 40. Scatterplots of the one-, two- and three-parameter model standardized residuals with ability for item 49 (r =.53, p=.62).

In general, the one-parameter SR vary greatly from zero, indicating poor model fit, while the SR of the two-and three-parameter models tend to cluster about zero, characterizing models that fit reasonably well. Because the two-parameter model takes discrimination into account, the SRs for this model are smaller than those obtained for the one-parameter model. The three-parameter model, accounting for both discrimination and lower asymptote, provides the best fit and most consistent estimates of performance. Notice that the one- and two-parameter model provide poor estimation for the lowest ability group, while the three-parameter model makes significantly better estimations.

In the effort to determine whether the Rasch model fits the DRP, one must consider whether the use of a discrimination parameter and lower asymptote is advantageous. The benefit of a discrimination parameter is highlighted in Figures 33 and 34 where the curvilinear relationship between the residuals and item difficulty dissipates when a discrimination parameter is introduced. The Rasch model's inability to account for varying item discrimination resulted in large standard residuals for most items. On the other hand, the two- and three-parameter models fit the DRP better than the Rasch model. This is evidenced by the small number of large residuals . The three-parameter model is the most parsimonious. The most prominent feature of this model is that it did a better job than the other logistic models in providing accurate estimates of ability, especially for the extreme ability groups where estimation is most difficult.