

Running Head: ROBUSTNESS OF LORD'S FORMULAS

Robustness of Lord's Formulas for Item Difficulty and Discrimination

Conversions between Classical and Item Response Theory Models

Teresa Dawber

W. Todd Rogers

Michael Carbonaro

Centre for Research in Applied Measurement and Evaluation (CRAME)

University of Alberta

Paper presented at the annual meeting of the American Educational Research
Association, San Diego, CA, April 12, 2004.

Abstract

Lord (1980) proposed formulas that provide direct relationships between IRT discrimination and difficulty parameters and conventional item statistics. The purposes of the present study were to determine (1) the veracity of the two formulas within the context that Lord proposed, and (2) the robustness of the formulas beyond the initial and restrictive conditions identified by Lord. Simulation and real achievement data were employed. Results from the simulation study indicate that the a -parameters were recovered quite well for low to moderately discriminating items regardless of ability distribution and the b -parameters were recovered quite well for the range typically found for achievement tests. Results of the real data were consistent with that found for the simulation study.

Robustness of Lord's Formulas for Item Difficulty and Discrimination

Conversions between Classical and Item Response Theory Models

The field of psychometrics encompasses different models that offer alternative frameworks for performing test and item analyses. The classical test score theory (CTST) model, the foundation of which was provided by Charles Spearman in 1904, is the traditional means of conducting item and test analyses. The family of item response theory (IRT) models, first introduced by Lord in 1952 for dichotomously scored items, was developed to circumvent the limitations of CTST. However, Lord (1980; also see Lord & Novick, 1968) proposed formulas that link the item difficulty and item discrimination indices of the CTST and the two-parameter IRT model.

Lord (1980, pp. 33-34) stipulated that under certain conditions the difficulty and discrimination indices derived from the two measurement frameworks are connected. That is to say, the classical item discrimination parameter may be used to predict the IRT discrimination parameter and the classical item difficulty parameter may be used to predict the IRT difficulty parameter.

For item discrimination, to the extent that number correct score x is a measure of ability (θ), the biserial correlation between the item and test score (ρ'_{ix}) is an approximation to the correlation between the item and ability estimate ($\rho_{i\theta}$). The association yields a relationship between the conventional biserial item-test correlation and the IRT discrimination index (a_i):

$$a_i \cong \frac{\rho'_{ix}}{\sqrt{1 - \rho'^2_{ix}}} .$$

Therefore, the IRT item discrimination parameter and the biserial correlation are approximately monotonic increasing functions of each other. Lord stated that the relationship is “valid only for the case where θ is normally distributed and there is no guessing” (p. 33). However, Lord qualified that the approximations are crude and may fall short because (1) the test score x contains errors of measurement while θ does not, and (2) x and θ have differently shaped distributions, since the relation between x and θ is nonlinear.

Lord also proposed a monotonic relation between the IRT difficulty index (b_i) and the classical difficulty index (π_i) when all items are equally discriminating. The relationship between the difficulty indices is described as $b_i \cong \frac{\gamma_i}{\rho'_{ix}}$. The difficulty parameter b_i is proportional to γ_i , the cut point on the continuous normal distribution underlying the binary item that separates the proportion of incorrect answers ($1 - \pi_i$), and the proportion of correct answers (π_i). Both b_i and γ_i decrease as π_i increases.

Review of Studies Using Lord's Formulas

The formulas provided by Lord (1980) were first presented in Lord and Novick (1968). Even though the relationships described are the same, the only qualifying condition in the earlier writing was that θ be normally distributed with a mean of zero and unit variance. Several studies were conducted in the mid to late 1970s using the formulas, also referred to as the heuristic method, within the framework of the three-parameter model. In 1980, Lord added the stipulation that the formulas were only applicable for the 2PL model. Despite the use of an incorrect IRT model, the following studies provide insight into how the formulas may function in the intended context.

Description of Studies

Using the formulas proposed by Lord and Novick (1968), Urry (1974) developed a graphical method. He devised graphs that consisted of mapping a grid system to model a - and b -parameters onto a coordinate system where the population point-biserial correlation, rather than the biserial correlation, is the ordinate and the population proportion passing an item is the abscissa. By plotting the data points for a given item using the conventional indices, the values of a and b may be interpolated. When there is no guessing, the graph is symmetric. When there is guessing, the graph is displaced to reflect inflation in the proportion passing the item and attenuation in the point-biserials through error due to guessing.

Urry (1974) proposed that the following four conditions needed to be met for effective application of the graphical method: (1) the latent trait is normally distributed; (2) the classical indices are based on large samples ($N = 2,000$) in order to approximate the set of parameters; (3) the items in the test must be homogenous ($K-R 20 \geq 0.90$); and (4) the items in the test must be of sufficient number ($n = 80$) for the point-biserial correlation between item and total test score to bear a close relationship to the correlation between the item and the latent ability measured by the test.

Urry examined the graphical approximations using data from 4,950 examinee responses to 98 unscreened mathematics items from the Washington Pre-College Test Battery, a highly reliable test ($K-R 20 = 0.93$). Correlations between the a - and b -parameters derived from the graphs and their maximum likelihood (ML) estimates were 0.89 and 0.97, respectively. Urry concluded that the correlation coefficients indicated a strong degree of accord between the heuristic approximations and the ML estimates.

In a theoretical paper Schmidt (1977) proposed that the graphical procedure proposed by Urry tends to systematically underestimate a_i and overestimate $|b_i|$ and the variance of b_i because the point-biserial correlation between the item score and the estimated latent trait (i.e., total test score), $r_{i\hat{\theta}}$, is taken as an estimate of the point-biserial correlation between the binary item and the perfectly reliable latent trait, $\hat{\rho}_{i\theta}$. Values of $r_{i\hat{\theta}}$ are attenuated because of guessing on item i , and the unreliability of $\hat{\theta}$. Schmidt argued that Urry's four criteria for the total test score to be an estimate of latent trait score, $\hat{\theta}$, would minimize rather than eliminate the effect. Schmidt pointed out that that increased values of the biserial correlation imply larger \hat{a}_i and smaller $|\hat{b}_i|$.

The heuristic method has also been used with simulation data. Jensema (1976) simulated data and compared the parameter estimates set during the data generation phase to the estimates derived from the heuristic method and ML estimation. Forty-eight data sets were created with a total of 2,800 items and 44,000 simulated examinees. True abilities of examinees were normally distributed. The simulation design consisted of: sample sizes of 250, 500, 750, and 1000; test lengths of 25, 50, and 100; a -parameters of 0.5, 1.0, 1.5, and 2.0, consistent within a dataset; b -parameters between -2.4 and 2.4 at intervals of 0.2; and c -parameters of 0.2. Parameter values derived from the heuristic method were used as starting values for the ML procedure.

The overall correlations between the true and heuristic estimates were 0.80 and 0.96 for the a - and b -values respectively, while the overall correlations between the true and ML estimates were 0.86 and 0.97 for the a - and b -values, respectively. Jensema concluded that the heuristic estimates were "surprisingly accurate" (p. 713). The correlations revealed that the true parameters and the corresponding estimates derived

from the heuristic method increased with greater sample size and a greater number of test items, as initially suggested by Urry. Jensema concluded that the heuristic method may be used as a convenient technique for examining the worth of an item pool for tailored testing.

Ree (1979) also conducted a simulation study to assess the effectiveness of the heuristic method. The a - and b -values derived from the heuristic method and the a - and b -values derived from three common computer programs (i.e., ANCILLES, LOGIST, OGIVIA) were correlated with true item parameters. Using the 3PL model, data were generated for an 80-item test for normally distributed, positively skewed, and uniformly distributed groups of 2,000 examinees. The true item parameters represented real examination data and were normally distributed ($M_a = 0.95$, $SD_a = 0.28$; $M_b = 0.16$, $SD_b = 0.93$; $M_c = 0.20$, $SD_c = 0.05$).

The correlations between estimated and true parameters revealed that the b -values were more closely aligned to the true parameters than the a -values. The heuristic values and the values obtained from the computer programs yielded correlations equal to or higher than 0.90 for the b -parameters. Correlations of a -parameters and the values obtained from the heuristic method were 0.32, 0.35, and 0.59 for the skewed, normal, and uniform ability distributions, respectively. Correlations of a -parameters and the values obtained from the computer programs also were variable across ability distributions. The lowest correlations, ranging from 0.44 to 0.57, were observed for the skewed data, whereas high correlations were found for the normal distribution (range of 0.83 to 0.84), and the uniform distribution (range of 0.87 to 0.90).

Contributions and Shortcomings of the Studies

Even though Lord's formulas were used in the context of the three-parameter model, the studies suggest that the transformation procedures from the classical item indices to the corresponding IRT item indices may provide some promise as a heuristic technique under certain conditions. General findings indicated that the b -parameters derived from the heuristic method were highly correlated with true or ML estimates of b -values regardless of the shape of the ability distribution, whereas the correlations for the a -parameters were moderately to highly correlated.

Correlations were presented as evidence of the accuracy of the heuristic method to reproduce the item parameters in the studies reviewed above. However, high correlations only indicate that sets of values are strongly linearly related; they provide no evidence of parameter recovery. Ree (1981) noted that a correlation between a set of parameters and their estimates would be misleading if systematic bias was present, such as consistent over- or under-estimation.

