ED 478 076	TM 035 017
AUTHOR	Zhu, Renbang; Yu, Feng
TITLE	Exploring a Stable Difficulty Index for Polytomous Essay Items.
PUB DATE	2003-04-00
NOTE	26p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Difficulty Level; *Essay Tests; Indexes; Measurement Techniques; *Regression (Statistics); *Reliability; Simulation; Test Items
IDENTIFIERS	*Polytomous Items

ABSTRACT

To ensure fairness, it is of critical importance that testing programs make sure that essay items given to examinees are equivalent in difficulty. The purpose of this study was to evaluate the stability and accuracy of a logistic regression based polytomous essay difficulty index. Preliminary results from a simulation study (9 conditions with a 100 replications of each condition) and real data from prior studies suggest that the developed index exhibits three desired statistical properties. First, the index was sensitive to the change of essay difficulty. In addition, the index could be computed easily through logistic regression analysis based on a relatively small sample size. Finally, the index was reasonably stable across different examinee ability distributions. However, the degree of stability was not as good as expected, particularly when examinees' ability distributions differed markedly from the normal condition. Therefore, further modification of the difficulty index and analyses needs to be done before the index can be recommended for practical use.(SLD)

ETS Educational Testing Service

Exploring a Stable Difficulty Index for Polytomous Essay Items

Renbang Zhu

Feng Yu

Educational Testing Service

Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) held between April 21 to 25, 2003, in Chicago, IL.

Abstract

Writing assessment has been introduced into more and more large-scale, high stake tests in the past several years. To ensure that the fairness issue will not be challenged, it is of critical importance that testing programs make sure that essay items given to examinees are equivalent in difficulties. Currently, most practitioners adopt average essay scores, based on pretest data, to index essay difficulties. For paper-pencil tests, this may not appear to be problematic because few essay items are needed while large pretest samples are available to ensure the accuracy and stability of the index. For CBT administration, however, a large supply of essay items is required to support test delivery on a daily basis. Under such circumstance, mean essay score is no longer an adequate index to label essay difficulty because the examinee sample size per item is small when many new essay items need to be pre-tested at the same time. In addition, if new essay items are pre-tested during different periods, examinee ability distributions may differ substantially. As a result, mean essay scores cannot be as stable and accurate as needed. The purpose of this study was to evaluate the stability and accuracy of a logistic regression based essay difficulty index. Preliminary results from the simulation study and real data analysis suggested that this index exhibited three desired statistical properties. First, the index was sensitive to the change of essay difficulty. Further, the index could be easily computed through logistic regression analysis based on a relatively small sample size. Finally, the index was reasonably stable across different examinee ability distributions. However, the degree of stability was not as good as expected, particularly when examinees' ability distributions differed remarkably from the normal condition. Therefore, further modification on the difficulty index and analyses need to be done before recommending for practical use.

Introduction

In the last decade, the use of polytomous essay items has been growing dramatically in order to measure examinees' underlying abilities more accurately. For CBT and some frequently administered tests, it is of critical importance that testing programs make sure that the essay items given to examinees are equivalent in difficulties so that fairness issue will not be challenged. Currently, most practitioners adopt average essay scores, based on pretest data, to index essay difficulties. For traditional paper-pencil tests, few essay items are needed because normally only a limited number of forms are administered in a year. Therefore, using mean essay score as a difficulty index may not appear to be so problematic as long as large typical examinee samples are available to ensure the accuracy and stability of the mean essay score. However, for some of the large-scale CBT testing programs offering tests on a daily basis, large numbers of essay items are required for test security reason. Under such circumstance, a mean essay score is no longer an adequate index to label essay difficulty due to two main reasons. First, when many more new essay items need to be pre-tested at the same time, the examinee sample size per item becomes much smaller. Secondly, new essay items are pre-tested during different time periods of test administrations, across which examinee ability distributions may differ substantially. As a result, mean essay scores cannot be as stable and accurate as needed for the purposes of item banking and item monitoring in test development.

One way to resolve this problem is to apply IRT based models to parameterize essay difficulties. IRT suggests that item difficulty parameter estimates are relatively stable across different examinee ability distributions. In other words, IRT based model parameters are relatively sample independent while the mean item scores are sample dependent. Thus, an IRT method provides test developers more confidence to assemble multiple test forms with equivalent overall difficulties. However, IRT based model parameter estimation is also constrained by sample size requirement. In addition, as a result of parameter scaling, IRT difficulty parameter estimates from different calibrations are not comparable unless an adequate equating is performed.

Logistic regression (Hosmer & Lemeshow, 1989) is another attractive procedure to parameterize essay difficulties. One clear advantage of logistic regression is that it is close to

IRT models in form, but not limited by sample size and model-data fit as required in the IRT method (French & Miller, 1996; Breland, Muraki, Lee, Najarian & Beyer, 2000). A study by Breland et al. indicated that logistic regression methodology functioned well and was easy to use.

The purpose of the current study was to conduct a preliminary investigation on a logistic regression based polytomous essay difficulty index, initially developed by Breland and his colleagues in 2000, through a set of more realistic simulation conditions. The focus this study was on the evaluation of the stability and accuracy of the proposed difficulty index across different examinee ability distributions and sample sizes.

Computation of Polytomous Difficulty Index

A logistic regression equation for a dichotomous item can be expressed as,

$$P(U_j \mid x) = \frac{e^{[U_j g_j(x)]}}{1 + e^{[U_j g_j(x)]}} \qquad (1)$$

where U_j represents measured dichotomous score (0 or 1) of item j, x is an independent variable related to score U_j, and function $g_j(x)$ is called a logit, which can be defined as,

$$g_j(x) = \alpha_j + \beta_j x \qquad (2)$$

where α_j is the intercept, β_j is the slope parameter associated with independent variable x.

The dichotomous model shown in Equation 1 can be extended to polytomous essay items by dichotomizing the polytomous score categories. Agresti (1990) recommended three different ways to dichotomize polytomous score categories: continuation ratio logits, cumulative logits, and adjacent categories. The cumulative logits model was applied in this study. The cumulative logits model follows a stepwise dichotomization procedure. It re-categorizes a polytomously scored item with K score categories into K-1 sets of dichotomous scores. The lowest (or the highest) possible value of the essay score is designated as the reference score category. The probability of obtaining the reference score category is first compared with the probability of obtaining other score categories and then the probability of obtaining each score category is compared with the probability of obtaining all of the categories with higher (or lower) numeric codes than the present score category. Table 1 gives an example for dichotomizing a polytomous essay item with eleven score categories.

Insert Table 1 about here

As shown in Table 1, the score 1.0 category is compared to all other score categories in the first regression analysis; then the score 1.0 and the score 1.5 categories combined is compared to all other score categories, ... and so on.