In 1980, Lord clarified the circumstances for which the formulas were relevant by stating that they are "valid only for the case where θ is normally distributed and there is no guessing" (p. 33). The accuracy of the formulas under these two conditions has not been determined. Instead, attention has been given to the comparability of CTST and IRT item indices determined by analyzing the same data set with both models and using correlational techniques to determine the degree of association (e.g., Fan, 1998; MacDonald & Paunonen, 2002; Stage, 1998a, b, 1999) Although not directly related to the purposes of the present study, the findings of this research reveals that the difficulty and discrimination indices of the two models are linearly related.

Purpose

In 1974, Urry pointed out the efficacy of using his item parameter approximations, based on the Lord and Novick formulas, was an open empirical question. To date the question has not been answered. The purposes of the present study are to determine (1) the veracity of the two formulas within the context that Lord (1980) proposed, and (2) the robustness of the formulas beyond the initial and restrictive conditions identified by Lord. The conditions that ability (θ) is normally distributed with mean zero and unit variance, and no guessing imply that the two-parameter IRT model be used. An additional condition, specified for the difficulty parameter, is that all items are equally discriminating.

The formulas will be investigated as a function of shape of ability distribution, test length, and item discrimination. The findings of the present study will be used to develop guidelines as to if and when the formulas proposed by Lord accurately reproduce IRT item indices.

Method

Lord's formulas were investigated using simulated data and actual achievement data. The simulated data was used to examine the behaviours of the formulas under different experimental conditions, where the population parameters are known. The achievement data was used to examine the extent to which the simulation results are generalizable to real data.

Simulation Study

Research Design

The research design was a 3 (ability distribution) x 3 (test length) x 2 (item discrimination) fully crossed design, comprising 18 cells. The conditions were designed to represent realistic response data.

Ability distribution. Given that ability is likely not normally distributed for most groups of examinees (Lord, 1980), two skewed distributions were modeled as well the normal distribution. The skewed ability distributions were generated using the beta probability density function. The positively skewed distribution, defined as beta (2.9, 5.7), achieved an expected skewness of 0.40 and an expected kurtosis of -0.30; the negatively skewed distribution, defined as beta (5.7, 2.9), achieved an expected skewness of -0.40 and an expected kurtosis of -0.30. The beta distributions were rescaled so that the mean and standard deviation of the distribution of θ 's were 0 and 1, respectively.

Test length. Three test lengths were employed: a short exam of 20 items, a moderate exam of 40 items, and a long exam of 80 items. The short and moderate exam lengths are consistent with that frequently found in psychological and educational applications (Seong, 1990; Yen, 1987). The longest exam is consistent with Urry's requirement for the item-test biserial correlation to be a close approximation to the item-latent trait correlation.

Item discrimination. Two conditions of item discrimination were investigated. One condition maintained a constant value of 1 for a -parameters, which adheres to Lord's stipulation that there is a monotonic relation between b_i and π_i when items are equally discriminating. Traub (1983) commented on the appropriateness of the assumption that

all item discrimination parameters are equal. Considering the abundance of empirical evidence, he stated: “The fact that otherwise acceptable achievement items differ in the degree to which they correlate with the underlying trait has been observed so very often that we should expect this kind of variation for any set of achievement items we choose to study” (p. 64). Therefore, variable discrimination values were modeled. A log normal distribution was chosen, given it is the default distribution for slopes in BILOG (Mislevy & Bock, 1990) and has been selected by other researchers modeling achievement data (e.g., D. L. Henderson-Montero, personal communication, May 2, 2003; Seong, 1990).

Item difficulty. Hambleton and Swaminathan (1985) noted that when a test is designed for norm-referenced interpretations, items are generally chosen to have a narrow range and medium level of difficulty, such as b -values between -2 to 2. This strategy will differentiate examinees on their competence in a given subject area, producing a broad range of scores that maximize discrimination among examinees. However, to observe the effect of more extreme b -parameters on the formulas, b -values were selected from a normal distribution ($\mu = 0, \sigma = 1$) in the present study.

Data Generation

The item response generation technique described by Harwell et al. (1996) was used to create the data. The procedure entailed four steps. Step 1 involved the generation of true ability scores (θ). In Step 2, tests were created according to the test specifications (test length, a -parameters, and b -parameters). For each cell in the experimental design, a unique test was created by deriving new sets of a - and b -parameters consistent with the specifications for that cell. In Step 3, response probabilities of a population of 10,000 examinees to the n (20, 40, 80) items based on the 2PL IRT model were determined,

producing a $10,000 \times n$ matrix. In Step 4, the matrix of response probabilities was translated into a $10,000 \times n$ data matrix of 0/1 responses. Each response probability was compared to a random number drawn from a uniform distribution of values in the closed interval $[0,1]$. If the response probability was equal to or greater than the random number, a '1' was assigned for that item, whereas if the response probability was less than the random number, a '0' was assigned for that item.

Replications

The benefits of replicated over non-replicated IRT simulation studies are the same as that observed in empirical studies; aggregating results over replications produces more stable and reliable findings. The number of replications influences the precision of the estimated parameters. Therefore, increasing the number of replications is an attractive technique for reducing the variance of estimated parameters (Harwell et al., 1996). In the present study, 100 random samples of 1,000 examinees were drawn from the population for each experimental condition¹.

Computer Programs

Mathematica for Students (Version 4, Wolfram, 2000) was used to generate the response data matrices. *LERTAP* (Version 5, Nelson, 2000), an *Excel* application, was employed to obtain the classical item analyses and Lord's estimation of the a - and b -parameters. Random sampling from the 10,000 examinees was performed with an *Excel* (Version 5) macro program.

¹ Two other sample sizes – 500 and 250 examinees – were examined as well. Only the results for the sample size of 1,000 examinees are reported in the present paper.

Diploma Examination Study

Lord's formulas were applied to actual achievement data sets. These data sets consisted of the item scores obtained on provincial examinations by students who wrote the biology exam ($N = 9,030$), representing the sciences, and the English exam ($N = 13,375$), representing the humanities (Alberta Learning, 1999a, b). The exams are high school graduation examinations, which contribute 50% towards students' final course grades. Only the multiple-choice components of the exams were used, which comprised 48 items for the biology exam and 70 items for the English exam.

The assumptions underlying the two-parameter IRT model were assessed. The shape of the scree plot yielded by a principal component analysis showed a dominant first component and the difference between the second and third components, third and fourth, and so forth were small in comparison to the difference between the first and second components (Hambleton, Swaminathan, & Rogers, 1991), suggesting each examination was essentially unidimensional (Nandakumar, 1994). Speed was not a factor since students are given up to 30 minutes of additional time when required.

Descriptive statistics for the total test and ability distributions, and item information are presented in Table 1. The classical item information reveals that the items were of moderate difficulty, as judged by the mean p -values. The observed score distributions for both exams were negatively skewed and platykurtic. Item requirements for the exams include minimum and maximum acceptable difficulty levels 0.30 and 0.85, respectively, and a minimum acceptable point-biserial correlation of 0.20 (Alberta Education, 1999), which approximately corresponds to a biserial correlation of between

0.25 and 0.30. The mean biserial correlations met the criterion of 0.40 and above to be considered high (Nelson, 2001).

Examination of the IRT information reveals that the ability distributions for both exams were positively skewed and leptokurtic. The IRT item information indicates that the mean difficulty and discrimination parameters for the achievement data are lower than that modeled in the simulation study.

Table 1.

Psychometric Properties of the Biology and English Exams using CTST and IRT Analyses

	CTST		IRT	
	Biology	English	Biology	English
Test Level Information				
Mean	33.48 (3 to 48)	44.26 (12 to 69)	0.08 (-4.00 to 4.02)	0.03 (-3.15 to 4.00)
Median	34	44	-0.10	-0.08
SD	7.83	11.08	1.21	1.09
Reliability (α)	0.86	0.89		
Skewness	-0.37	-0.13	0.78	0.66
Kurtosis	-0.52	-0.69	0.53	0.35
Item Level Information				
<i>M</i> Difficulty	0.70 (0.39 - 0.88)	0.63 (0.35 to 0.86)	-1.12 (-2.44 to 0.90)	-0.87 (-2.82 to 0.86)
<i>M</i> Discrimination ^a	0.50 (0.30 to 0.73)	0.46 (0.21 to 0.67)	0.56 (0.25 to 1.09)	0.47 (0.14 to 0.89)

^a r_b is presented for the CTST information.

Statistical Analyses

The item parameters derived from Lord's formulas were compared to the parameters used to generate the data matrices for the simulated data. Since the true item parameters for the achievement exams were not known, the item parameters derived from BILOG using the two-parameter model were used to evaluate the formulas.

Dependent Variables.

Bias. Parameter recovery is generally assessed by comparing the difference between an item parameter estimate and the corresponding parameter value (Harwell et al., 1996). Estimation bias is defined as the mean difference between the estimated and true parameter value for an item across 100 replications. Bias in each a_i is calculated by:

$$\text{Bias } a_i = \frac{\sum_{r=1}^{100} (\hat{a}_{ir} - a_i)}{100}.$$

Bias in b_i is calculated similarly. Smaller differences indicate the estimates are closely aligned to the parameter values compared to larger differences. Maintaining the valence of the difference enabled determination of whether the estimates systematically overestimated (positive bias) or underestimated (negative bias) the parameter value. Examining the nature of the bias is particularly important in light of Schmidt's (1977) contention that Lord's formulas tend to systematically underestimate a_i and overestimate $|b_i|$.

Standard Errors. Empirical standard errors were calculated to determine how variable the estimates were over replications. Gifford and Swaminathan (1990) presented the following formula for variance of the estimates across replications:

$$\hat{\sigma}_{a_i}^2 = \frac{\sum_{r=1}^{100} (\hat{a}_{ir} - \bar{\hat{a}}_i)^2}{100},$$

where $\hat{\sigma}_{a_i}^2$ is the variance error the estimated item discrimination for item i and $\bar{\hat{\alpha}}_i$ is the mean of the estimated a -parameters for item i across 100 replications. The variance error for b_i was calculated similarly. Smaller values of sampling variance suggest that the estimates are fairly stable and reliable, whereas larger values indicate the estimates may be unreliable. Standard errors were determined by taking the square root of the mean of the variance error for each condition. The standard error was used to construct 95% confidence intervals around the bias of zero.

Results

The results are presented in graphical form, rather than by summary data, to illustrate the patterns of bias. The scatter plots display estimation biases of IRT item parameters using Lord's formulas across the spectrum of parameter values. The simulated data results are presented first, followed by the achievement data results.

Simulation Study

The simulation study examined the behaviours of the formulas under conditions congruent and incongruent with that prescribed by Lord. Although length of test was an independent variable, bias patterns were consistent across the 20, 40, and 80 item tests. The 80 item tests were used in the present paper to illustrate these results. The scales of the ordinate and abscissa were kept constant for the discrimination figures and the difficulty figures to facilitate comparisons across conditions. Bias, documented on the y -axis, is presented in increments of two standard errors, which corresponds approximately to the 95% confidence interval.

Item Discrimination

Lord's formula for discrimination is solely a function of the biserial correlation. Because discrimination values typically range between 0 and 2 (Hambleton, Swaminathan, & Rogers, 1991), mean standard errors were computed from the items within that range. The mean standard errors for the normal, positively skewed, and negatively skewed ability distributions were 0.1064, 0.0966, and 0.1001, respectively, yielding an overall mean of 0.1010. Rounding to two decimal places, the standard error of 0.10 was used.

Figure 1 shows estimation bias as a function of the true a -parameter when ability is normally distributed. The results indicate that when true a -values were less than or equal to 1.50, biases were within 2 standard errors of the true value for all but three items. Two of these items were close to 1.50 ($a_i = 1.46$, bias = 0.23; $a_i = 1.48$, bias = 0.27), whereas the other was close to 1.00 ($a_i = 0.99$, bias = -0.23). When a_i exceeded 1.50, more extreme biases were observed for all 11 items. Items that tended to yield high bias values also tended to have high biserial correlations. For example, the most discriminating item ($a_i = 3.08$) had a bias of -1.30 and a biserial correlation of 0.92.

The data for the positively and negatively skewed distributions are presented in Figures 2 and 3, respectively. Similar patterns were observed. For the positively skewed distribution, 4 of the 57 items (7%) with true a -values less than 1.20 had biases greater than two standard errors of the true value ($a_i = 0.99$, bias = -0.34; $a_i = 1.05$, bias = 0.24; $a_i = 1.11$, bias = 0.22; $a_i = 1.19$, bias = -0.52). When a -parameters were equal to or greater than 1.20, 13 of the 23 items (57%) had biases greater than two standard errors. The bias of greatest magnitude, -0.96, was observed for an a -value of 3.08 ($r_b = 0.90$). For the negatively skewed distribution, 3 of the 57 items (5%) with true a -values less than 1.20

had biases greater than two standard errors of the true value ($a_i = 0.70$, bias = -0.21; $a_i = 1.12$, bias = 0.26; $a_i = 1.18$, bias = 0.25). When a -parameters were equal to or higher than 1.20, 14 of the 23 items (61%) had biases that exceeded two standard errors. The bias of greatest magnitude, 0.80, occurred for an a -value of 2.04 ($r_b = 0.94$).

Item Difficulty

The patterns of bias for the difficulty parameter are described and illustrated as a function of true b -parameter. In all cases, curvilinear relationships were found. The mean standard errors for the normal, positive and negatively skewed distributions for the condition of unit discrimination were 0.0828, 0.0957, and 0.1070 and for the condition of variable discrimination were 0.0827, 0.0963, and 0.1566, respectively. The mean standard error across these six conditions was 0.1035. Rounding to two decimal places, the standard error of 0.10 was used.

Unit discrimination. Lord (1980) stipulated that the discrimination be constant when using the formula for difficulty. Figure 4 provides a scatter plot of the item difficulty biases for the conditions of normal ability distribution and unit discrimination. The estimated values were within 0.20, or two standard errors, of true values when b -parameters were less than 2.00. The most difficult item ($b_i = 2.44$, p -value = 0.05) yielded the most extreme bias, 0.47. For this set of b -parameters, no true values were less than -2.00; therefore, the performance of the difficulty formula could not be evaluated for easy items (i.e., $b_i < -2.00$) when ability is normal and discrimination is constant.

Figure 5 displays the biases of the estimated values for the positively skewed ability distribution. Estimated values were within two standard errors of true values when the b -parameters were greater than -1.50. The best fitting curve tended to arc downwards

for items with b -parameters less than -1.50 . The four items that had true difficulty values of less than -1.50 were underestimated by more than two standard errors. The easiest item ($b_i = -2.71, p\text{-value}=0.97$) was underestimated by 1.10.

While extreme biases were observed for easy items in the positively skewed distribution, extreme biases were observed for difficult items in the negatively skewed distribution. Figure 6 shows that when b -parameters were less than or equal to 1.50, estimated difficulty was within two standard errors of true values. The best fitting curve tended to arc upwards for items with b -parameters greater than 1.50. All eight items with true difficulty greater than 1.50 were overestimated by more than two standard errors. The b -values for the two most difficult items ($b_i = 2.32, p\text{-value} = 0.05$; $b_i = 2.54, p\text{-value}=0.04$) were overestimated by 1.05 and 1.10, respectively.

Variable discrimination. When discrimination was variable, bias patterns were consistent with that observed for unit discrimination. Figure 7 displays estimation bias for true difficulty when ability is normally distributed. Estimated difficulty for all items was within two standard errors of true values except the easiest item ($b_i = -1.92, p\text{-value} = 0.95, \text{bias} = -0.22$) and the hardest item ($b_i = 2.17, p\text{-value} = 0.07, \text{bias} = -0.31$).

Figure 8 illustrates the pattern of bias for the positively skewed distribution. Like the condition with the discrimination held constant, the estimated values were within two standard errors of true values when the b -parameters were -1.50 or greater. The six items that had b -values less than -1.50 were all negatively biased. The most extreme bias, -0.95 , was observed for the easiest item ($b_i = -2.27, p\text{-value} = 0.96$).

Figure 9 illustrates the pattern of bias for the negatively skewed distribution. All estimated b -values were within two standard errors of their true values when b_i was less

than or equal to 1.50. There were 7 items with b_i values higher than 1.50, of which 4 had biases that exceeded two standard errors. Only two of these items are shown on the figure. The other two items are not shown in Figure 9 due their large biases; the b -values and bias of these two items were, respectively, 2.88 (p -value = 0.01) and 1.86, and 2.73 (p -value = 0.01) and 1.61².

Achievement Data

Bias in the achievement data was calculated as the difference between Lord's estimate and the BILOG parameter estimate. The ability of Lord's formulas to recover a - and b -parameters is considered with respect to both the biology and English exams. Like the simulated data, the scales for the ordinate and abscissa were kept constant for the difficulty figures and the discrimination figures to facilitate comparisons across the two subject areas, and are different than that used to present the simulation data.

To calculate standard errors for the item parameters of the achievement data, 100 random samples of 1,000 examinees were performed. The standard errors for the a -parameters were 0.0577 for the biology exam and 0.0505 for the English exam. The standard errors for the b -parameters were 0.1461 for the biology exam and 0.1391 for the English exam. Standard errors of 0.05 and 0.14 were selected for the discrimination and difficulty parameters, respectively.

Item Discrimination

Figure 10 provides a scatter plot of the discrimination biases for the biology exam. Estimated a -values were within 0.10, or two standard errors, of BILOG values. Biases ranged between -0.035 and 0.08 . A similar trend was observed for the English

² Accommodating such extreme values would alter the scale in such a way that the patterns of the remaining items would be obscured.

exam (see Figure 11). All estimated a -values were within two standard errors of BILOG values, and were positive, ranging between 0.016 and 0.080. Both figures demonstrated a trend such that bias decreased as discrimination increased.

Item Difficulty

Figures 12 and 13 illustrate estimation bias of the difficulty parameter for the biology and English exams, respectively. All item biases were within two standard errors, or 0.28, of BILOG difficulty values for the biology exam, but not for the English exam. In the case of the English exam, all items with b -values greater than -1.80 were within two standard errors, while 5 of the 10 items with b -values less than -1.80 level were overestimated by more than two standard errors. The patterns of biases suggest that there are linear relationships between bias and item difficulty for both exams.

Discussion

The purposes of the present study were two-fold. The veracity of the discrimination and difficulty formulas were assessed within the context of which Lord (1980) proposed. Data were simulated to replicate the ideal conditions of normally distributed ability (θ) with the two-parameter IRT model where guessing is not a factor in student responses, and equal discrimination across items, specified for the difficulty parameter. Given these conditions are unlikely to be met, the robustness of the formulas was assessed using simulation and actual exam data.