The cumulative logits model can be formally expressed as,

$$L_{jk} = \ln\left(\frac{\sum_{m=1}^{k} P_{j,m}(x)}{\sum_{n=k+1}^{K} P_{j,n}(x)} \right), \; k = 1, \ldots, K\text{-}1 \tag{3}$$

where L_{jk} stands for a logit in terms of log-odds for the k^{th} dichotomized regression for polytomous item *j*. As shown above, no data are lost in the cumulative logits coding scheme: the response probabilities are moved from the denominator to the numerator as successive stages of regression. For the k^{th} regression, Equation 2 can be rewritten as,

$$g_{jk}(x) = \alpha_{jk} + \beta_j x \tag{4}$$

where the fitted curves are assumed to be parallel for all *K-1* dichotomized regression equations. That is, β_j remains constant. Equation 4 can be further rewritten as,

BEST COPY AVAILABLE

$$g_{jk}(x) = \beta_j(x + \frac{\alpha_{jk}}{\beta_j}) = \beta_j(x - \xi_{jk}) \qquad (5)$$

Since ξ_{jk} in Equation 5 is analogous to the location parameter in IRT for the k^{th} regression equation, Breland proposed that the mean of the ξ_{jk} over all *K-1* equations be used as an overall difficulty index for polytomous essay item *j* (Breland *et. al.*, 2000). The difficulty index $\overline{\xi_j}$ is given as,

$$\overline{\xi_j} = \frac{1}{K-1}\sum_{k=1}^{K-1}\xi_{jk} \qquad (6)$$

Simulation Design and Real Data Analysis

A simulation study was conducted to investigate the stability and accuracy of the proposed difficulty index $\overline{\xi}$. A total of 20 essay items from a large-scale international admission test were selected for the study. Each essay item has 11 score categories, graded as 1.0, 1.5, …, 5.5, 6.0. Table 2 displays the corresponding slope, location, and categorical parameter estimates for each of the 20 items. These estimates were calibrated by using *PARSCALE* (Muraki, 1999). The calibration was based on the generalized partial credit model and a set of real essay writing data.

Insert Table 2 about here

The simulation conditions included in the study intended to mimic the writing section of the admission test. The first factor to be considered in examining the stability of the essay difficulty index was examinee sample size. Three different sample sizes were applied in this study: 500, 800, and 1,000 to represent small, medium, and large sample sizes in essay writing.

The second factor was the examinee writing ability distribution. Three different ability distributions, θ_w, were applied: standard normal, negatively skewed, and positively skewed. Considering the fact that a non-normal distribution in psychological data was typically found to have a skew index of less than 0.8 and a kurtosis between ±0.6 (Pearson & Please, 1975; Fleishman, 1978), coupled with what was observed in the real data of the large-scale admission test, the skewed distributions in this study were simulated with the skew index of 0.6 and –0.9 for the positively and negatively skewed distributions, respectively, and the kurtosis indices were between .01 and –0.5. Figures 1 presents three sample ability distributions based on three sets of simulated data.

Insert Figure 1 about here

Each simulee's essay score was generated using the generalized partial credit model, which can be written as,

$$P_{jk}(\theta_{wi}) = \frac{\exp[\sum_{v_j=0}^{k} 1.7a_j(\theta_{wi} - b_j + d_{jv_j})]}{\sum_{k=0}^{K_j} \exp[\sum_{v_j=0}^{k} 1.7a_j(\theta_{wi} - b_j + d_{jv_j})]}, \quad k = 0, 1, \ldots, K_j \qquad (7)$$

where P_{jk} (θ_{wi}) is the probability for simulee i with writing ability θ_{wi} to obtain a score of k for item j; a_j and b_j are the slope and location parameters of item j; d_{jv}'s are a set of categorical parameters of item j, with associated constraints $d_{j0} = 0$ and $\sum_{v_j=1}^{K_j} d_{jv_j} = 0$ (Muraki, 1992); and K_j represents the maximum possible score categories of item j.

The process of generating score on essay j for simulee i with writing ability θ_{wi} began with computing the categorical probabilities, P_{jk}'s (k = 0, 1, …, K_j), based on Equation 7. Then, the cumulative probabilities were obtained at each score level, as follows,

$$P'_{jk} = \sum_{k=0}^{k} P_{jk}, \quad k = 0, 1, \ldots, K_j. \qquad (8)$$

For each simulee, the final essay score was assigned by comparing a uniform random number between 0 and 1 to the cumulative probabilities. If the random number was smaller than P'_{j0}, then the simulee was given a score of 0, which corresponded to the minimum possible essay score of 1.0. Otherwise, if the randomly generated number was equal to or larger than P'_{jk} but smaller than $P'_{j(k+1)}$, then a score of $k + 1$ was assigned.

Further, the simulee's corresponding ability on the vocabulary section in the same test, θ_{vi}, which was used as the independent variable x in Equation 4 for the logistic regression analysis, was derived from,

$$\theta_{vi} = \rho\theta_{wi} + \sqrt{1-\rho^2}\,\theta'_{vi} \qquad (9)$$

where ρ was the correlation between θ_{wi} and θ_{vi}, which was fixed at 0.34 (based on real data analysis) in the study. θ_{vi}' was the simulee's initial vocabulary ability randomly generated from the same distribution as θ_{wi}.

In summary, the performance of the proposed difficulty index was evaluated under nine different simulation conditions (3 ability distributions × 3 sample sizes). Each condition was replicated 100 times.

Further, the operational data for 10 essay items from the writing section in the admission test were selected to assess the performance of the difficulty index. When selecting these essay items, priority was given to the consideration that these items must represent different writing tasks in the actual writing assessment. Also, each essay item should be responded by a reasonably large number of examinees. Finally, the selected essay items must vary in difficulty levels in terms of the mean essay scores calculated from real operational data. In order to compute the difficulty index for an essay item as defined in Equations 4 through 7, the examinees' estimated verbal ability on the same test were used as the independent variable in the logistic regression analyses. The correlation between the verbal ability and the writing score in the observed data was found to be 0.57.

The SAS LOGISTIC procedure was used to obtain the estimates for the slope parameter and the intercept parameters for the logistic regression model. The logistic regression package also provides the minus twice log likelihood, abbreviated as –2LL, which can be used to evaluate the goodness-of-fit for the logistic regression model. Log likelihood is negative, whereas -2LL is positive. Large values of –2LL indicate poor predication of the dependent variable. The SAS PROC LOGISTIC output contains the –2LL statistic for the model with no independent variables (i.e., intercept only) in the equation and the –2LL statistics for the full model that has both intercept and the independent variable(s). The absolute difference between the two –2LLs is a χ^2 random variable with a known distribution. If χ^2 is statistically significant, we may conclude that the inclusion of the independent variables has significantly improved the goodness-of-fit for the logistic regression model (Homes & Lemeshow, 1989; Menard, 2002).

It is possible that the model-data fit is statistically significant, especially for large sample, but the relationship between the dependent variable and the independent variable(s) could be insubstantial. Therefore, it is necessary to further investigate the association between the independent variable(s) and the dependent variable in the logistic regression model. The likelihood ratio test statistic, or R^2_L, can be used for this purpose (Menard, 2000, also see Agresti, 1990; Homes & Lemeshow, 1989). R^2_L is a measure of the proportional reduction in the –2LL when comparing the full model with the predictors and the model without. A large R^2_L indicates a substantial improvement of model-data fit.

Results

The mean model-data fit statistics for the simulated items under each ability distribution condition, with sample size 1000, are reported in Table 3. Similar patterns were observed under the other two sample size conditions.

Insert Table 3 about here

As shown in the footnote, all χ^2 values were highly significant (df =1; $p \leq 0.0001$), which suggested that inclusion of the vocabulary variable as the independent variable had significantly improved the model-data fit under all ability conditions.

As also shown in Table 3, generally speaking the R^2_L statistics were the largest under the normal distribution, while the R^2_L statistics under the two non-normal conditions were similar. Theoretically, R^2_L falls between 0 (the independent variables in the model are useless in predicting the dependent variable) and 1 (the model predicts the dependent variable with perfect accuracy). However, it is very unlikely to find zero or perfect prediction in a practical setting. The R^2_L statistics for most items in this study fell between 0.05 and 0.08 under normal condition and between 0.04 and 0.05 under non-normal conditions. The R^2_L statistics with this magnitude were certainly not very impressive. However, as a reference point, a logistical regression model with four predictors, proposed by Menard in 2002 and claimed to be working "fairly well," yielded a R^2_L value of 0.28. On average, each predictor added to the model brought about a reduction in –2LL by 0.07. Therefore, considering the fact that the logistic regression model in our study contained one predictor only, it appeared that the predictor variable was doing quite a good job.