Item Discrimination

The behaviour of Lord's formula for item discrimination was examined using variable discrimination parameters for the normal and two skewed ability distributions (see Figures 1 through 3) in the simulation study. How well the formula worked was

shown to be a function of true discrimination. Estimation bias was minimal, defined as within two standard errors of the true values, for a -values less than or equal to 1.50 for the normal ability distribution and less than 1.20 for the skewed ability distributions. Biases of greater magnitude, both positive and negative, were observed for discrimination values outside of these limits. Hence, it would appear that Lord's formula for item discrimination works well for the normal distribution and $\hat{a}_i \leq 1.50$, and for skewed distributions and $\hat{a}_i \leq 1.20$.

Estimation using the achievement data suggested that the formulas performed very well. None of the values was misrepresented by as much as 0.10, a predictable outcome considering the simulation results because a -values for both exams ranged from approximately 0.20 to 1.00. Without higher discriminating items, the formula performed impressively well. This finding is especially noteworthy since the ability distributions were leptokurtic, a condition not modeled in the simulation study, and more positively skewed than that modeled in the simulation study. However, the results from the simulation study indicate that the formula does not perform well for more discriminating items. Whether this finding will hold for actual achievement test items with higher discriminations than those observed in the two examinations needs to be verified.

Item Difficulty

Lord prescribed the conditions of normal ability distribution and constant item discrimination for the difficulty formula. Although large bias was observed for the most difficult item, the bias was within two standard errors of true values for the range of b 's typically found in achievement tests (i.e., -2 to 2; Hambleton & Swaminathan, 1985).

The skewed distributions with unit discrimination parameters (Figures 5 and 6) were recovered quite well; however, bias was differentially affected by the direction of the skewness. Bias was most pronounced and negative for the easy items in the positively skewed distribution and most pronounced and positive for the difficult items in the negatively skewed distribution. These results may seem counterintuitive since the tail of the distributions results in lower bias than the non-tailed region. The answer is found in the nature of the ability distributions. There are few ability values in the non-tail. In the present study, there were few ability values greater than 2.20 for the negatively skewed population and few true ability values less than -2.20 for the positively skewed population, resulting in a ceiling effect and a floor effect, respectively. The effect on b -parameter estimation is dramatic. The numerator of the formula is a z -score, which changes more rapidly as p -values reach very high and very low levels, driving up the absolute value of the z -score. As a consequence, b_i is overestimated when most examinees answer incorrectly and b_i is underestimated when most examinees answer correctly. Although high and low p -values (i.e., $0.10 < p < 0.90$) are not desirable item characteristics, the findings highlight the limitation of the formula to accurately predict IRT difficulty in such circumstances.

The patterns of bias in the b -estimates observed for the conditions of variable discrimination were comparable to their counterparts for unit discrimination. Seemingly Lord's restriction that the discrimination be held constant is not required.

Estimation using the achievement data suggested that the formulas again performed quite well. The b -parameters were estimated within two standard errors of BILOG values for all items on the biology exam and all but five of the easier items on the

English exam. The bias for the easy items was positive, rather than negative, as found in the simulation for the positively skewed distributions. The achievement tests were somewhat easy, with the majority of items assuming negative BILOG b -values. In contrast, the simulated tests were of moderate difficulty with approximately the same number of items above and below zero. Lastly, while the distributions of ability estimates, $\hat{\theta}_i$, were platykurtic for the simulated skewed data, the distribution of the ability estimates were leptokurtic for the achievement data. Clearly, more work is needed to understand the influence of kurtosis on the estimates of b_i produced by the formula.

Hambleton (1989) commented on problems associated with the heuristic method. He noted that the relationships between model parameters and conventional item statistics are based on highly restrictive assumptions, and the usefulness of the parameter estimates is reduced if the assumptions made are not met. Results from the present study suggest that violations of the prescribed assumptions appear not to have a detrimental effect on parameter estimation. Hambleton also pointed out that parameter estimates produced by the heuristic method do not have known sampling distributions, and therefore the standard errors associated with the estimates are unknown. Although standard errors were empirically derived for the present study, more work needs to be done to derive theoretical standard errors.

To conclude, the present study explored Lord's formulas by comparing the difference between parameter estimates and corresponding parameter values. Previously, correlational data had been used to assess the formulas. Schmidt (1977) contended that a_i would be systematically underestimated and $|b_i|$ would be systematically overestimated using the three-parameter IRT model. Within the context of the two-parameter model, the

results suggest that this is not the case. Rather, parameter recovery was a function of true item values. Taken together, and notwithstanding the call for more work, Lord's formulas produced very satisfactory results for items with item difficulties ($-2 < b_i < 2$) and discriminations ($a_i < 1.2$) generally found in achievement tests. To echo the sentiments of Jensema, the heuristic estimates are "surprisingly accurate."

References

- Alberta Education (1999). Alberta Education Annual Report 1998-1999. Edmonton, AB: Author.
- Alberta Learning (1999a). *Biology 30 Grade 12 Diploma Examination, June 1999*. Edmonton, AB: Author.
- Alberta Learning (1999b). *English 30 Grade 12 Diploma Examination, June 1999*. Edmonton, AB: Author.
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58, 357-381.
- Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement*, 14, 33-43.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 147-200). New York: McMillan.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff.
- Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.
- Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in Item Response Theory. *Applied Psychological Measurement*, 20, 101-125.

Jensema, C. (1976). A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement*, 36, 705-715.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New Jersey: Lawrence Erlbaum.

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

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62, 921-943.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd Edition). Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Bock, R. D. (2000). *BILOG 3.2: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software, Inc.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses – Comparison of different approaches. *Journal of Educational Measurement*, 31, 17-35.

Nelson, L. R. (2000). *Laboratory of Education Research Test Analysis Package (LERTAP, Version 5)* [Computer program]. Curtin University of Technology: Perth, Western Australia.

Nelson, L. R. (2001). *Item Analysis for Tests and Surveys Using Lertap 5*. Curtin University of Technology: Perth, Western Australia.

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement*, 3, 371-385.

Ree, M. J. (1981). The effects of item calibration sample size and item pool size on adaptive testing. *Applied Psychological Measurement*, 5, 11-19.

Schmidt, F. L. (1977). The Urry method of approximating the item parameters of latent trait theory. *Educational and Psychological Measurement*, 37, 613-620.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, 14, 299-311.

Stage, C. (1998a). *A comparison between item analysis based on Item Response Theory and Classical Test Theory. A study of the SweSAT subtest ERC*. (Educational Measurement No. 30). Umea, Sweden: University of Umea, Department of Educational Measurement.

Stage, C. (1998b). *A comparison between item analysis based on Item Response Theory and on Classical Test Theory. A study of the SweSAT subtest WORD*. (Educational Measurement No. 29). Umea, Sweden: University of Umea, Department of Educational Measurement.

Stage, C. (1999). *A comparison between item analysis based on Item Response Theory and Classical Test Theory. A study of the SweSAT subtest READ*. (Educational Measurement No. 33). Umea, Sweden: University of Umea, Department of Educational Measurement.

Stone, C. A. (1992). Recovery of Marginal Maximum Likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, 16, 1-16.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, 34, 253-269.

Wolfram, S. (2000). *Mathematica for Students* (Version 4) [Computer program]. Wolfram Research.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, 52, 275-291 .

Figure 1. Bias in Discrimination Estimation as a Function of True Discrimination for the Normal Ability Distribution

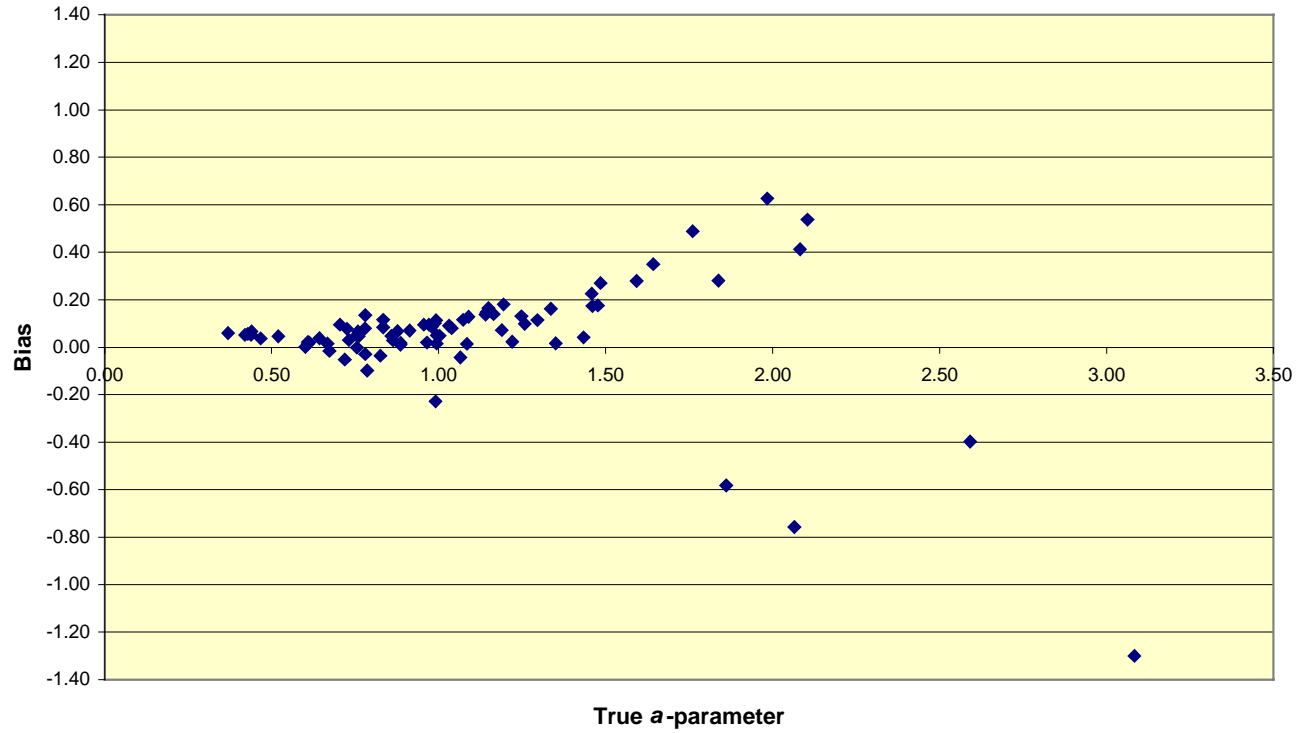


Figure 2. Bias in Discrimination Estimation as a Function of True Discrimination for the Positively Skewed Ability Distribution

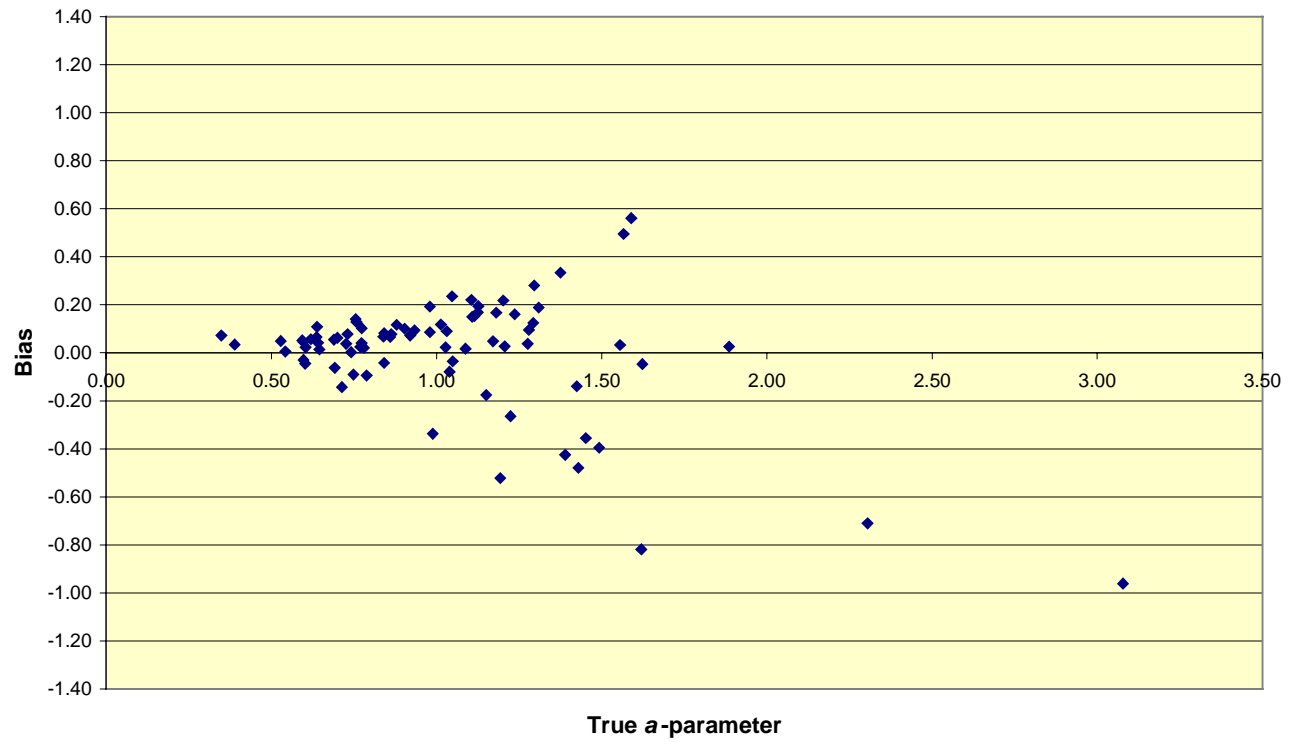


Figure 3. Bias in Discrimination Estimation as a Function of True Discrimination for the Negatively Skewed Ability Distribution

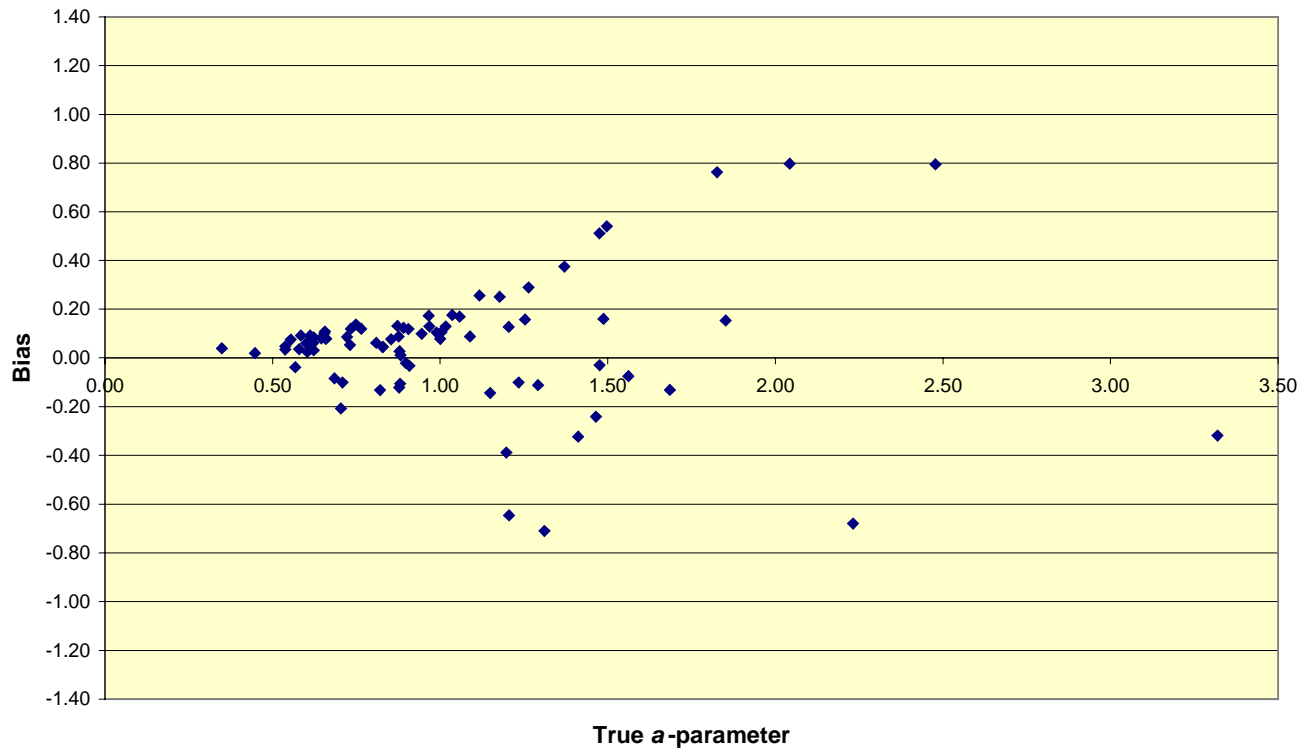


Figure 4. Bias in Difficulty Estimation as a Function of True Difficulty for the Normal Ability Distribution and Unit Discrimination