Table 4 contains the model-data fit statistics for items based on the real data analyses.

Insert Table 4 about here

The results from the real data analyses showed that the χ^2 statistics were all highly significant (df=1; $p \leq 0.0001$), which indicated that verbal ability was a good predictor in our study. On average, the inclusion of this predictor variable in the model reduced the –2LL by about 10 percent. Noted in Table 4, both the χ^2 and the R^2_L statistics computed based on real data were much higher than the values based on the simulation data. This was probably because the correlation between the verbal ability and the writing score in the observed data (r = .57) was much larger than the correlation between the vocabulary ability and the writing score in the simulation data (r = .34).

Table 5 below contains the logistic regression based difficulty indices as well as the mean essay scores for each essay item under each simulation condition. The results shown here were averaged across 100 replications.

Insert Table 5 about here

It was our expectation that the difficulty index developed through logistic regression analysis could possess such a statistical property that the index value monotonically increases as the degree of item difficulty level increases. The simulation results in Table 5 indicated that the difficulty index was fairly sensitive to the change of mean essay scores for the 20 simulation items. Figure 2 provides a visual presentation of the relationship between the measured difficulty index value and the mean essay score under each simulated ability distribution condition, based on the sample size of 500.

Insert Figure 2 about here

Except for one item (i.e., item #16) in the normal and negatively skewed distribution conditions, it can be seen in the figure that, overall, the difficulty index monotonically increased as the mean essay score increased. The figure also shows that the mean essay scores were pretty sensitive to the variation of ability conditions, especially from the positively to the negatively skewed distribution. Under the positively skewed condition, the mean essay scores fell between 3.7 and 3.9, while under the negatively skewed condition the mean essay scores fell between 3.9 and 4.1.

Another way of examining the sensitivity of mean essay score and logistic regression difficulty index to ability distribution was to compare the percentage of change across different ability distributions. For the mean essay score as well as the difficulty index, the values from the normal condition were used as a baseline to compute the percentage of change under the skewed distribution conditions. Then, the difference between the percentage of change in mean essay score and the percentage of change in the difficulty index (PCTDIFF) was calculated. A positive value in PCTDIFF suggested that the difficulty index is relatively more stable than the mean essay score. The results of comparing the percentage of change based on sample size of 1000 are shown in Table 6 and Figure 3. Similar patterns were observed under the other two sample size conditions.

Insert Table 6 and Figure 3 about here

As shown in Table 6 and Figure 3, about two-thirds of the items had a positive value in PCTDIFF, which indicated that the difficulty index appeared to be more stable than the mean essay score when examinees' ability distribution varied from normal to skewed conditions.

Figure 4 gives an item-by-item comparison of logistic regression based difficulty indices across three ability distributions under each sample size condition.

Insert Figure 4 about here

As can be seen, the performance of the difficulty index was fairly stable. For most items, the index values differed no more than half a point across different ability distributions, which was equivalent to a 0.125 change in index value. An additional feature revealed from this comparison is that the index performance tended to be more consistent between the normal and the negatively skewed distributions. This finding is important because empirical data in essay writing suggested that item level score distributions are usually normal or slightly negatively skewed.

As also displayed in Figures 4, the index stability improved slightly when the sample size increased, but the improvement was noticeable only when the sample size changed from 500 to 800. Figure 5 provides a comparison of index performance across different sample size conditions under each ability distribution condition.

Insert Figure 5 about here

Figure 5 shows that sample size variable had very little impact on the stability of the difficulty index. For almost all items under each ability distribution condition, the variation of difficulty index was less than 0.2 point in absolute value. About half of the simulation items had a discrepancy of 0.1 point in index values from one sample condition to the other.

Finally, the computed difficulty indices for 10 items based on the operational data are reported in Table 7. Also included are the mean essay score, standard deviation, and sample size for each item.

Insert Table 7 about here

Due to the availability of operational data, only a small number of real items administered within a limited period of time were selected for the current study. The results in Table 7 clearly indicated that the index was capable of reflecting the change in difficulty level of real essay items because, overall, the index values monotonically increased when the observed mean essay score increased. Figure 6 describes the relationship between the logistic regression difficulty index and the mean essay score.

Insert Figure 6 about here

Conclusion

Writing assessment has been introduced into more and more large-scale, high stake tests in the past several years. To ensure that all examinees are tested fairly by administering essay items with comparable difficulties in the same or in different test administrations, it is important for testing programs to have a reliable and adequate statistical procedure in place to provide accurate and stable measure on item difficulties.

The results of this study suggested that a difficulty index for polytomously scored essay items obtained through logistic regression methodology seemed to be promising. The first desirable feature of this difficulty index was its stability across different simulated ability distributions. However, it is worth noting that this stability holds only when the distributions of examinees' abilities do not differ remarkably from the normal condition.

Experience in the current study suggested that, when the ability distributions were highly skewed (e.g., skew > 1.0), the difficulty index could differ significantly from what was desired. A highly skewed distribution, which is typically associated with one thin tail, may also lead to large standard errors relative to the size of the coefficients.

An additional nice property found in this study was the insensitivity of the difficulty index to sample size variation. The results indicated that using a sample of 500 examinees in real life situation is probably good enough for obtaining a reliable estimate of the difficulty parameter for an essay item. Allowing a relatively small sample for item analysis has a significant implication for test development because many more new essay items can be pre-tested at the same time. Of course, the examinee sample size per item must not be so small as to limit the capability of logistic regression analysis, which normally requires observations across all categories for a polytomously scored essay item.

The results from the operational essay writing data analyses also turned out to be quite encouraging, in that the difficulty index seemed to be related to the observed mean essay score fairly well. When applying the logistic regression methodology to the real data, it is of critical importance that the independent variable should be sufficiently predictive so that the explanatory power of the variation in the dependent variable can be maximized. Also, in order to maintain the same scale for the difficulty indices estimated over time, care must be taken to ensure that the predictors used in the logistic regression do not change in measurement scale or in other statistical properties.

Since, as mentioned above, the stability of the logistic regression difficulty index was not as good as desired, particularly when examinees' ability distributions differed substantially from the normal condition, further modification on the difficulty index and analyses need to be done before recommending for practical use.

References

Agresti, A. (1990). *Categorical data analysis.* New York, NY: Wiley.

Breland, H., Muraki, E., Lee, Y., Najarian, M., and Beyer, J. (2000). Comparability in TOEFL CBT essay prompts using the logistic regression method. Paper presented at the annual conference National Council on Measurement in Education (NCME) in New Orleans, Louisiana.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532.

French, A.W., and Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 33 (3), 315-332.

Hosmer, D. W., and Lemeshow, S. (1989). *Applied logistic regression*. New York, NY: Wiley.

Mendard, S. (2002). *Applied logistic regression analyses*. Sage Publications, Inc.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., and Bock, R. D. (1999). *PARSCALE* (Version 3.2): *Analysis of graded responses and ratings* [Computer program]. Chicago, IL: Scientific Software International, Inc.

Pearson, E.S., and Please, N.W. (1975). Relation between the shape of population distribution of four simple test statistics. *Biometrika, 62*, 223-241.

Table 1 Dichotomizing a Polytomous Essay Item with Eleven Score Categories

Type of Regression	Polytomous Score Categories	Regression Equation #	Score Categories Recategorized as 0	Score Categories Recategorized as 1
Cumulative Logits	1.0			
	1.5	1	1.0	1.5, 2.0, …, 5.5, 6.0
	2.0	2	1.0, 1.5,	2.0, 2.5, …, 5.5, 6.0
	2.5	3	1.0, 1.5, 2.0	2.5, 3.0, …, 5.5, 6.0
	3.0	4	1.0, 1.5, 2.0, 2.5	3.0, 3.5, …, 5.5, 6.0
	3.5	5	1.0, 1.5, 2.0, 2.5, 3.0	3.5, 4.0, …, 5.5, 6.0
	4.0	6	1.0, 1.5, …, 3.0, 3.5	4.0, 4.5, 5.0, 5.5, 6.0
	4.5	7	1.0, 1.5, …, 3.5, 4.0	4.5, 5.0, 5.5, 6.0
	5.0	8	1.0, 1.5, …, 4.0, 4.5	5.0, 5.5, 6.0
	5.5	9	1.0, 1.5, …, 4.5, 5.0	5.5, 6.0
	6.0	10	1.0, 1.5, …, 5.0, 5.5	6.0

Table 2 Item Parameters for Simulating Essay Scores using Generalized Partial Credit Model

Item	a	b	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
1	1.099	-0.806	2.097	1.542	0.823	1.059	0.499	0.058	-0.896	-1.153	-1.781	-2.247
2	1.401	-0.915	1.803	2.007	1.184	1.090	0.226	0.119	-0.916	-1.209	-1.883	-2.421
3	0.874	-0.808	2.185	1.910	1.481	1.000	0.334	0.010	-0.912	-1.402	-1.974	-2.632
4	2.059	-0.927	2.242	1.524	1.043	0.961	0.232	-0.140	-0.770	-1.184	-1.659	-2.250
5	0.278	-0.788	-0.579	4.823	0.602	2.101	0.532	0.478	-1.541	-1.204	-2.380	-2.832
6	1.884	-0.954	1.997	1.651	1.338	0.903	0.397	0.004	-0.761	-1.258	-1.830	-2.442
7	1.210	-0.859	1.939	1.843	1.223	1.110	0.208	0.104	-0.817	-1.219	-1.948	-2.443
8	1.223	-0.826	1.830	1.752	1.295	0.949	0.231	-0.015	-0.890	-1.157	-1.724	-2.270
9	1.244	-0.893	1.975	1.656	1.842	1.052	0.182	-0.028	-0.945	-1.301	-1.909	-2.525
10	1.773	-0.949	2.094	1.658	1.371	0.982	0.317	-0.021	-0.882	-1.233	-1.894	-2.394
11	0.937	-0.834	2.211	2.063	1.180	1.143	0.233	0.126	-0.949	-1.410	-2.012	-2.586
12	0.841	-0.824	1.984	2.297	1.292	1.064	0.228	0.128	-1.071	-1.336	-2.106	-2.480
13	1.146	-0.844	2.103	1.859	1.285	1.019	0.264	0.059	-0.910	-1.252	-1.969	-2.457
14	1.302	-0.823	1.967	1.914	0.850	0.887	0.163	0.024	-0.780	-1.090	-1.736	-2.199
15	0.531	-0.782	2.136	2.579	0.764	1.112	-0.130	0.363	-1.162	-0.979	-2.320	-2.362
16	0.164	-0.934	2.548	5.771	-0.929	1.890	0.067	0.509	-2.032	-1.125	-3.298	-3.402
17	0.829	-0.786	2.188	2.084	0.615	1.182	0.000	0.144	-0.865	-1.020	-2.052	-2.276
18	0.445	-0.781	2.124	2.877	0.292	1.185	-0.119	0.382	-1.054	-1.030	-2.221	-2.437
19	0.960	-0.805	2.329	1.997	0.762	1.025	0.198	0.084	-1.007	-0.977	-1.968	-2.444
20	0.526	-0.760	2.069	2.554	0.634	1.076	-0.075	0.222	-1.096	-0.865	-2.054	-2.466

Table 3 -2 Log Likelihood, χ^2 Statistics and R^2_L Coefficients, averaged across 100 Replications, for Each Item under 3 Ability Distribution Conditions (Simulation Data, N=1,000)

ABILITY	NORMAL				NEG_SKEW				POS_SKEW			
Item	Intercept Only	Intercept and Covariate	χ^{2*}	R^2_L	Intercept Only	Intercept and Covariate	χ^{2*}	R^2_L	Intercept Only	Intercept and Covariate	χ^{2*}	R^2_L
1	4127.5	3833.3	294.2	0.07	4495.2	4307.5	187.7	0.04	4650.3	4432.5	217.8	0.05
2	3986.3	3683.4	303.0	0.08	4396.2	4199.4	196.9	0.04	4557.7	4327.6	230.1	0.05
3	4057.9	3796.4	261.4	0.06	4447.7	4269.2	178.6	0.04	4635.5	4425.1	210.4	0.05
4	4106.1	3774.4	331.7	0.08	4505.6	4303.1	202.5	0.05	4553.7	4323.4	230.3	0.05
5	4174.5	4049.2	125.3	0.03	4466.9	4346.2	120.7	0.03	4598.2	4440.6	157.6	0.03
6	3981.1	3656.3	324.8	0.08	4386.3	4188.0	198.3	0.05	4590.5	4362.7	227.8	0.05
7	4036.3	3743.4	292.9	0.07	4414.7	4225.5	189.2	0.04	4620.0	4401.5	218.5	0.05
8	4142.7	3847.3	295.4	0.07	4508.8	4315.2	193.6	0.04	4642.4	4428.3	214.1	0.05
9	4013.8	3720.6	293.3	0.07	4404.6	4209.3	195.3	0.04	4516.4	4297.4	219.0	0.05
10	3983.1	3659.0	324.1	0.08	4405.0	4205.7	199.3	0.05	4550.3	4326.0	224.3	0.05
11	4019.7	3759.1	260.6	0.06	4426.6	4251.2	175.4	0.04	4603.6	4394.9	208.7	0.05
12	4043.7	3781.7	262.0	0.06	4441.1	4265.5	175.6	0.04	4593.4	4377.0	216.3	0.05
13	4046.1	3755.7	290.4	0.07	4443.8	4253.6	190.2	0.04	4616.7	4394.2	222.5	0.05
14	4190.9	3883.7	307.2	0.07	4521.4	4323.9	197.5	0.04	4601.8	4372.7	229.0	0.05
15	4224.2	4007.1	217.1	0.05	4538.8	4374.6	164.3	0.04	4636.7	4437.6	199.1	0.04
16	4371.1	4291.2	79.9	0.02	4528.3	4442.1	86.2	0.02	4598.2	4485.2	113.0	0.02
17	4192.1	3929.9	262.2	0.06	4509.3	4324.8	184.5	0.04	4625.3	4414.2	211.1	0.05
18	4287.5	4086.0	201.5	0.05	4564.9	4409.7	155.2	0.03	4645.5	4459.4	186.1	0.04
19	4137.4	3859.1	278.3	0.07	4477.9	4290.6	187.3	0.04	4613.8	4398.8	215.0	0.05
20	4294.3	4073.4	221.0	0.05	4575.0	4413.8	161.2	0.03	4657.3	4459.3	198.0	0.04

* The Chi-Square test statistics, with $df=1$, are all significant at *0.0001* level.