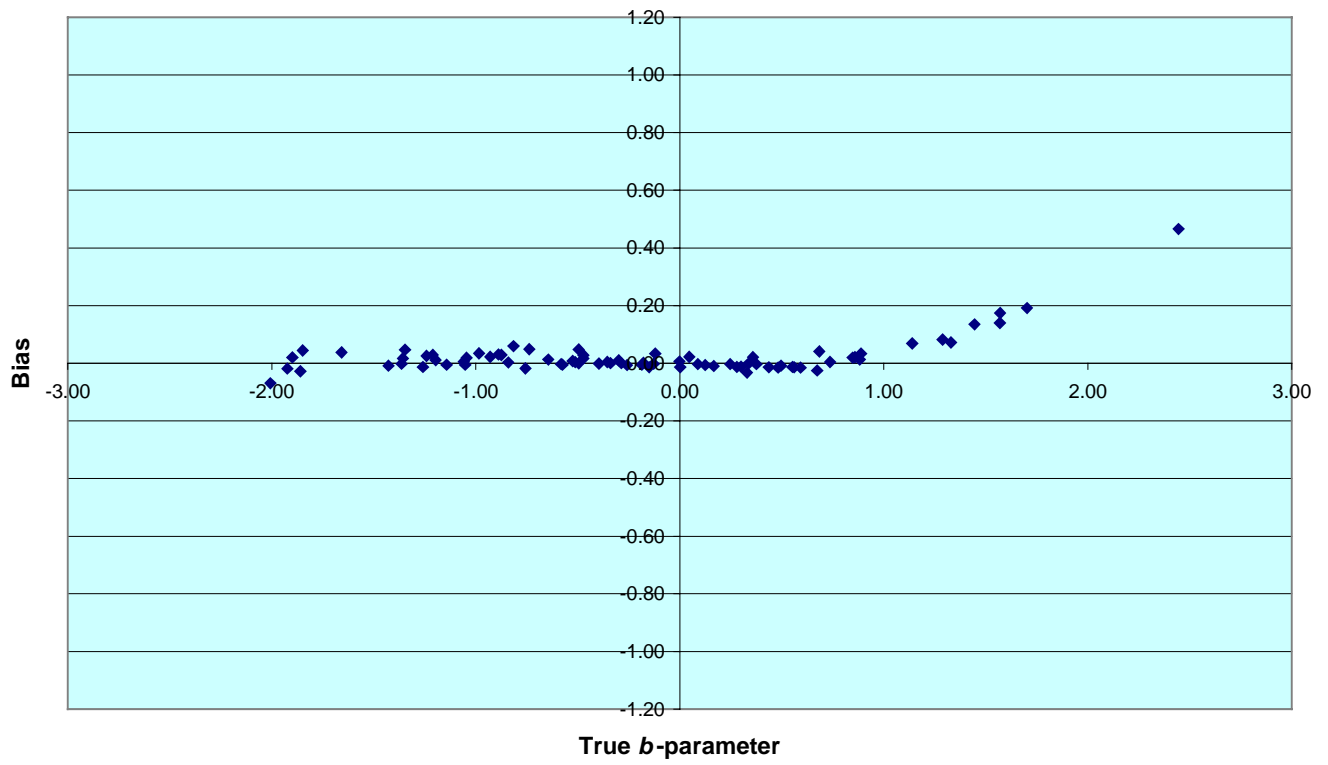


Figure 5. Bias in Difficulty Estimation as a Function of True Difficulty for the Positively Skewed Ability Distribution and Unit Discrimination

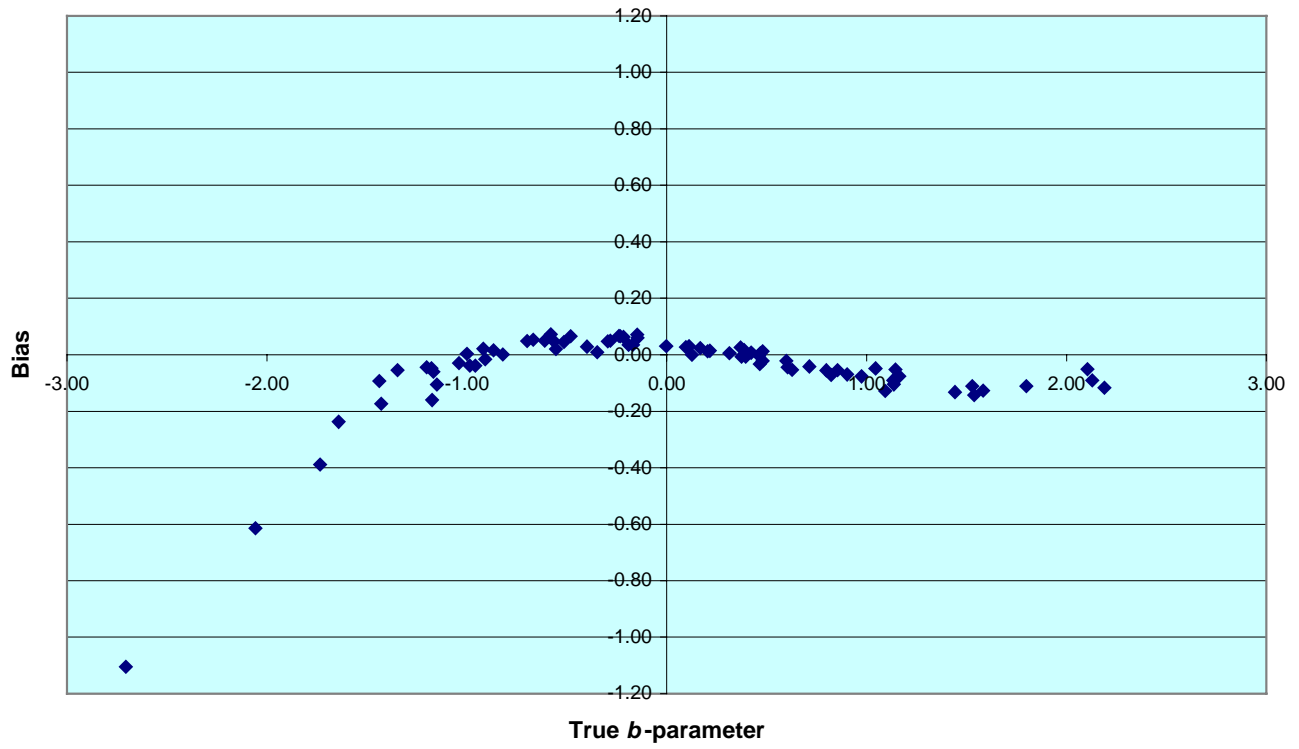


Figure 6. Bias in Difficulty Estimation as a Function of True Difficulty for the Negatively Skewed Ability Distribution and Unit Discrimination

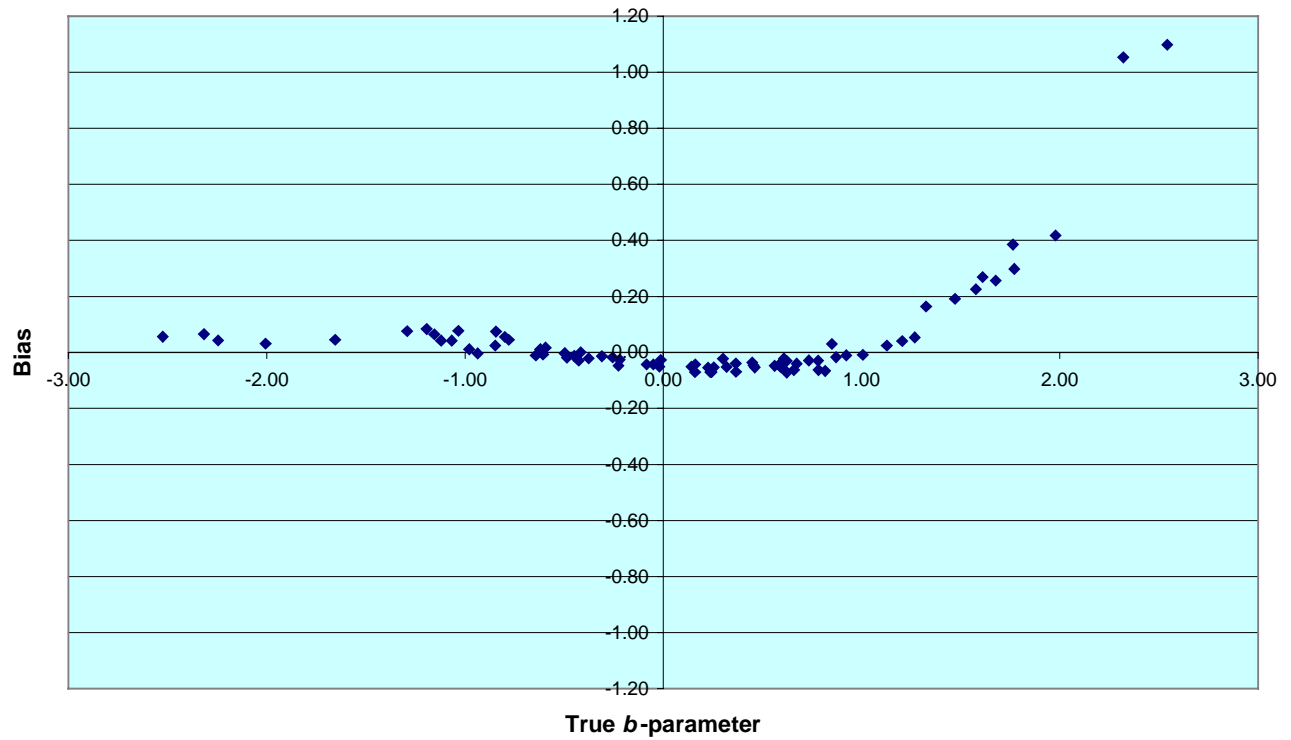


Figure 7. Bias in Difficulty Estimation as a Function of True Difficulty for the Normal Ability Distribution and Variable Discrimination