Table 4 -2 Log Likelihood, χ^2, and R^2_L Statistics for Each Essay Item (Operational Data)

Item	Intercept Only	Intercept and Covariate	Chi-Square	*df*	Prob > Chi-Square	R_L^2
1	4853.4	4360.9	492.5	1	< 0.0001	0.10
2	6211.6	5664.7	546.9	1	< 0.0001	0.09
3	7999.5	7300.7	698.8	1	< 0.0001	0.09
4	7534.2	6880.9	653.4	1	< 0.0001	0.09
5	9104.1	8149.6	954.5	1	< 0.0001	0.11
6	11142.2	10207.8	934.4	1	< 0.0001	0.08
7	4809.5	4374.4	435.1	1	< 0.0001	0.09
8	10075.6	9098.3	977.3	1	< 0.0001	0.10
9	7349.4	6600.4	749.0	1	< 0.0001	0.10
10	12238.1	11246.1	992.0	1	< 0.0001	0.08

Table 5 **Logistic Regression Difficulty Index and Mean Essay Score, averaged across 100 Replications, under Each Ability and Sample Size Condition (Simulation Data)**

ABILITY	NORMAL						NEG_SKEW						POS_SKEW					
N	500		800		1000		500		800		1000		500		800		1000	
Item	Index	Score	Index	Score	Index	Score	Index	Score	Index	Score	Index	Score	Index	Score	Index	Score	Index	Score
1	2.66	4.32	2.73	4.33	2.74	4.32	2.62	3.95	2.72	3.96	2.71	3.96	2.08	3.75	2.12	3.76	2.09	3.76
2	3.13	4.39	3.17	4.39	3.28	4.39	2.90	4.02	2.92	4.03	3.05	4.03	3.34	3.87	3.23	3.88	3.30	3.88
3	2.71	4.23	2.96	4.24	2.85	4.24	2.31	3.89	2.42	3.90	2.33	3.90	2.76	3.79	2.82	3.80	2.71	3.78
4	2.94	4.42	3.13	4.43	3.17	4.43	3.31	4.07	3.40	4.08	3.37	4.07	3.21	3.85	3.33	3.86	3.28	3.85
5	3.20	4.17	3.23	4.17	3.14	4.17	2.92	3.87	3.05	3.89	3.00	3.88	3.01	3.82	2.88	3.83	2.93	3.82
6	3.14	4.45	3.30	4.45	3.21	4.44	3.17	4.08	3.06	4.06	3.18	4.08	3.45	3.91	3.55	3.92	3.52	3.91
7	2.94	4.33	2.95	4.35	2.92	4.34	2.74	3.98	2.84	3.99	2.62	3.97	2.78	3.84	2.81	3.83	2.81	3.82
8	2.88	4.33	2.86	4.33	2.83	4.32	2.80	3.98	2.81	3.98	2.76	3.96	2.31	3.79	2.40	3.79	2.41	3.79
9	3.17	4.33	3.13	4.32	3.19	4.32	2.75	3.98	2.76	3.99	2.74	3.98	3.47	3.87	3.43	3.88	3.46	3.87
10	3.05	4.42	3.27	4.42	3.35	4.41	3.07	4.05	3.11	4.05	3.13	4.05	3.70	3.91	3.64	3.90	3.65	3.91
11	2.75	4.26	2.94	4.26	2.95	4.26	2.38	3.92	2.46	3.91	2.47	3.92	2.91	3.80	2.95	3.80	3.03	3.80
12	2.83	4.24	2.91	4.23	2.88	4.23	2.37	3.90	2.42	3.90	2.49	3.90	2.91	3.80	2.90	3.79	2.90	3.79
13	2.89	4.30	2.91	4.30	2.95	4.30	2.59	3.95	2.57	3.95	2.55	3.95	2.74	3.80	2.89	3.82	2.91	3.81
14	2.72	4.34	2.74	4.34	2.73	4.34	2.90	3.99	2.88	3.98	2.85	3.98	2.34	3.77	2.27	3.75	2.45	3.77
15	2.63	4.19	2.68	4.18	2.66	4.18	2.54	3.87	2.58	3.87	2.66	3.87	2.33	3.72	2.29	3.72	2.28	3.72
16	2.86	4.00	3.04	4.01	3.03	4.00	3.49	3.79	3.26	3.79	3.29	3.78	2.89	3.73	3.15	3.74	2.96	3.73
17	2.56	4.26	2.67	4.26	2.59	4.26	2.44	3.91	2.55	3.92	2.48	3.91	2.16	3.72	2.24	3.73	2.13	3.72
18	2.55	4.16	2.62	4.18	2.59	4.18	2.64	3.87	2.66	3.87	2.71	3.87	2.03	3.70	2.12	3.69	2.14	3.70
19	2.61	4.27	2.78	4.28	2.76	4.28	2.55	3.93	2.57	3.94	2.57	3.93	2.43	3.75	2.48	3.75	2.47	3.75
20	2.47	4.17	2.58	4.19	2.55	4.19	2.56	3.87	2.53	3.88	2.57	3.88	2.02	3.70	2.07	3.69	2.09	3.70

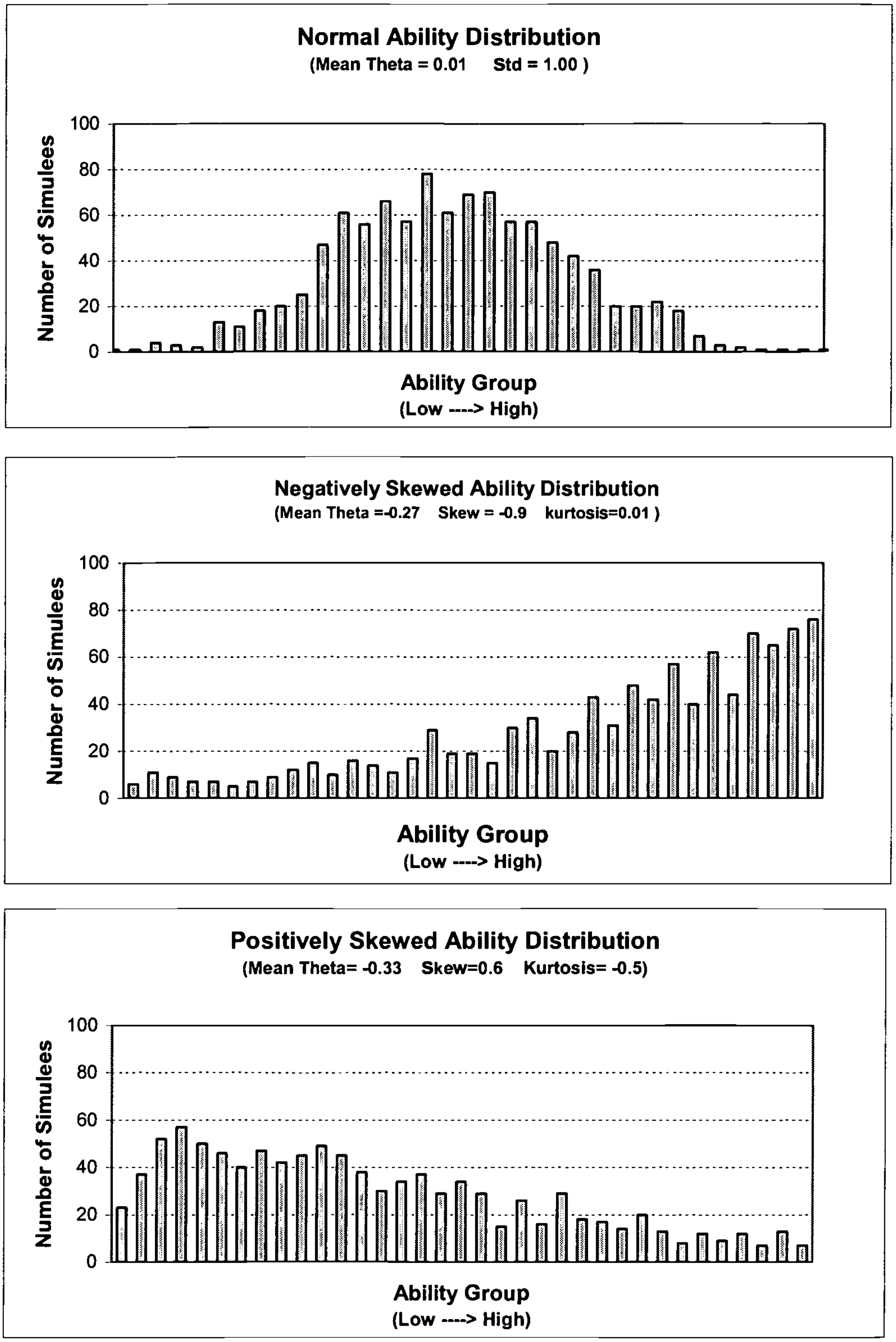
Table 6 Percentage of Change in Mean Essay Score and Difficulty Index (Baseline=Normal)

ABILITY	NEG_DIFF			POS_DIFF		
Item	% Change in Mean Score	% Change in Difficulty Index	PCTDIFF	% Change in Mean Score	% Change in Difficulty Index	PCTDIFF
1	0.08	0.01	0.07	0.13	0.24	-0.11
2	0.08	0.07	0.01	0.12	0.01	0.11
3	0.08	0.18	-0.10	0.11	0.05	0.06
4	0.08	0.07	0.01	0.13	0.03	0.10
5	0.07	0.04	0.03	0.09	0.07	0.02
6	0.08	0.01	0.07	0.12	0.10	0.02
7	0.08	0.10	-0.02	0.12	0.04	0.08
8	0.08	0.03	0.05	0.12	0.15	-0.03
9	0.08	0.14	-0.06	0.10	0.09	0.01
10	0.08	0.06	0.02	0.11	0.09	0.02
11	0.08	0.16	-0.08	0.11	0.03	0.08
12	0.08	0.13	-0.05	0.10	0.01	0.09
13	0.08	0.13	-0.05	0.11	0.01	0.10
14	0.08	0.04	0.04	0.13	0.10	0.03
15	0.07	0.00	0.07	0.11	0.14	-0.03
16	0.06	0.09	-0.03	0.07	0.02	0.05
17	0.08	0.05	0.03	0.13	0.18	-0.05
18	0.07	0.05	0.02	0.11	0.18	-0.07
19	0.08	0.07	0.01	0.12	0.11	0.01
20	0.07	0.01	0.06	0.12	0.18	-0.06

Table 7 Difficulty Index and Mean Essay Score for 10 Real Essay Items (Operational Data)

Item	Difficulty Index	Mean Essay Score	STD	N
1	**2.64**	*4.32*	1.016	1175
2	**3.21**	*4.24*	0.952	1525
3	**2.99**	*4.43*	1.008	1942
4	**3.48**	*4.44*	0.925	1891
5	**3.11**	*4.33*	1.015	2193
6	**3.25**	*4.41*	0.933	2791
7	**2.96**	*4.23*	0.954	1185
8	**2.72**	*4.34*	1.051	2400
9	**2.66**	*4.26*	1.049	1751
10	**2.36**	*4.18*	1.101	2850

Figure 1 Three Ability Distributions for the Simulation Study of Essay Difficulty Index



Normal Ability Distribution

(Mean Theta = 0.01 Std = 1.00)

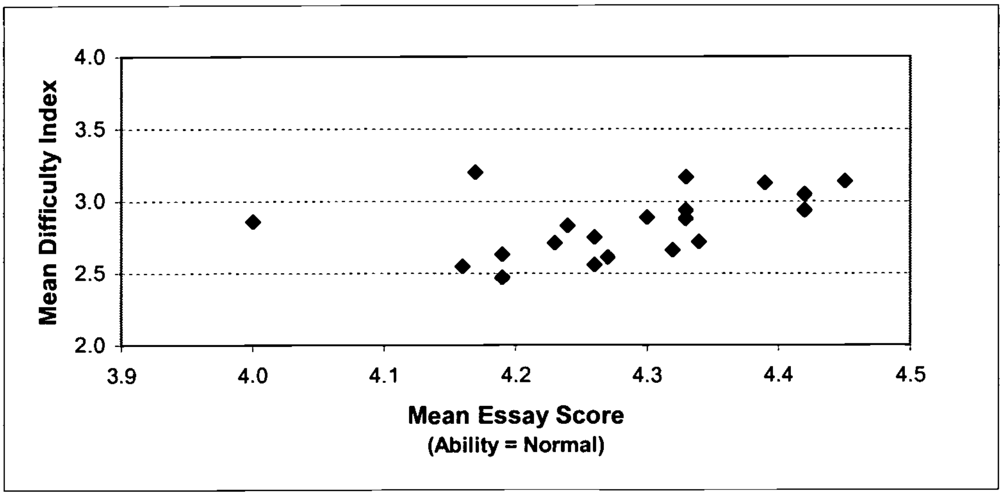
Negatively Skewed Ability Distribution

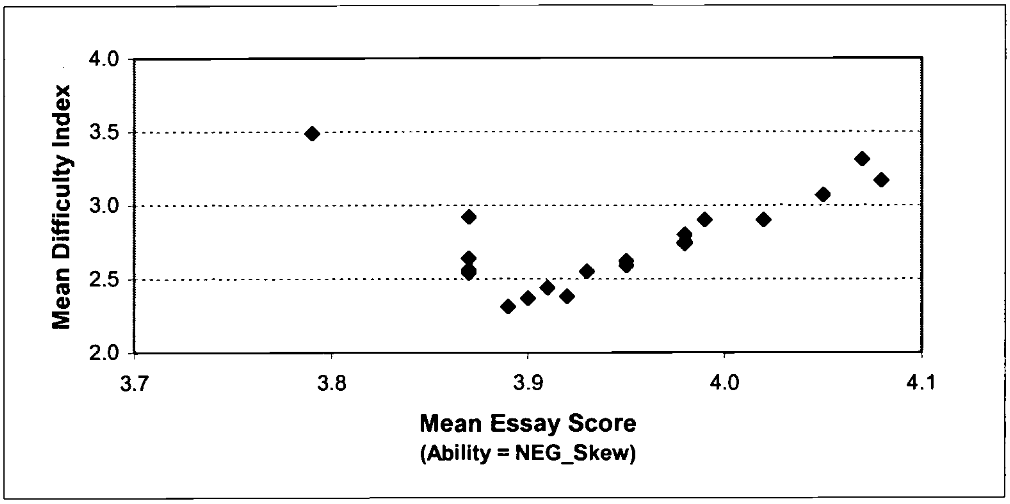
(Mean Theta =-0.27 Skew = -0.9 kurtosis=0.01)

Positively Skewed Ability Distribution

(Mean Theta= -0.33 Skew=0.6 Kurtosis= -0.5)

Figure 2 Plot for Difficulty Index and Mean Essay Score (Simulation Data, N=500)