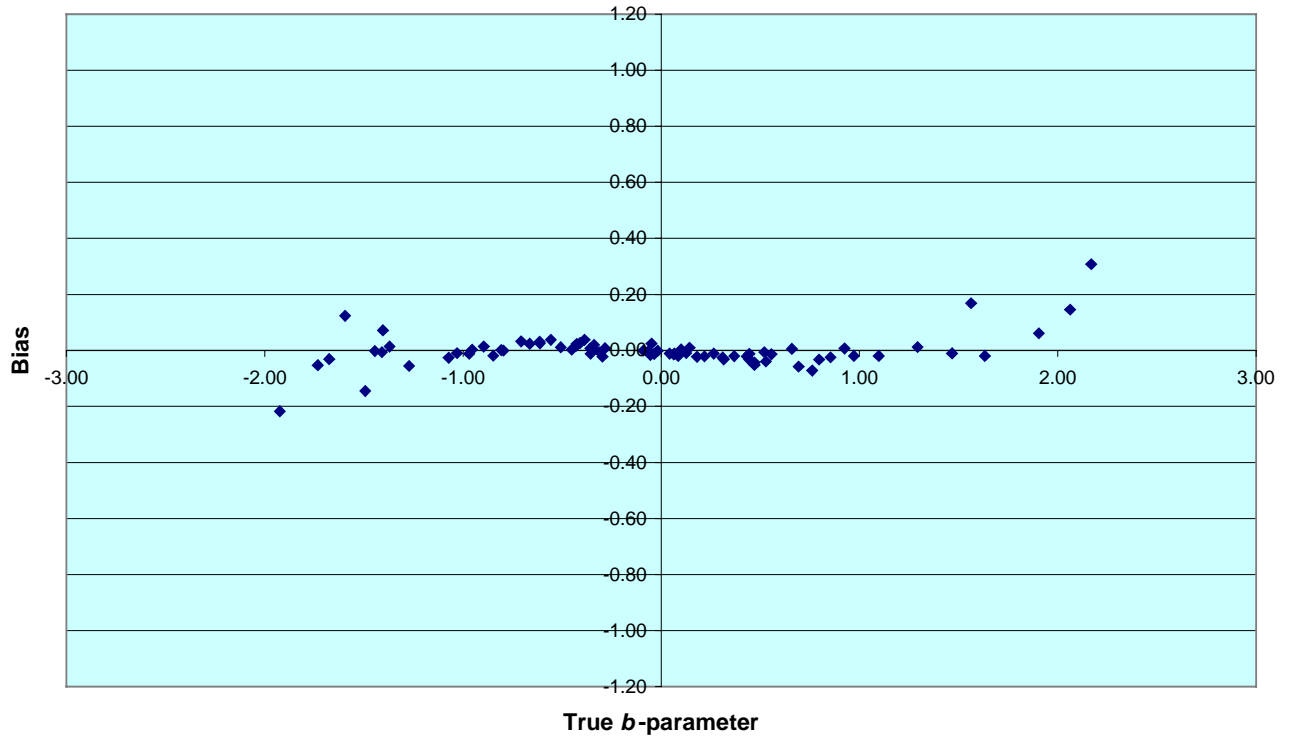


Figure 8. Bias in Difficulty Estimation as a Function of True Difficulty for the Positively Skewed Ability Distribution and Variable Discrimination

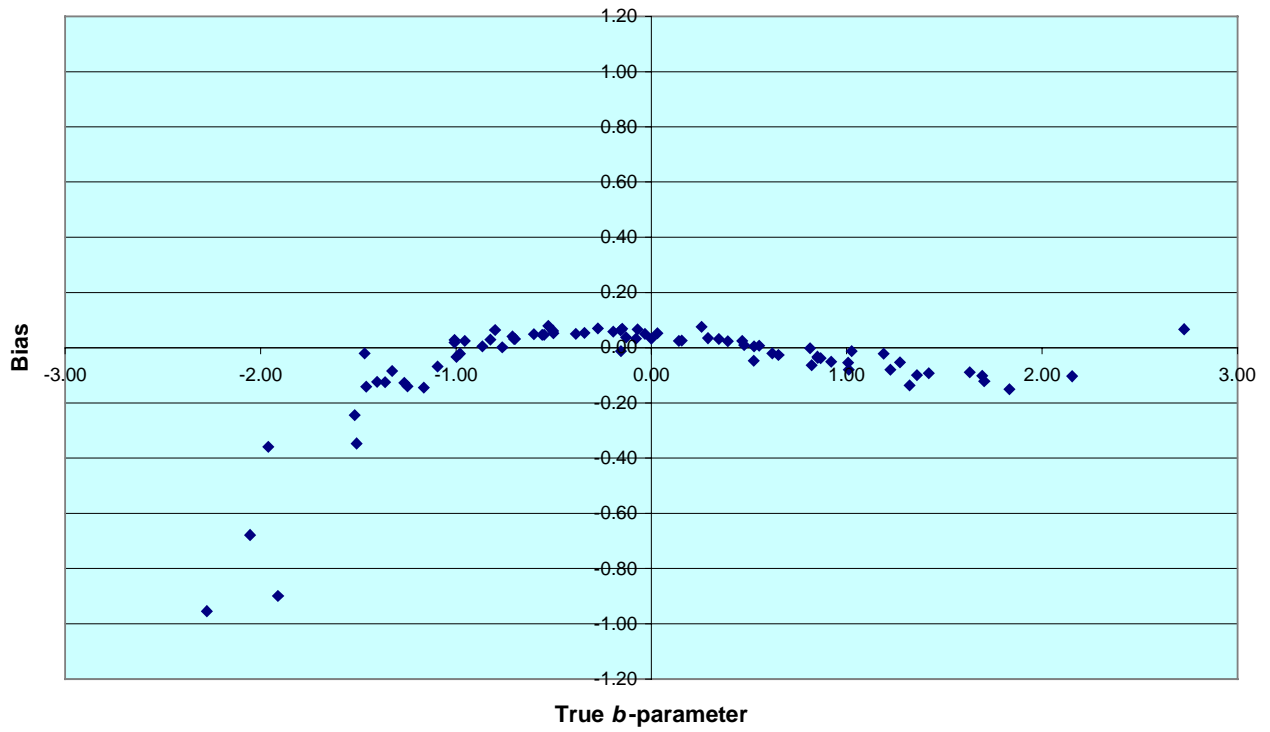
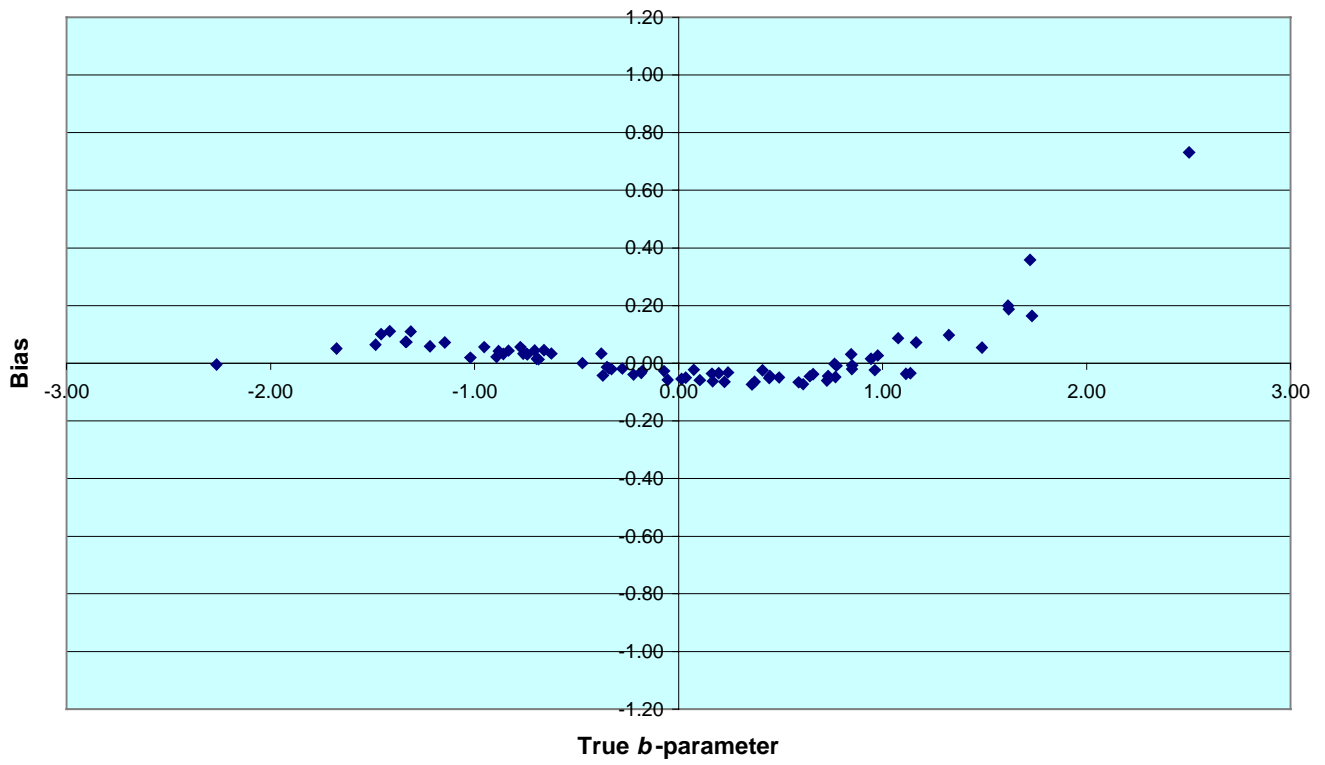


Figure 9. Bias in Difficulty Estimation as a Function of True Difficulty for the Negatively Skewed Ability Distribution and Variable Discrimination