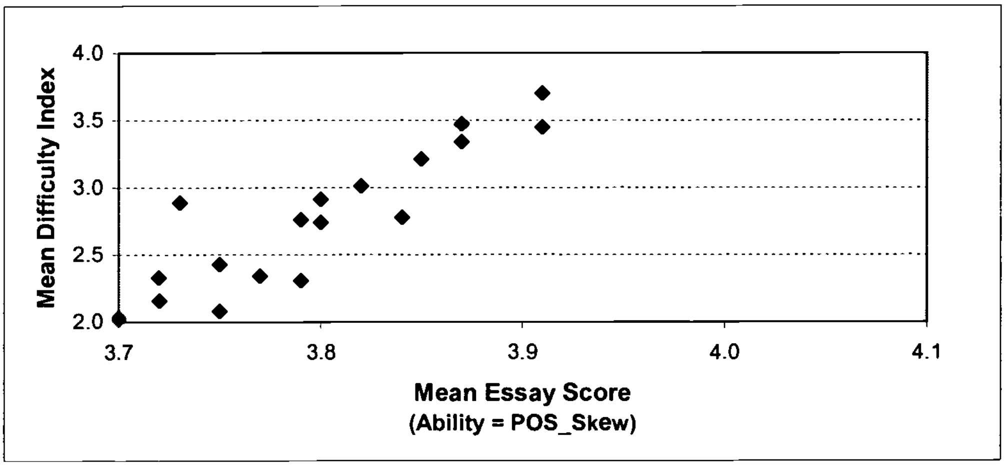


Figure 3 Difference of % Change for Mean Essay Score and Difficulty Index (Baseline=Normal)

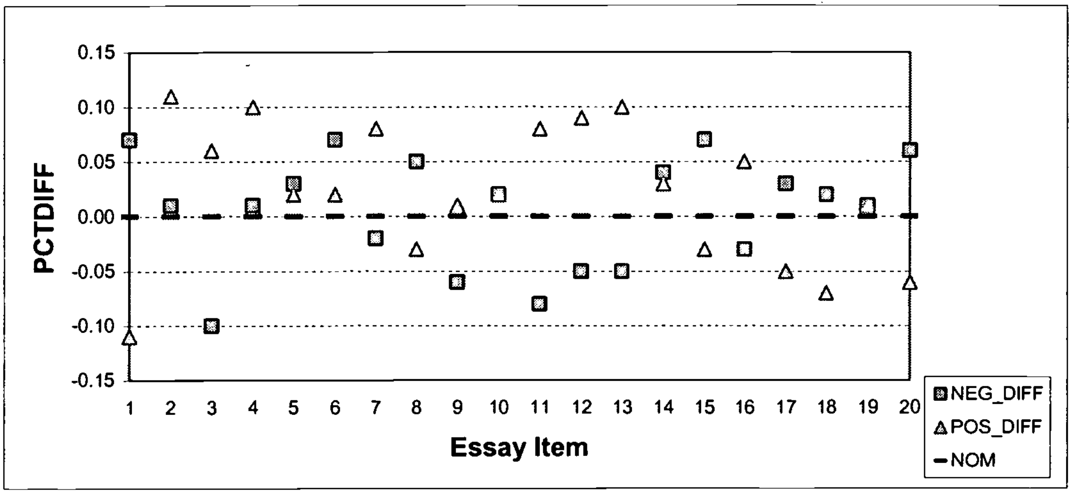


Figure 4 Stability of Difficulty Index across Different Ability Distribution Conditions (Simulation Data)

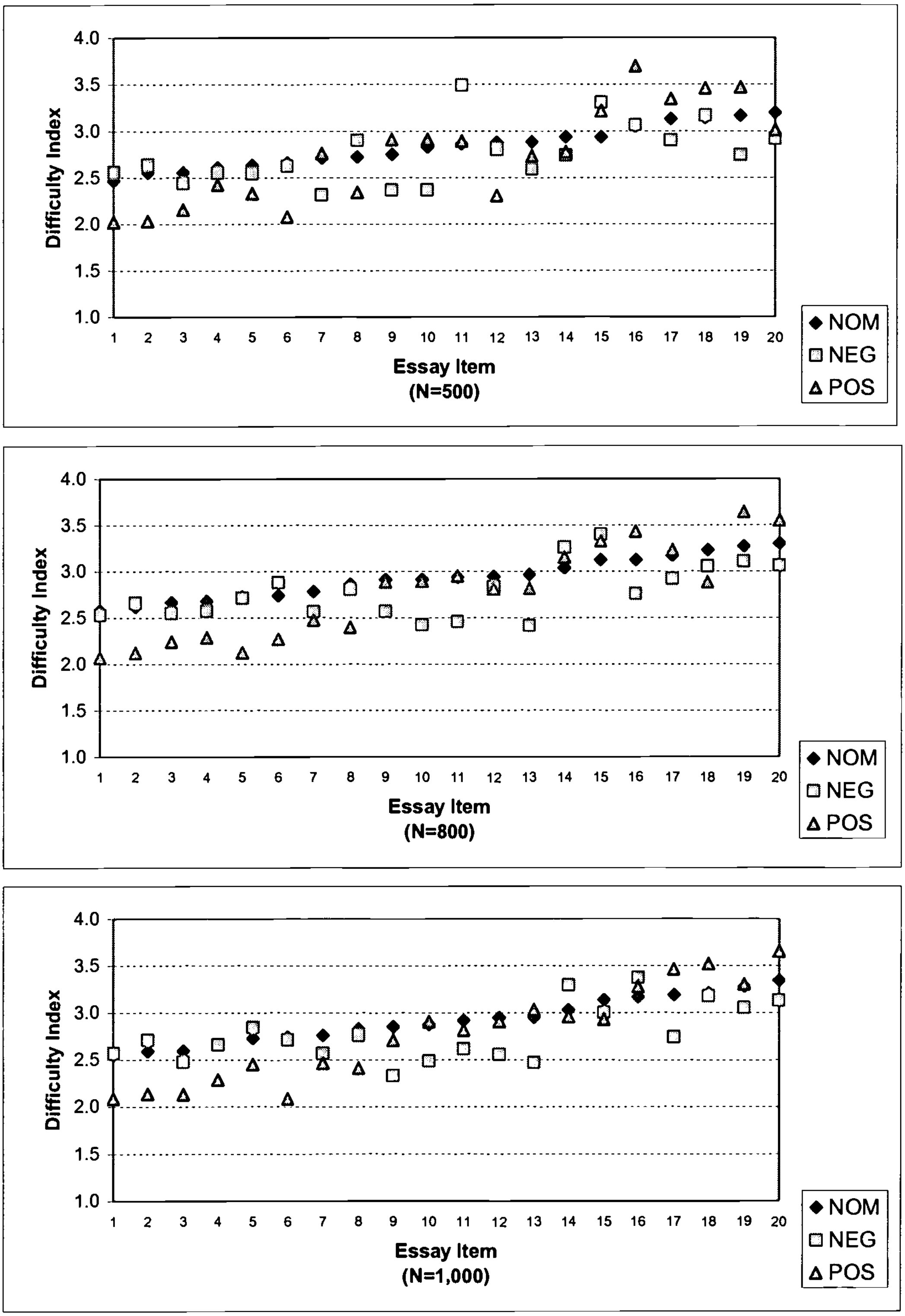
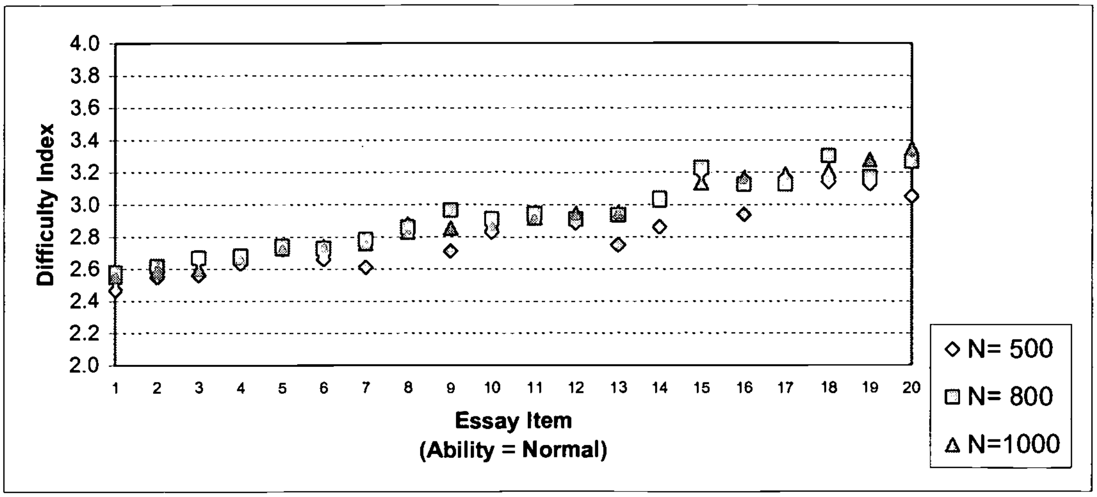
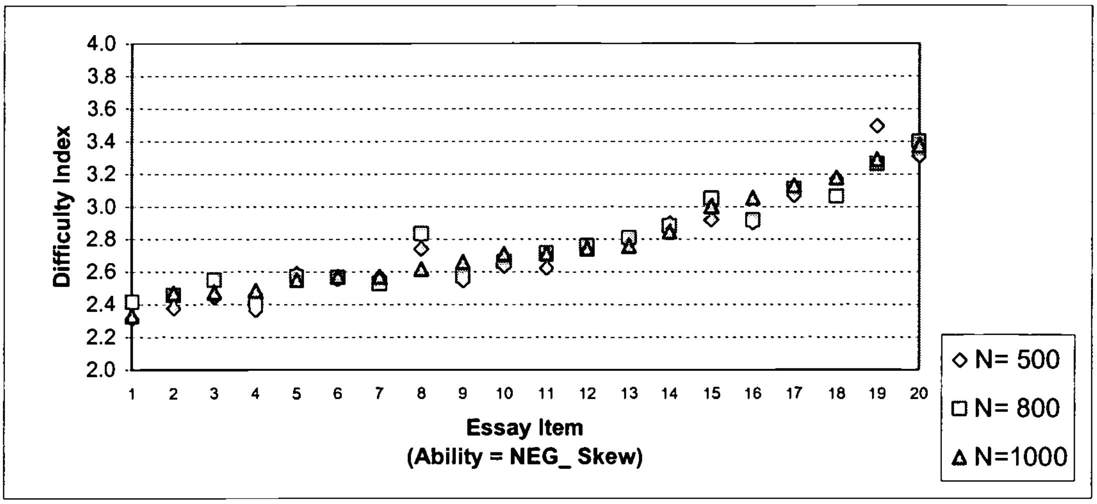


Figure 5 Stability of Difficulty Index across Different Sample Size Conditions (Simulation Data)





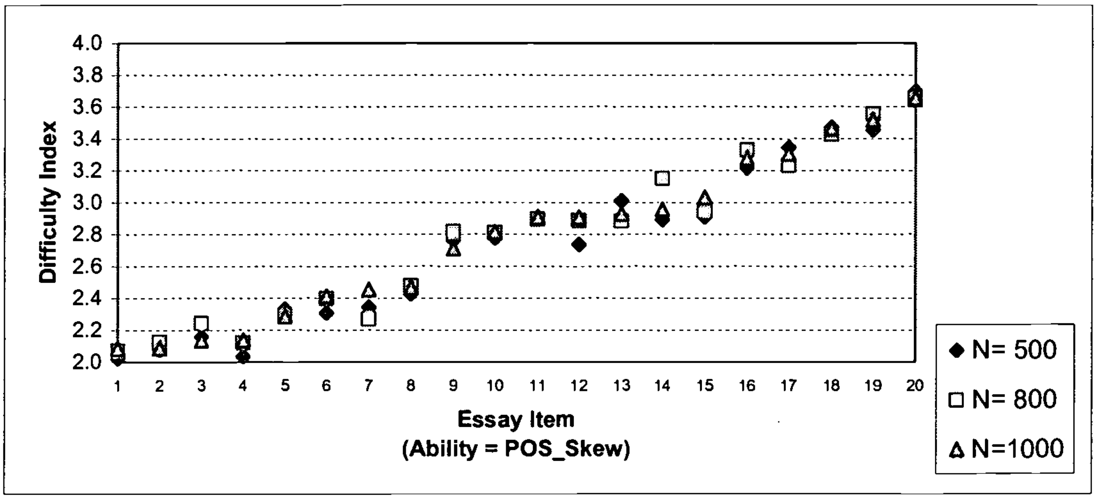
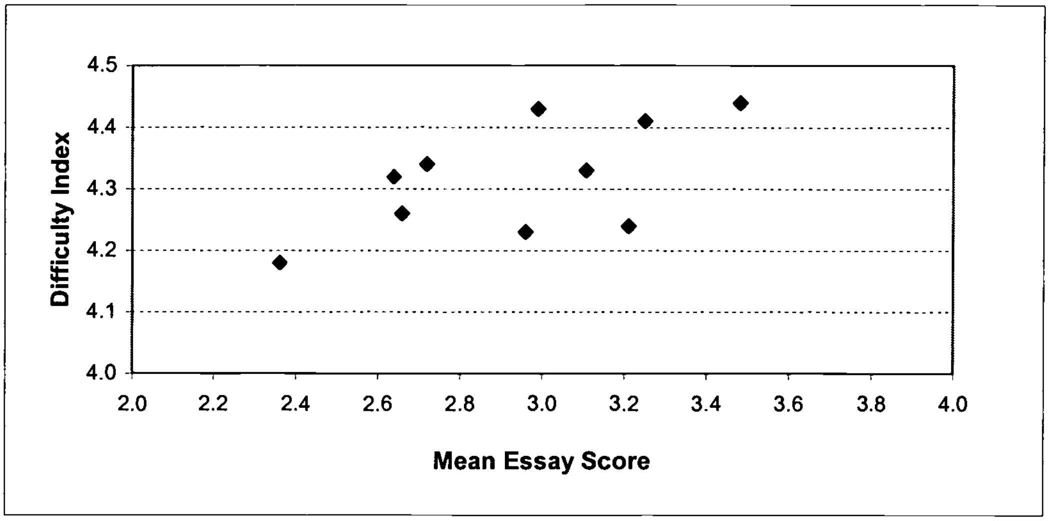


Figure 6 Difficulty Index and Mean Essay Score for Real Essay Items (Operational Data)



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035017

I. DOCUMENT IDENTIFICATION:

Title: Exploring a stable dificulty index for polytomous essay items	
Author(s): RENBANG ZHU, FENG YU	
Corporate Source: Educational Testing Service	Publication Date: AERA, April, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educationel Resources Informetion Center (ERIC) nonexclusive permission to reproduce end disseminete this document es indiceted ebove. Reproduction from the ERIC microfiche or electronic medie by persons other then ERIC employees end its system contractors requires permission from the copyright holder. Exception is mede for non-profit reproduction by libraries end other service egencies to setisfy information needs of educetors in response to discrete inquiries.

Sign here, please

Signature: Renbang Zhu	Printed Name/Position/Title: RENBANG ZHU Measurement Statistician	
Organization/Address: Educational Testing Service Rosedale Rd. MS40-L, Princeton, NJ08541	Telephone: (609) 683-2829	FAX: (609) 683-2130
	E-Mail Address: rzhu@ets.org	Date: 06/03/2003

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org