**Figure 10. Bias of the IRT Discrimination Parameter as a Function of BILOG
Discrimination for the Biology Exam**

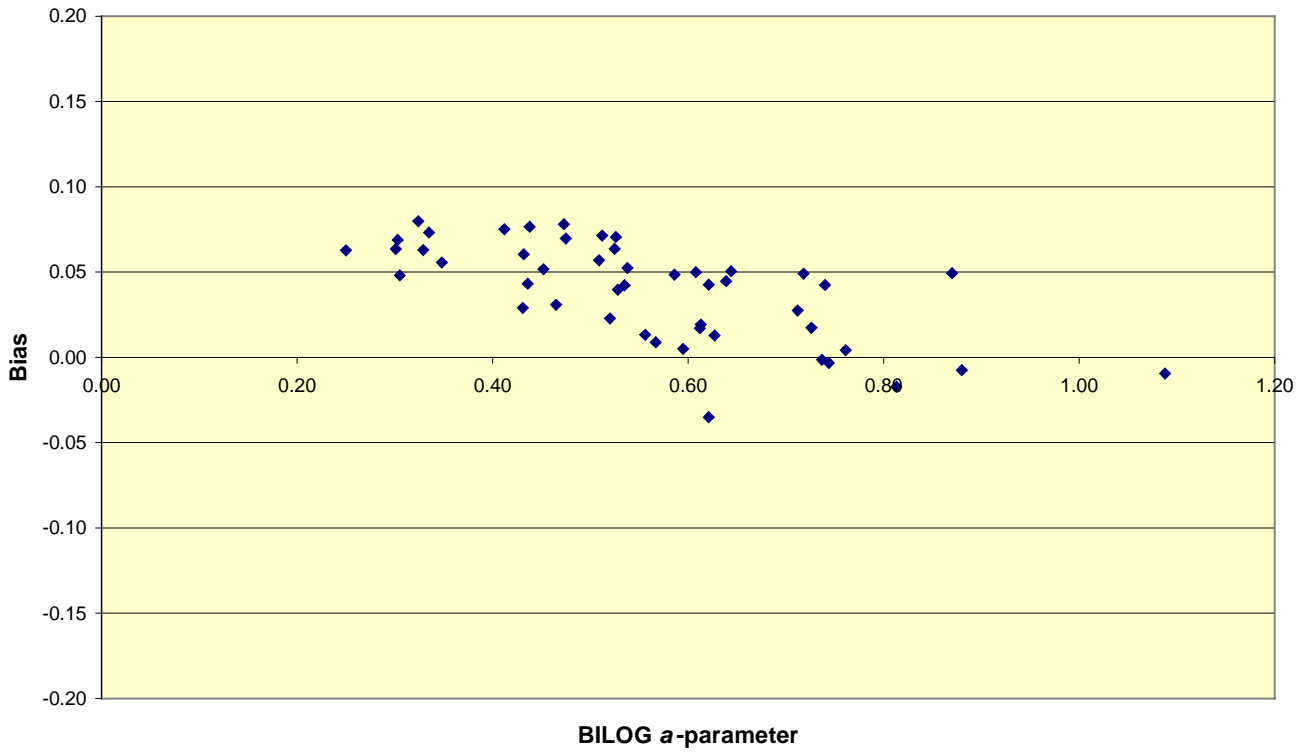


Figure 11. Bias of the Discrimination Parameter as a Function of BILOG Discrimination for the English Exam

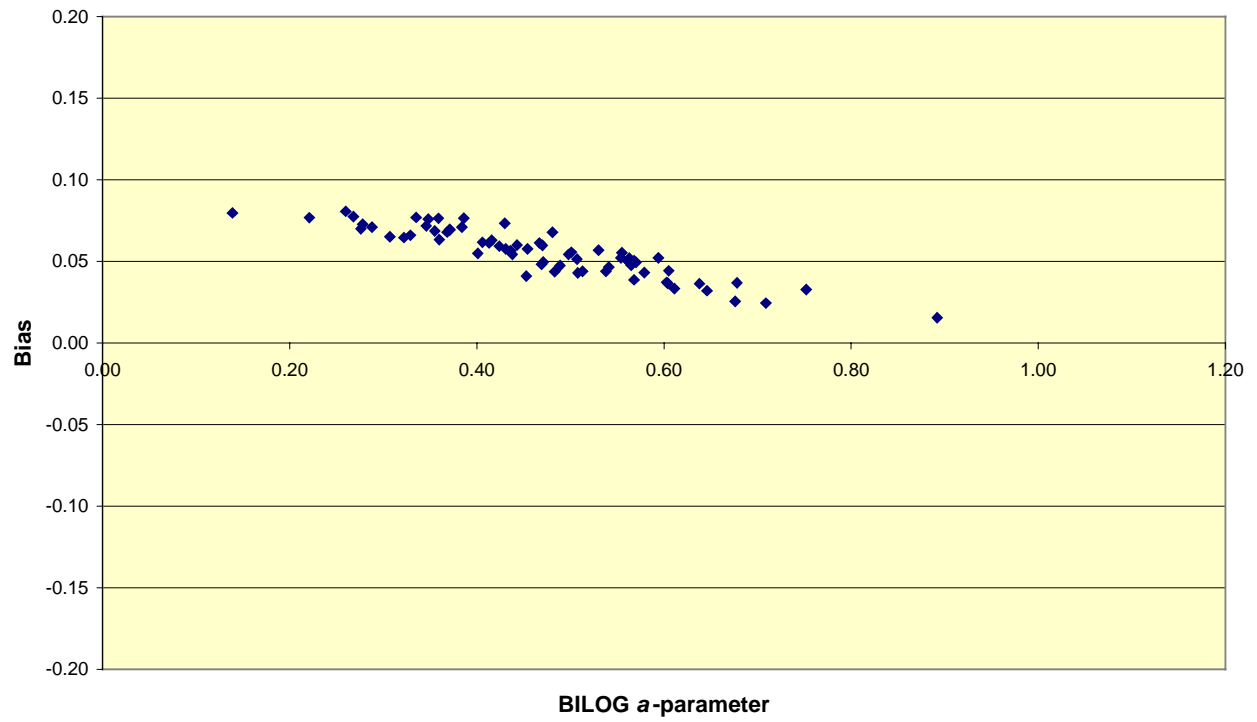


Figure 12. Bias of the IRT Difficulty Parameter as a Function of BILOG Difficulty for the Biology Exam

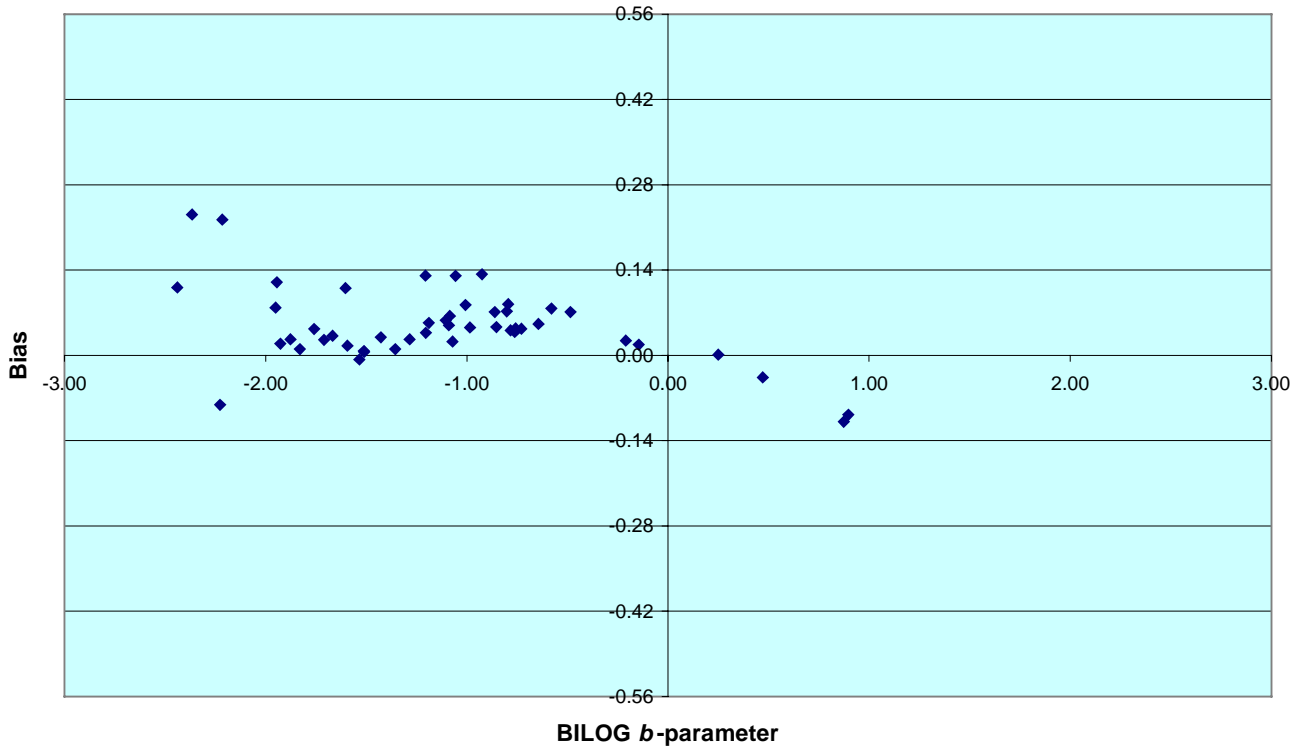


Figure 13. Bias of the Difficulty Parameter as a Function of BILOG Difficulty for the English Exam

